Harvard Data Science Review • Special Issue 6: Data Privacy for Social Sciences

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

Jordan Stanley[1] **Evan Totty**[1]

[1]United States Census Bureau, Suitland, Maryland, United States of America

The MIT Press

Published on: Jan 05, 2024

DOI: https://doi.org/10.1162/99608f92.c168891c

ABSTRACT

The U.S. Census Bureau has expressed interest in using modern synthetic data modeling techniques for privacy and confidentiality protection in future microdata releases. 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 the SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB) 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 synthetic and confidential results in terms of both magnitudes and statistical conclusions, but the synthetic data is not a perfect replication of the confidential data. 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.

Keywords: synthetic data, data privacy, data accuracy, statistical disclosure limitation, labor economics, applied microeconomics

---

Media Summary

Facing increasingly complex privacy threats, data providers must weigh appropriate options for safely providing data to the public. At the same time, data providers cannot lose sight of producing useful data and statistics. The U.S. Census Bureau has produced synthetic data by modeling internal survey data and administrative records. The synthetic data share statistical characteristics with the internal data without revealing individual respondent records. Data users can obtain results based on the internal data after first developing their analysis and code on the synthetic data. Concerns about synthetic data accuracy as well as the tradeoffs between protecting privacy and providing accurate statistical estimates are valid and worth investigating. So, this article performs statistical analysis to test the accuracy of a synthetic data product: the Survey of Income and Program Participation Synthetic Beta (SSB). We conduct numerous statistical analyses on the public synthetic data and on its private counterpart: the Survey of Income and Program Participation Gold Standard File. We find that the estimates and statistical conclusions are often similar across data sets; however, the results using synthetic data fall short of perfect replication of the results derived from the internal data. Overall, the median difference between a statistic based on the internal data and its synthetic counterpart is 8% for descriptive analysis and 24% for regression analysis. Descriptive analysis performs better for medians than for means or measures of variance. Regression analysis perform better when it does not rely on merged external data after synthesis or within-person models. We also find that regression results based on the synthetic data produce the same sign as results based on the internal data 79% of the time and the same statistical conclusion 63% of the time. This article cannot come close to testing every potential analysis, but these early returns are encouraging. Given advances in synthetic data modeling since the inception of the SSB, we view the accuracy estimates as a floor for the potential accuracy of future synthetic data.

1. Introduction

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 such 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 that are inherently tied to the accuracy of the information. Concerns over the accuracy of data exist independent of privacy concerns; survey nonresponse and measurement error have been increasing over time (Meyer et al., 2015). While increasing data protections could further reduce accuracy of the final data product, not having public trust that data are protected could also affect data accuracy through inaccurate responses or through declining response rates (Cohn et al., 2020; Couper et al., 2008).[1]

Data providers must balance the need to protect the information for millions of respondents with producing accurate statistics and data for public use. Synthetic data present one avenue for protecting privacy (Abowd et al., 2019, 2020). 'Synthetic' refers to a data set where some or all the data released are based on modeled estimates from confidential data.[2] 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 to reduce potential accuracy loss. Synthetic data can come from partial synthesis in which only some variables 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 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 synthetic data is an internal validation 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.[3] 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.[4]

For more than a decade, the U.S. Census Bureau has run a validation system with the 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). The GSF links microdata from select panels of the SIPP to microdata from several administrative data sources such as the Social Security Administration (SSA) and the Internal Revenue Service (IRS). Because of confidentiality issues related to linking SIPP data with administrative records, it was not feasible to release the GSF to the public. The SSB was created and made publicly available instead, allowing for unprecedented access to linked survey-administrative microdata that otherwise carry much higher barriers to access. The SSB is a synthesized version of the GSF featuring a subset of variables found on the internal data sets. This synthetic data set was created by modeling the GSF through sequential regression multivariate imputation (SRMI).[5] The SSB has evolved over time (see Benedetto et al., 2018, for more details), but in the current version (Version 7), four synthetic files are created from a 'snapshot' of the internal GSF. The snapshot serves as the data set for internal validations. Additional technical information on the development of the SSB is available in Benedetto et al. (2018).

In practice, data users can apply for access to the publicly available SSB.[6] 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 et al. (2015); Carr et al. (2022); Carr and Wiemers (2022); Juhn and McCue (2016); Hampton and Totty (2023); Henriques (2018); Kejriwal et al. (2020); Neumeier et al. (2018).

Still, little research exists that shows how empirical analyses differ between synthetic and confidential data. Many researchers have conducted research using the SSB and GSF or other similar synthetic data sets with validation, but rarely have authors presented results from both the confidential and synthetic data in their papers.[7] With the validation option, it is not necessary that the SSB or other synthetic data sets with validation accurately replicate every statistical relationship between variables in the confidential data; however, synthetic data accuracy is still beneficial even when a validation option is available. At a minimum, maintaining structural relationships is important for users developing code and learning and understanding the data. Beyond the structure that is inherent in the survey or census, synthetic data that maintain evidence of relationships between variables can be useful for generating and vetting research ideas and plans for analysis. Furthermore, a potentially large benefit of accurate synthetic data is generating useful statistics at relatively lower costs. 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. The synthetic data will likely not stand alone for all use cases, but relatively accurate statistical estimates from synthetic data that are useful to researchers can simultaneously help meet privacy constraints and reduce the need for validation. 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. Carr et al. (2023) also compare results generated from the SSB to results generated from the GSF. Their analysis provides a deeper assessment of the usefulness of the SSB for a particular application (earnings dynamics), whereas our analysis provides a broader view of the SSB's strength and weaknesses across a wide range of topics and statistics.

The main contribution of our article is to present new evidence on how research using a synthetic data set compares to the same research performed using the corresponding confidential data. 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.[8] Most of these differences are consistent with expected and interpretable patterns related to modeling decisions when synthesizing the data, merging external data to the synthetic data after synthesis, and using statistical methods that are sensitive to outlier observations, which are likely to be perturbed during synthesis. For example, our analyses show that medians are much more similar between the GSF and SSB than means or measures of variance. This is consistent with the fact that outlier values, which affect means but not medians, are likely to be perturbed by privacy protection methods and the idea that privacy protection introduces some 'noise,' which may affect measures of variance. For model-based results, our analyses show that methods that rely on merged external data and/or within-person methods are much less similar between the GSF and SSB than other model-based results. The relatively poor performance of models that rely on merged external data is due to the simple fact that the external data were not used in the synthesis process and therefore their relationship with GSF variables is inherently altered. The relatively poor performance of the within-person models is likely related to the fact that the SSB modeled longitudinal earnings in levels rather than in year-over-year changes.

The overall accuracy of the SSB can be evaluated using competing benchmarks. One might care about the exact magnitude of statistics, in which case we find that the median absolute relative error for a statistic based on the SSB compared the GSF is 8% for our descriptive analysis and 24% for our model-based analysis. One may also care about confidence interval comparisons. For our model-based results, we find that the SSB confidence interval overlaps with the GSF confidence interval 52% of the time, covers the GSF coefficient estimate 35% of the time, and covers 33% of the GSF confidence interval on average. Finally, when validation is available and synthetic data is only needed for developing an analysis plan, we may instead prefer to assess sign and statistical significance. In this case, we find that the SSB produces the same sign as the GSF 79% of the time for our results and the same statistical conclusion 63% of the time. When the statistical conclusion is not the same, it is almost always because a statistically significant relationship in the GSF became insignificant in the SSB.

We have inferred three key takeaways from our results. First, modeling decisions inherently prioritize particular use cases, so feedback from data users is mutually beneficial. Second, the science for generating and evaluating synthetic data has advanced in the years since the SSB was first developed and is still evolving. Third, validation and/or verification are important complements to synthetic data. We will discuss these points in greater detail later in the article. One important caveat is that our study 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 data sets or synthetic data in general. While SRMI was the frontier of science for generating synthetic data when the SSB was first developed in 2003, nonparametric classification and regression tree (CART) methods can generate more accurate synthetic data while also being easier to implement (Drechsler & Reiter, 2011; Reiter, 2005; Reiter & Kinney, 2012). We also want to emphasize that our characterization of accuracy is based on the extent to which results from the SSB replicate results from the GSF. Implicit in this characterization is the assumption that the confidential data represent 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 article, but we acknowledge this limitation and provide more discussion later in the article.

We next provide an overview of our data samples and analyses in Section 2 with the details and results of every individual use case. We summarize the differences between the GSF versus SSB results and discuss takeaways and caveats related to our results in Section 3. Finally, we summarize our conclusions and discuss future research goals in Section 4.

2. Quantitative Analysis on Synthetic Data and Internal Reference File

In this section, we first provide information on modeling details of the SSB, selection of our analytical samples, and construction of confidence intervals with synthetic data. We then discuss the details and results of each analysis we performed.

The GSF was synthesized four separate times by building up the synthetic data as a series of sequential conditional marginals, using only previously synthesized variables as explanatory variables. After estimating a given conditional model on the GSF, a synthetic value was imputed for each record based upon the most up-to-date synthetic data. Hence, while the synthetic variables were not used in the model estimation, they were used to impute other synthetic values in order to keep the synthetic data internally consistent. For version 7.0 of the SSB, missing values due to nonresponse were left in the GSF. Indicators for missing values were then modeled and imputed along with the other possible variable values. Four different modeling features were specified for each variable: (1) model type (ordinary least squares [OLS], logistic, or Bayesian bootstrap); (2) structural relationships; (3) variable value restrictions; and (4) grouping and conditioning variables. The modeling step was used to estimate a posterior predictive distribution (PPD) for each variable. Draws from this PPD were then used as the synthetic values. Relevant modeling details for the longitudinal earnings variables include that they were modeled in levels rather than changes; each time period was modeled separately, but the earnings models did condition on earnings in the previous period; and the earnings variables underwent a transformation process before modeling so that their distribution more closely resembled a standard normal (and the transformation was undone after imputation). More information on modeling details can be found in Benedetto et al. (2018); however, many modeling details were intentionally withheld to ensure that original records in the GSF cannot be reconstructed nor can the SSB be linked to the publicly available nonsynthetic SIPP microdata.

In the present article, we perform several types of analyses—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 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 social science disciplines. There are thus many use cases that we can test. Second, earnings is a continuous measure, so we can test 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. The GSF and SSB have data from roughly 783,000 individuals, and the linkage rate between the SIPP and the administrative records varies by SIPP panel (75% in 1984, 78%–84% for the 1990, 1991, 1992, and 1993 panels, 47% for 2001, 72% for 2004, and 82% for 2008). In assessing missing data patterns, we utilize the full GSF and SSB data sets. Our primary analytical samples consist of individuals who have nonmissing annual 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 nonmissing SIPP earnings data (which include reported $0) for all 12 months in a given calendar year. We make such a restriction in the SIPP data for this study for the sake of getting the most 'like versus like' comparisons to keep the emphasis on assessing differences between estimates from the GSF and estimates from the SSB. Researchers comparing the two sources of earnings data would typically calculate annual earnings by summing monthly earnings data and requiring that all 12 months have nonmissing values; therefore, we defined those variables and samples accordingly. We acknowledge that this choice could bias toward more similar estimates in the SIPP versus the DER for the analyses that rely on this distinction. The 783,000 starting count is reduced to roughly 492,000 person-year observations when we restrict to individual-year observations with annual measures of earnings from both the DER and the SIPP (as described above).

In many of our analyses, we focus on a sample 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 a common regression outcome of interest, this subsample typically serves as the basis for our regression samples. Each regression analysis includes additional restrictions such as nonmissing covariates and assorted age range limitations. These sample definitions are described in the respective subsections of Section 2 pertaining to each separate regression analysis. Note that for select analyses that replicate tables or figures from other papers, we will have analytical subsamples that do not directly stem from our main samples. Analytical sample sizes are included as notes in the tables and figures in the Appendix.

Results based on the SSB must be combined across all four synthetic data sets for proper estimation and inference. The GSF is synthesized multiple times so that users can account for synthesis uncertainty. Statistics such as means, medians, and coefficient estimates are estimated on all four data sets and then simply averaged together. The corresponding measure of variance used to compute standard errors and confidence intervals is based on Reiter (2004), which is the approach recommended to SSB users (Benedetto et al., 2018). The variance statistic developed in Reiter (2004) combines the variance of the statistic across the four data sets and the average across the four data sets of the estimated variance of the statistic on each individual data set. The intuition for this method is that it accounts for synthesis uncertainty by taking into account the variation in the statistic across multiple implicates of synthetic data. We adhere to these methods in producing our SSB point estimates, standard errors, and confidence intervals.

The descriptive analysis includes earnings distributions, means, standard deviations, medians, ratios, and counts. Our regression analyses pertain to research topics such as wage premiums, returns to schooling, minimum wage policy, retirement, work disability onset, the Vietnam War draft lottery, and intra-household earnings. In total, we computed more than 500 statistics pertaining to high priority and common use cases of the GSF and SSB data, including replications of published papers covering more than 10 distinct topics from applied microeconomics research. We will now proceed with summarizing our findings (with most of our results tables and figures included in the Appendix), discussing the results, and reflecting on potential takeaways and future considerations.

2.1. Descriptive Analysis

2.1.1. Earnings Summary Statistics

Figure A1 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 earlier in this section. We then drop the top and bottom 5% of annual earnings observations for each data set 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, we can say that 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 A2 and A3 study how well the SSB replicates differences between the SIPP and DER earnings. Figure A2 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 overreporting in survey earnings in the lower part of the DER distribution and underreporting 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 overreporting versus underreporting 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 A3 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 person-year observation 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 A4 through A7 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 A4 and A5 correspond to the full sample described in the main text, while Figures A6 and A7 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 A4 and A6 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 A5 and A7. 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 pattern holds 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 A8 and A9 show statistics for differences between the administrative record values for earnings (the DER) and the self-reported values from the SIPP. Figure A8 presents the ratio between these earnings values, while Figure A9 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 A8, 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 A5 and A7. Turning to the absolute differences (see Figure A9), 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 nonmissing 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 A1. 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 largest differences between the SSB and GSF results 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.

2.1.2. Positive, Nonpositive, and Missing Earnings Patterns

This subsection shares results from analysis of different categorizations of earnings values—positive earnings (i.e., nonmissing and greater than zero), nonpositive earnings (i.e., nonmissing 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 nonmissing 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 was not 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 GSF and full SSB 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 nonmissing 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 years 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 nonmissing 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 nonmissing is 0.14 while the same estimate for DER nonmissing 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.[9]

We also perform descriptive analysis of positive vs. nonpositive 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 A10 shows the proportions of records with different combinations of positive and nonpositive earnings in the different data sources and data sets. The figure format is like that in Figures A4 through A9. Looking within assorted samples and subsamples, we measure the proportion of records fitting the given characteristic (e.g., positive SIPP earnings and nonpositive 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 nonpositive DER earnings, while the third and fourth bars show records with nonpositive SIPP earnings and positive DER earnings. The "positive DER and nonpositive SIPP" proportions are usually within a percentage point or two comparing GSF to SSB within each *x*-axis category; however, the "positive SIPP and nonpositive 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 than or equal to zero for the purposes of our analysis, an individual would need to have twelve nonmissing months of SIPP earnings that sum to no greater than zero.

To help assess what may be driving this finding, we generate some additional statistics concerning positive vs. nonpositive earnings. For our research sample where both SIPP and DER earnings are nonmissing, the GSF has positive DER earnings in 68.55% of such cases while the SIPP annual earnings are positive in 68.05% of cases. In the SSB, DER earnings are positive in 70.25% of cases, and SIPP annual earnings are positive in 75.34% 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% to 75.34%) than in the frequency of positive DER earnings (68.55% to 70.25%). Looking at monthly SIPP earnings in the full samples, 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 nonpositive DER" cases in the SSB relative to the GSF.[10]

The final check we perform in this vein is to assess nonpositive DER observations over time. Figure A11 shows the comparison of the proportion of nonpositive DER earnings in the full GSF sample to that in the full 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 nonpositive DER earnings in the GSF and the proportion of nonpositive 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 article, an interesting finding of the analysis presented in Figure A11 is that the GSF initially shows a higher percentage of nonpositive 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 versus nonmissing and positive versus nonpositive. We do see closer replication of the GSF patterns for static individual estimates (e.g., overall frequencies of missing values, proportions of nonpositive DER earnings by year) than for estimates involving within-person dynamics (e.g., frequencies of individuals with full calendar years of nonpositive monthly earnings). We feel more research in this area is warranted but dedicate the remainder of this article to comparing GSF and SSB results from regression analysis of social science questions.

2.2. Model-Based Output

2.2.1. 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 A2 shows predictors of having missing SIPP monthly earnings or DER annual earnings. We regress binary indicators for missing SIPP or DER earnings on demographic characteristics and census region indicators. 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 that many variables 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 still statistically significant). Overall, the SSB replicates many of the demographic correlates with missing earnings present in the GSF data.

2.2.2. Mincer Regressions

Next, we perform analysis in the style of the seminal Mincer (1974) model and many subsequent labor economics analyses. Table A3 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 nonmissing values for all covariates.

In Table A3, 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% for high school degree, 67% for some college, 141% for college degree, and 328% for advanced degree when using DER earnings as the outcome. The corresponding estimates from the SSB are 49%, 92%, 206%, and 453%.

To summarize, the SSB shows larger premia across education categories with the bigger differences occurring as education increases. For example, the expected earnings premium from an advanced degree is roughly 38% 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. Overall, while the estimate magnitudes differ to varying degrees, the statistical takeaways from these Mincer-style regression analyses are the same comparing SSB to GSF in terms of sign and significance.

2.2.3. Time Series Evidence: Wage Gaps Over Time and Lifecycle Earnings

Figures A12–A15 show time series results for several constructed variables or wage gaps. The results are generated 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 & Ransom, 2019; Card & Lemieux, 2001).[11]

Figure A12 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.[12] 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. There is also some evidence of a flattening 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 also 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 & Welch, 1990).

Figure A13 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; however, 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 important 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 & Lewis, 2012; Blau & Kahn, 1997; Mulligan & Rubinstein, 2008). The Black–White wage gap has been widening since 1980 (Daly et al., 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.[13] 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 A14 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. The main difference is that the SSB results do not show as clear of a delineation between a shrinking wage gap from 1980 to 2000 and a stable wage gap since 2000. Figure A15 depicts 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 this set of time series results 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.

2.2.4. Returns to Schooling

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

Early studies on the returns to schooling were based on OLS estimates of regressions that attempt to explain wages or earnings as a function of schooling and experience (Mincer, 1974).[14] Decades of work since then has focused on omitted ability bias in the 'Mincer equation.' Ordinary least squares 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 et al., 2006). As a result, a large body of work has emerged using a variety of econometric techniques in an 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 & 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 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 the calendar year and individuals without missing covariates. The results are shown in Table A4. 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% 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% 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% return in the GSF versus 13–16% in the SSB, both statistically significant), but the IV estimates in the SSB are smaller than OLS (11–13% return for 2SLS versus 13–16% for OLS) and not statistically significant.

Panel C of Table A4 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 versus SSB.[15] 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 with the interactive fixed effects estimators from Bai (2009) and Pesaran (2006) to account for omitted ability bias. The Panel A portions of Tables A5–A7 reproduce several of the main results in their paper, while the Panel B portions attempt to replicate their results when we run their code on the SSB.[16] The analysis in Table A5 is similar to the analysis described in the previous paragraphs: the authors estimated Mincer equations with OLS and IV (based on quarter of birth) using both cross-section and panel data samples. Table A6 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 A7 shows results for heterogeneous coefficient panel data models that include interactive fixed effects.

The GSF results in Panel A of Table A5 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 A6 and A7, 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%. The SSB results are always small and close to zero. Only one of the 10 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 A5–A7 that results that 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 A5 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 relatively close between the GSF and SSB (9.2% GSF return versus 7.0% SSB return for column 1 [24% reduction]; 10.5% GSF return versus 9.3% SSB return for column 4 [11% reduction]). The specification that does rely on person fixed effects shows a much larger attenuation of the return to schooling (7.7% GSF return versus 4.2% SSB return in column 3 [45% reduction]). All the results in Tables A6 and A7 rely on within-person earnings variation. The Table A6 results include both individual fixed effects and interactive individual and time fixed effects, which capture time-varying returns to unobserved individual characteristics. The Table A7 results include interactive 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 that 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.

2.2.5. Vietnam War Draft Lottery and Civilian Earnings

In this subsection, 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 nonrandom 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.[17] 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.[18] Later, a number from 1 to 366 was chosen and all individuals whose RSN was below that 'draft eligible' ceiling were selected as draft-eligible.[19] 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 Earnings Record (SER) to build a panel of individual-level annual earnings from 1960 to 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 A8 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 10 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.

2.2.6. 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 policymakers. The GSF presents an opportunity to analyze the long-term earnings effects of SSDI application and receipt. Our 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 6 years or more before disability onset, and then relative year dummy variables are created for years between 5 years prior to disability onset and 10 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 nondisabled 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 A16 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 nondisabled 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 A9.

In both the event study and the difference-in-differences analysis, the signs and significances for most of the estimates are the same between the GSF and SSB. 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 A16). 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 A16). 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 A9). 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 nondisabled individuals (or individuals who could be identified as lacking work-limiting disabilities) in the GSF sample.

These findings make sense in the context of some of our other analyses. The SSB does capture the strong and expected negative relationship between SSDI and likelihood of positive earnings, but it does not 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 versus positive distinction (e.g., zero earnings in the GSF versus $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 an 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.

2.2.7. 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 et al., 2011, 2017; Dube et al., 2010; Neumark et al., 2014a, 2014b). 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.[20] We then merge publicly available data on the minimum wage in each state over time to the SSB and GSF data by matching on the state and year variables.[21] 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-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 A10. 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, while Panel B shows results based on DER earnings.[22]

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 to 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).[23] 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 A11–A13 reproduce their results from the GSF for employment, permanent exit, and benefit claiming, respectively.[24] The Panel B portion shows the results from the SSB.

The employment results in Table A11 are broken up into three different outcomes: any employment (indicating positive DER earnings in a given year), full-time employment, and part-time employment.[25] Each of the regressions are balanced person-year panel regressions over ages 62–70. We show three different specifications for each outcome.

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). Here, 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 10% 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.

These estimates are all much closer to zero and not close to statistical significance at conventional levels. One coefficient even changes signs compared to its GSF counterpart.

The results for permanent exit from employment in Table A12 are assessed via two different outcomes—partial exit and full exit. Each of those is based on the amount of a person's earnings in a given year relative to their lifetime maximum amount.[26] 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 A8, 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 A13. 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. Again, these hazards models rely on within-person earnings dynamics. 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 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.[27]

In summary, while the SSB did replicate some evidence of a relationship between minimum wages and increased employment for older workers (column 1 of Table A11), most of the statistically significant relationships between the merged external minimum wage data and labor market outcomes in the GSF became insignificant in the SSB. The discrepancy between the GSF and SSB results was even greater when the models relied on within-person earnings dynamics, either as a key source of identifying variation or in the measurement of the outcome variable.

2.2.8. 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 more than the male's. The authors used the GSF to show that this pattern exists in administrative earnings data from the United States and is thus not just a result of measurement error in self-reported survey earnings data. While the authors did not find the discontinuity at 0.5 when first using the SSB, the validated results did confirm the existence of such a discontinuity.

Figure A17 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 figures in the top row, which are 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 (see lower row) 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.

The Bertrand et al. (2015) use case 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 been unable to use administrative data on earnings in the United States or would have needed to find a more difficult and costly path to accessing those sensitive data (e.g., gaining access through a Federal Statistical Research Data Center [FSRDC]). 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.

3. Summary of Results and Discussion

Results from the assorted analyses performed in this article have several implications related to our goal of assessing the comparability between estimates derived from the SSB and those generated from the same analyses using the confidential GSF. In the prior section we summarized the assorted analyses we performed, leaving the corresponding tables and figures in the Appendix. In order to give an overview of the entire set of analyses and overall accuracy of the SSB, Figures 1–4 and Table 1 summarize the descriptive and model-based results. After discussing these summaries, we will highlight some additional patterns that can be found across the use cases.

Figure 1 shows a scatterplot of all the GSF versus SSB results discussed in the prior section and found in the Appendix. The figure shows descriptive and model-based results separately. These scatterplots illustrate the association between the magnitude of GSF and SSB estimates. While we can see 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 and the SSB. In particular, the descriptive estimates are strongly correlated as evidenced by the bunching on and around the 45-degree line. The model-based estimates also cluster near the 45-degree line in many cases, although there is greater deviation away from the 45-degree line for the model-based results.

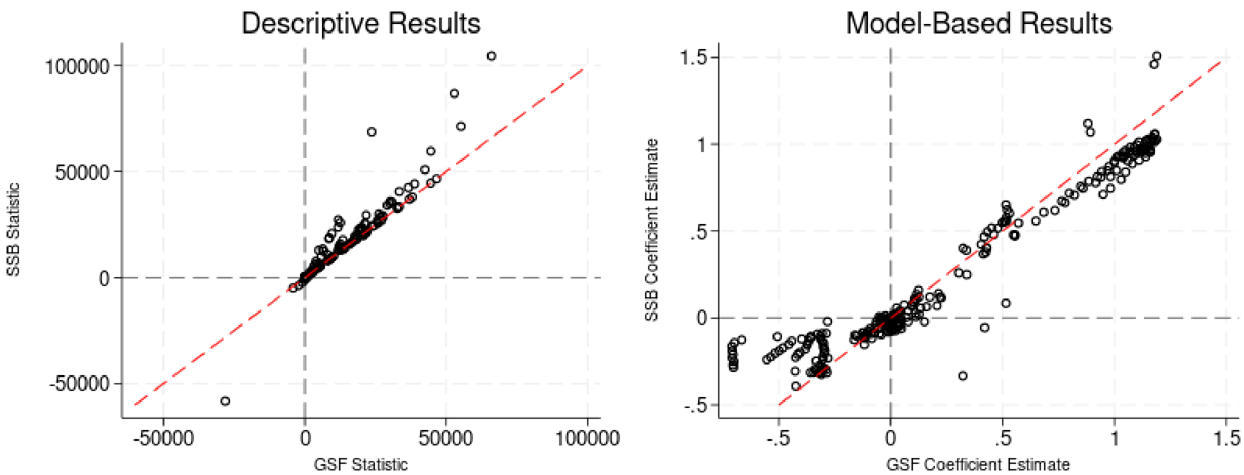

Figure 1. Scatterplot of GSF and SSB results. The right figure plots the GSF versus SSB results for the regression-based results shown in the Appendix. The left figure plots the remaining statistics in the article (e.g., means, medians, ratios, and counts). In each, the *x*-axis is the estimate using the internal GSF, the *y*-axis is the estimate using the SSB, and the red line is the 45-degree line. Figures A1 and A17 in the Appendix are excluded because we only released the figures themselves and not the underlying statistics. The coefficients from Figure A13 were adjusted by subtracting nine (the starting point for the *y*-axis) before including them so that all coefficients from the Appendix could fit in the same figure. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) 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.

The descriptive results illustrate a slope of association greater than the 45-degree line. One potential explanation is based on the fact that some of the descriptive statistics relate to measures of variance—either across individuals, within individuals over time, or between the two sources of earnings information. It is reasonable to expect measures of variance to increase if we think of synthesis as introducing some noise into the data. Figure 2 groups the descriptive results by the type of statistic to investigate this potential explanation. The measures of variance (standard deviations, earnings gaps between individuals or data sources, and ratios of earnings between individuals or data sources) all show the increased slope of association. Means also show an increased slope though, indicating that this is not just a story about variance. The only type of statistic without such an association is medians, which appear to lie directly on the 45-degree line, further illustrating the strong performance of medians as we discussed in the previous section.

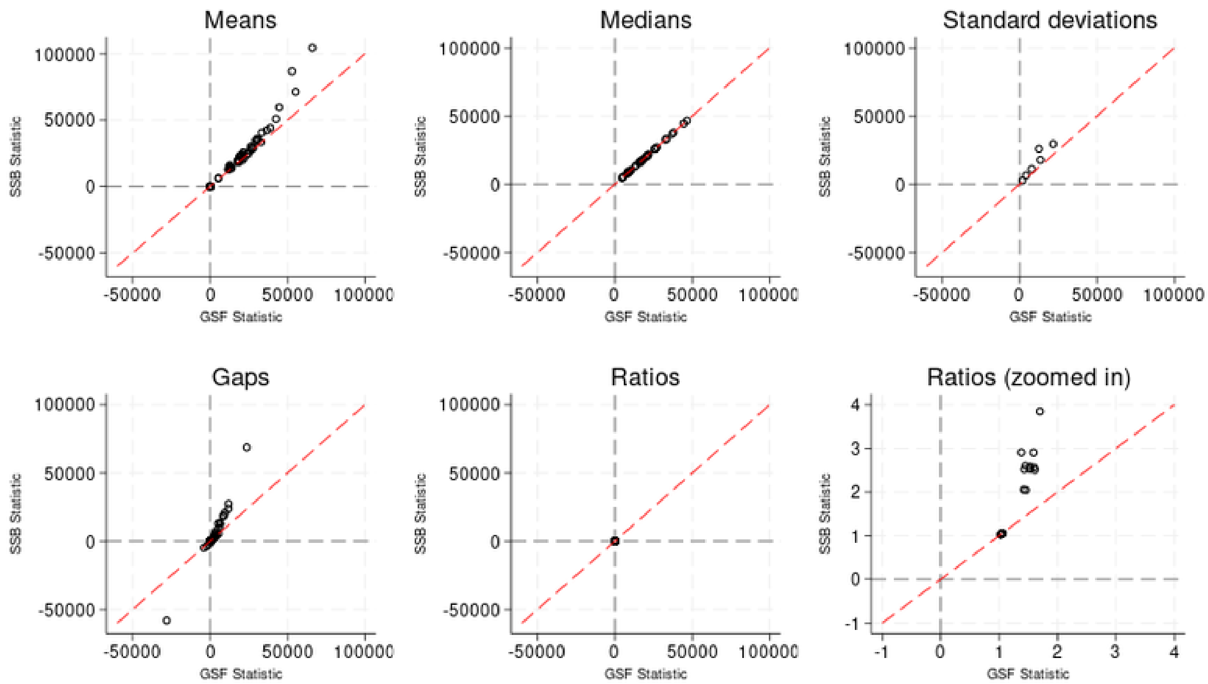

Figure 2. Scatterplot of descriptive results by type of statistic. This figure groups the Descriptive Results in Figure 1 by the type of statistic. See Figure 1 for additional details about the construction of the figure. We classified means as Figures A4, A6, and A11; medians as Figures A5 and A7; standard deviations as Table A1; gaps as Figures A2, A3, A9, and A10; and ratios as Figure A8. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) 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.

The model-based results shown in Figure 1, on the other hand, illustrate a flattening of the slope of association between the GSF and SSB results relative to the 45-degree line. This suggests that model-based results tend to exhibit some attenuation bias, which would be consistent with thinking of the data synthesis process as introducing some noise and weakening some covariate relationships. To investigate this further, Figure 3 groups the model-based results by different types of models. Our results in Section 2 suggested that models incorporating merged external data that was not part of the synthesis process and so-called within-person models that include individual-level fixed effects, interactive individual- and time-level fixed effects, or individual-level hazard models perform worse than other model-based results. Figure 3 confirms this takeaway. Models with merged external data exhibit little-to-no association between the GSF and SSB results and within-person models exhibit a strong attenuation of the relationship. The results based on all other models exhibit a slope of association that is much closer to the 45-degree line.

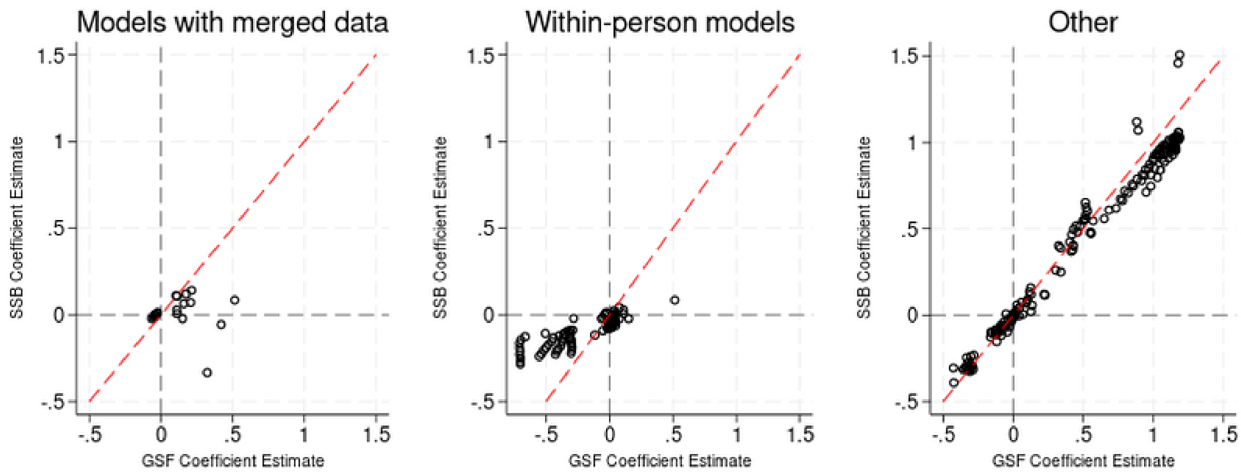

Figure 3. Scatterplot of model-based results by type of model. This figure groups the Model-Based Results in Figure 1 by the type of model. See Figure 1 for additional details about the construction of the figure. Some model-based results qualify as both models with merged external data and within-person models. We classified models with merged external data as Table A10, Table A11, Table A12, and Table A13. We classified within-person models as column (3) of Table A5; Table A6; Table A7; Table A8; Table A9; columns (3), (6), and (9) of Table A11; Table A12; Table A13; and Figure 16. 'Other' includes all other estimates meeting the conditions for inclusion in Figure 1. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) 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.

We further assess the difference in estimate magnitudes by computing the absolute relative error in estimates derived from the SSB compared to their counterparts in the GSF. Figure 4 shows the cumulative density function for absolute relative error (i.e., the absolute value of the difference between the SSB estimate and GSF estimate divided by the GSF estimate). As seen qualitatively in Figure 1, the SSB is more accurate for the descriptive analyses performed in this article than for the model-based analyses. Figure 4 confirms this quantitatively: the median absolute relative error is 0.08 for our descriptive results and 0.24 for our model-based results. So, more than half of our descriptive results derived from the SSB are within 10% of their GSF counterparts, while the majority of our model-based results derived from the SSB are within 25% of their GSF counterparts.[28]

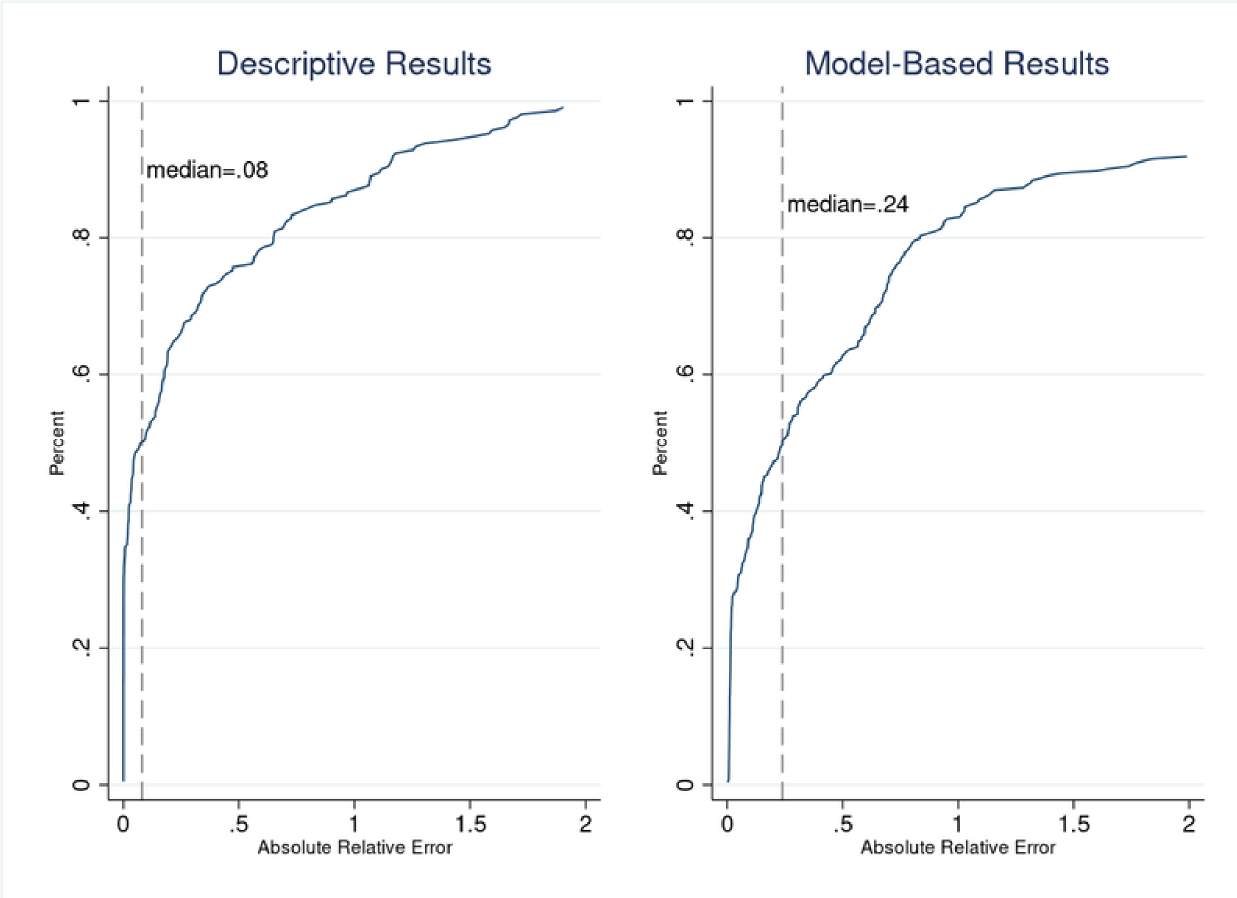

Figure 4. Distribution of absolute relative errors. The right figure plots the Gold Standard File (GSF) versus SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB) results for the regression-based results shown in the Appendix. The left figure plots the remaining statistics in the article (e.g., means, medians, ratios, and counts). The *x*-axis is the absolute relative error comparing the estimate from the SSB to the corresponding estimate from the GSF. The absolute relative error is computed as the absolute value of the difference between the SSB estimate and GSF estimate divided by the GSF estimate. The median absolute relative error is indicated by the dotted vertical line and corresponding value. The figures are truncated at 2 for presentation clarity. Figure A1 and Figure A17 in the Appendix are excluded because we only released the figures themselves and not the underlying statistics.

Finally, Table 1 summarizes differences in the inferences drawn based on the GSF versus SSB. One benchmark is to compare confidence intervals between the two data sets (Karr et al., 2006). Better SSB accuracy would mean that the SSB confidence intervals have a high degree of overlap with the GSF results. Panel A of Table 1 reports results based on this benchmark using 95% confidence intervals. Confidence intervals in the SSB are approximately twice as wide as the GSF confidence intervals on average. The wider confidence intervals are due in part to the use of the confidence interval construction methodology from Reiter (2004) that accounts for synthesis uncertainty. The SSB confidence interval overlaps with some part of the GSF confidence interval

52% of the time and it covers the GSF coefficient estimate 35% of the time. Overall, the SSB confidence interval overlaps with 33% of the GSF confidence interval on average.

Table 1. SSB versus GSF inference comparison.

		(1)	(2)
Panel A: Confidence Interval Comparison			
GSF CI average width			0.069
SSB CI average width			0.129
Proportion of models with any CI overlap			0.521
Proportion of models with GSF coefficient inside SSB CI			0.351
Average fraction of GSF CI overlapped by SSB CI			0.331
Panel B: Sign and Significance Comparison		Count	Percent
(1) Same sign and significance		56	59.57%
(2) Same sign, change significance			
	(2a) GSF significant, SSB not	18	19.15%
	(2b) SSB significant, GSF not	0	0.0%
(3) Change sign, neither significant		3	3.19%
(4) Change sign and significance			
	(4a) GSF significant, SSB not	13	13.83%
	(4b) SSB significant, GSF not	2	2.13%
(5) Change sign, both significant		2	2.13%
Total		94	100%

Panel C: Statistical Conclusion Comparison	Count	Percent
Same statistical conclusion [(1) + (3)]:	59	62.67%
Failed to replicate relationship due to synthesis [(2a) + (4a)]	31	32.99%
Spurious relationship due to synthesis [(2b) + (4b)]	2	2.13%
Opposite relationship due to synthesis (5)	2	2.13%

Note. The comparison includes all regression-based results shown in the Appendix except for those that do not report a standard error or confidence interval (Figures A12–A16 of the Appendix). From 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.

An alternative benchmark is to compare sign and statistical significance between the GSF and SSB. The relative merit of this approach is twofold. First, some results that have little or no confidence interval overlap may not be qualitatively different to a researcher. This is especially true for models with very large sample sizes, as they may have qualitatively similar coefficient estimates but very tight confidence intervals. Second, the SSB was never intended to stand alone. It exists to provide easier and more equitable access to the GSF, and users were always encouraged to validate their results after developing their analysis on the SSB. Therefore, replicating the same sign and statistical significance is arguably a more useful benchmark than confidence interval overlap as synthetic data users with access to validation may find the removal/addition of statistical relationships more problematic than differences in effect size.

Panel B reports the differences in sign and statistical significance between the GSF and SSB. The SSB coefficient estimates have the same sign as the GSF estimates 79% of the time. Panel C summarizes the sign and significance differences in terms of statistical conclusions based on hypothesis testing. The SSB produces the same statistical conclusion for approximately 63% of the coefficient estimates, fails to replicate a statistically significant relationship for approximately 33% of estimates, produces a spurious relationship for approximately 2% of estimates, and produces a significant relationship in the opposite direction for approximately 2% of estimates.

The SSB performs better in terms of sign and statistical conclusions than when considering confidence interval overlap. When the statistical conclusions are different, it is almost always because a statistically significant relationship in the GSF became insignificant in the SSB. While less than ideal, it is not surprising that this will happen sometimes with privacy-protected data. On the other hand, it is reassuring that spurious and opposite relationships in the SSB are exceedingly rare (4% of the time in our models). The relatively poor results for confidence interval overlap despite wider SSB confidence intervals illustrates that methods that account for synthesis uncertainty cannot always make up for biases that arise due to synthesis.[29] The fact that we mostly had very large sample sizes and thus tight confidence intervals may also play a role in the relatively low overlap.[30]

In looking across all the estimates generated for this article, differences between the SSB and GSF are often consistent with interpretable and expected patterns. Statistics that are sensitive to outliers (e.g., means in Figures A4 and A6 of the Appendix) may be less likely to be replicated in synthetic data than statistics that are not sensitive to outliers (e.g., medians in Figures A5 and A7 of the Appendix) because synthetic data inherently attempt to mask sensitive values such as outliers. Figure 2 clearly illustrates this tendency via the strong performance of medians relative means. Additionally, regressions that rely solely on variables already in the data (e.g., in the regression analysis that produced Table A3 of the Appendix) may yield more replicable results than regressions that merge external data onto the synthetic data (e.g., the regression analysis that produced Table A10 of the Appendix). For the latter, variables merged onto synthetic data after synthesis are not used in the synthetic data modeling process, which inherently alters their relationship with the synthetic variables. Figure 3 clearly illustrates this altered relationship as models based on merged external data illustrate almost no association between the GSF and SSB.

Modeling decisions when creating the synthetic data can also explain some of the differences. For example, in Table A4 of the Appendix we see 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.[31] Additionally, Figure 3 clearly shows that results relying on within-person earnings dynamics tend to hold up less well. This pattern can also be tied back to a modeling decision described in Section 2: the synthetic data models for the SSB were primarily based on modeling variable *levels* rather than *year-over-year changes*. A similar data product from the Census Bureau known as the Synthetic Longitudinal Business Database (SynLBD) chose to model within-establishment year-over-year changes 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 article.[32]

Finally, we want to stress a caveat that relates to how we discuss and think about accuracy in this article. Throughout the article 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 nonresponse error, and measurement error (Abowd & Stinson, 2013; Meyer et al., 2015; Meyer & Mittag, 2021). These errors can impact important statistics (Bee & Mitchell, 2017; Meyer et al., 2022; Meyer &

Mittag, 2019). 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 underestimating 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.

4. Conclusion

It is increasingly difficult for data providers to protect the privacy of survey respondents due to the growing availability of public data sets, 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 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. Overall, there is a strong association between the GSF and SSB results. The median absolute relative error for a statistic based on the SSB compared to the GSF is 8% for our descriptive analysis and 24% for our model-based analysis. Descriptive statistics tend to perform better for medians than they do for means (which are influenced by outliers) or for measures of variance (which are influenced by additional noise). Model-based results tend to perform better when they do not rely on external data merged with the synthetic data after synthesis (which was not used in the synthetic modeling process by definition and thus inherently alters the relationship) or within-person methods (which may be due to the fact that SSB earnings were modeled in levels rather than within-person changes). For the model-based results, we also find that the SSB confidence interval overlaps with the GSF confidence interval 52% of the time, covers the GSF coefficient estimate 35% of the time, and overlaps with 33% of the GSF confidence interval, on average. 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: the SSB produces the same sign 79% of the time for our model-based results and the same statistical conclusion 63% of the time. It delivers opposite or spurious relationships only a combined 4% of the time.

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 article 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 researcher will not be considered so for a different use case or researcher. Having a feedback loop can help determine which use cases work best with a given synthetic model or process and possibly which use cases are worth prioritizing. Second, 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 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 nonparametric CART and machine learning are easier to implement and can generate more accurate synthetic data (Drechsler & Reiter, 2011; Reiter, 2005; Reiter & Kinney, 2012). Finally, 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.

Given the improvements in synthetic data modeling since the creation of the SSB and our strict approach to characterizing accuracy, we see our findings as a lower bound for the accuracy level that future synthetic data releases can attain. We will now conclude with a few final reflections and considerations for future applications. The SIPP Synthetic Beta was never intended to stand alone without a validation option, and many future synthetic data sets will likely pair a publicly available synthetic file with a validation or verification server run by the data provider. Not allowing external researchers access to the internal microdata may carry some costs (e.g., foregone research discoveries); however, synthetic data reduce privacy loss, and the combination with validation/verification generates numerous potential benefits. Synthetic data can be seen as a portal that permits wider access to variables, samples, or whole data sets that otherwise cannot be publicly available. While overall accuracy concerns are valid, the results discussed in this article as well as other analyses show that synthetic data can produce useful estimates in many cases. Modern synthetic data generation is more advanced than the methods used with the SSB, and it is likely that synthetic data accuracy will be higher in future synthetic data sets than what we have seen with the SSB. In situations where the synthetic data are not useful, validation and verification can assuage accuracy concerns. Synthetic data with validation/verification could also produce extra benefits such as increasing focus on science rather than results and reducing the prevalence of '*p*-hacking.' All things considered, synthetic data with validation/verification provides a strong option for balancing data privacy constraints and data utility considerations.

Acknowledgments

The authors would like to thank Gary Benedetto, Jason Fields, Caleb Floyd, David Johnson Robert Moffitt, Joanna Motro, Rolando Rodriguez, the editor and anonymous referees at *Harvard Data Science Review*, and seminar participants in the Census Economic Research Brown Bag Series, the National Academies Committee on National Statistics meeting for *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*, and the National Bureau of Economic Research meeting on *Data Privacy Protection and the Conduct of Applied Research* 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).

Disclosure Statement

The authors are both employees of the U.S. Census Bureau. They have no other disclosures to share for this article.

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. https://doi.org/10.1162/REST_a_00352

Abowd, J. M., Benedetto, G., Garfinkel, S. L., Dahl, S. A., Dajani, A. N., Graham, M., Hawes, M., Karwa, V., Kifer, D., Kim, H., Leclerc, P., Machanvajjhala, A., Reiter, J. P., Rodriquez, R., Schmutte, I. M., Sexton, W.N., Singer, P. E., & Villhuber, L. (2020). *The modernization of statistical disclosure limitation at the U.S. Census Bureau*. U.S. Census Bureau Working Paper. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0019793917692788

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security. *American Economic Review, 80*(3), 313–336. https://www.jstor.org/stable/2006669

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/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. https://doi.org/10.1016/j.econedurev.2019.02.003

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279. https://doi.org/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. https://doi.org/10.1257/app.6.2.178

Bee, A., & Mitchell, J. (2017). Do older Americans have more income than we think? In *Proceedings of the Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, *110*, 1–85. 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. 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*. U.S. Census Bureau. https://www2.census.gov/adrm/CED/Papers/CY18/2018-11-BenedettoStanleyTotty-Creation%20SIPP.pdf

Bertrand, M., Kamenica, E., & Pan, J. (2015). Gender identity and relative income within households. *The Quarterly Journal of Economics*, *130*(2), 571–614. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/01621459.1995.10476536

Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McLelland, R., Stallworth, P., Ueyama, K., & Williams, A. R. (2020). A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications. In J. Domingo-Ferrer & K. Muralidhar (Eds.), *Privacy in statistical databases* (pp. 257–270). Springer, Cham. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1162/00335530151144140

Carr, M. D., Moffitt, R. A., & Wiemers, E. E. (2022). Reconciling trends in male earnings volatility: Evidence from the SIPP survey and administrative data. *Journal of Business & Economic Statistics*, *41*(1), 26–32. https://doi.org/10.1080/07350015.2022.2126845

Carr, M. D., & Wiemers, E. E. (2022). The decline in long-term earnings mobility in the US: Evidence from survey-linked administrative data. *Labour Economics*, *78*, Article 102170. https://doi.org/10.1016/j.labeco.2022.102170

Carr, M., Wiemers, E., & Moffitt, R. A. (2023). *Using synthetic data to estimate earnings dynamics: Evidence from the SIPP GSF and SIPP SSB*. SSRN. https://dx.doi.org/10.2139/ssrn.4496224

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. https://doi.org/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. https://doi.org/10.3368/jhr.XXXVIII.3.618

Cohn, D., Brown, A. & Keeter, S., (2020). *Most adults aware of 2020 Census and ready to respond, but don't know key details*. Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/wp-content/uploads/sites/3/2020/02/PSDT_02.20.20_census_full_report.pdf

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. 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. 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. https://doi.org/10.48550/arXiv.2304.02107

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis, 55*(12), 3232–3243. https://doi.org/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. https://doi.org/10.1162/REST_a_00039

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *Theory of cryptography* (pp. 265–284). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11681878_14

Federal Committee on Statistical Methodology. (2005). *Report on statistical disclosure limitation methodology*. US Office of Management and Budget. 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. https://doi.org/10.2307/1913285

Hampton, M., & Totty, E. (2021). *Minimum wages, retirement timing, and labor supply*. U.S. Census Bureau working paper. https://doi.org/10.13140/RG.2.2.32952.67840

Hampton, M., & Totty, E. (2023). Minimum wages, retirement timing, and labor supply. *Journal of Public Economics, 224*, Article 104924. https://doi.org/10.1016/j.jpubeco.2023.104924

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.), *Handbook of economics of education* (Vol. 1, pp. 307–458). Elsevier. https://doi.org/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. https://doi.org/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. https://doi.org/10.1257/aer.p20161120

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*(3), 1–9. https://doi.org/10.1198/000313006X124640

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. https://doi.org/10.1002/jae.2759

Kinney, S. K., Reiter, J. P., & Miranda, J. (2014). SynLBD 2.0: Improving the Synthetic Longitudinal Business Database. *Statistical Journal of the IAOS*, *30*(2), 129–135. 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 Statistical Review*, *79*(3), 362–384. https://doi.org/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. 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. https://doi.org/10.1257/app.20170478

Meyer, B. D., & Mittag, N. (2021). An empirical total survey error decomposition using data combination. *Journal of Econometrics*, *224*(2), 286–305. https://doi.org/10.1016/j.jeconom.2020.03.026

Meyer, B. D., Mittag, N., & George, R. M. (2022). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *Journal of Human Resources*, *57*(5), 1605–1644. https://doi.org/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. https://doi.org/10.1257/jep.29.4.199

Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1086/298220

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

Neumark, D., Salas, J. M., & Wascher, W. (2014b). Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater? *ILR Review*, *67*(3), 608–648. https://doi.org/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. https://doi.org/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. https://doi.org/10.1111/j.1468-0262.2006.00692.x

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, *30*(2), 235–242. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7755-eng.pdf?st=8NPjUqX9

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, *21*(3), 441–462. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf

Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, *28*(4), 583–590.

https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/inferentially-valid-partially-synthetic-data-generating-from-posterior-predictive-distributions-not-necessary.pdf

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

Totty, E. (2017). The effect of minimum wages and employment: A factor model approach. *Economic Inquiry, 55*(4), 1712–1737. https://doi.org/10.1111/ecin.12472

Xie, X., & Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica 27*(4), 1485–1545. http://dx.doi.org/10.5705/ss.2014.067

Appendix

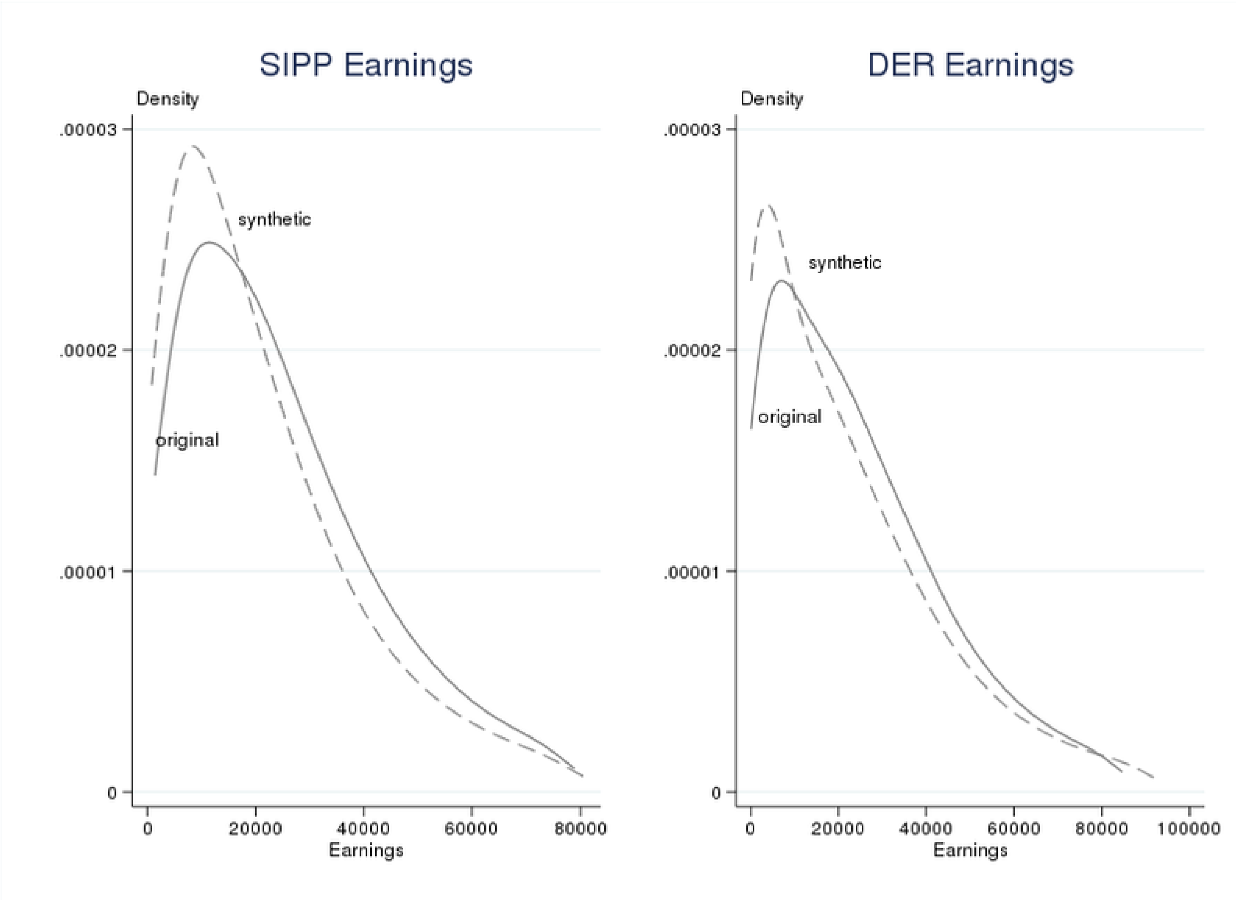

Figure A1. Self-reported and administrative earnings distributions. The figure shows a kernel density estimation (KDE) plot of the density for the Survey of Income and Program Participation (SIPP) and Detailed Earnings Record (DER) earnings in the original (GSF) and synthetic (SSB) data. The sample consists 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 nonmissing if all 12 months were nonmissing. The top and bottom 5% of earnings observations were trimmed for each earnings measure (DER or SIPP) and data source (GSF or SSB). Additional details in Section 2.1.1. From 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.

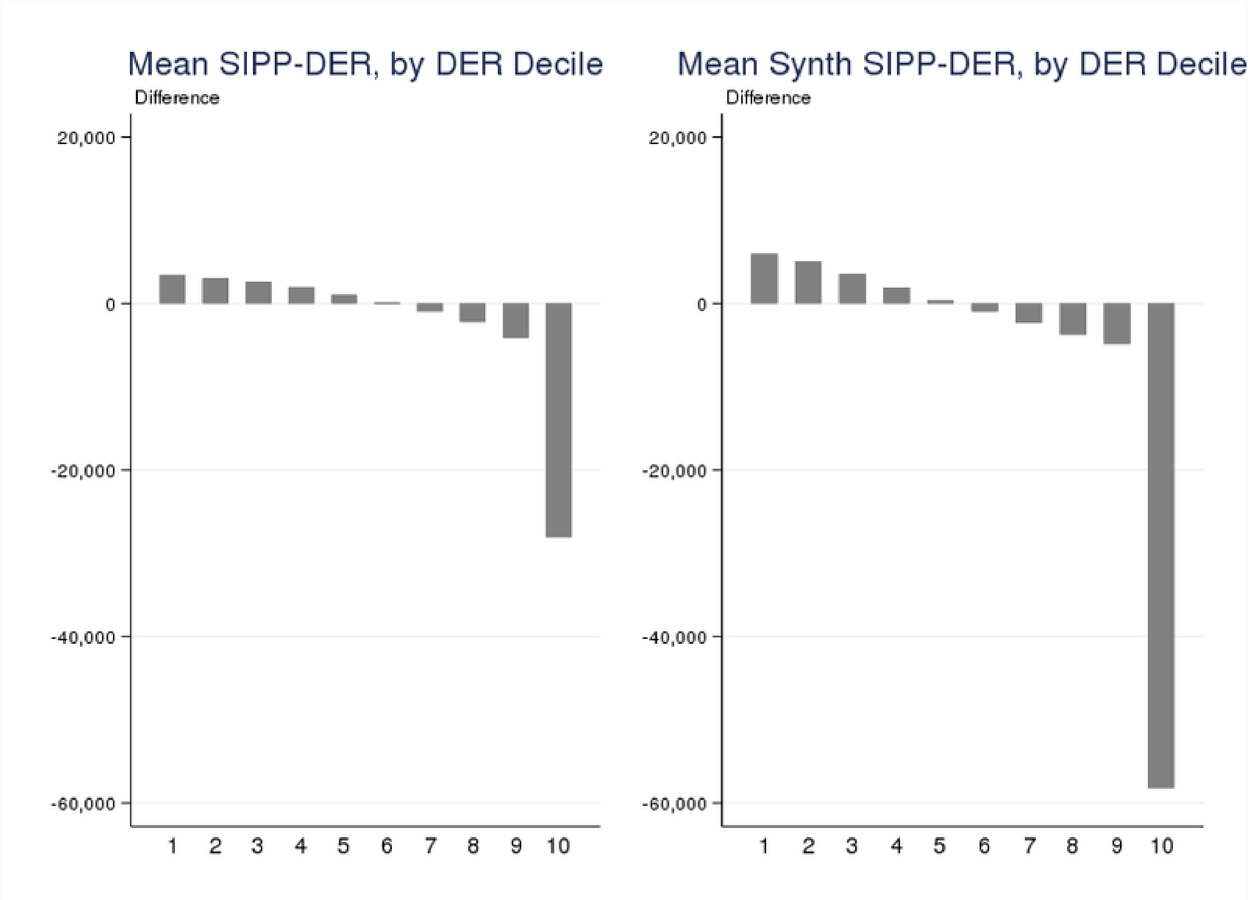

Figure A2. Earnings differences across data sets and sources. The figure shows the average person-level difference between the Survey of Income and Program Participation (SIPP) and Detailed Earnings Record (DER) earnings (SIPP minus DER) in each decile of DER earnings for the original (GSF) and synthetic (SSB) data. The sample consists of all person-year observations during which both DER and SIPP earnings are observed and are positive. Additional details in Section 2.1.1. From 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.

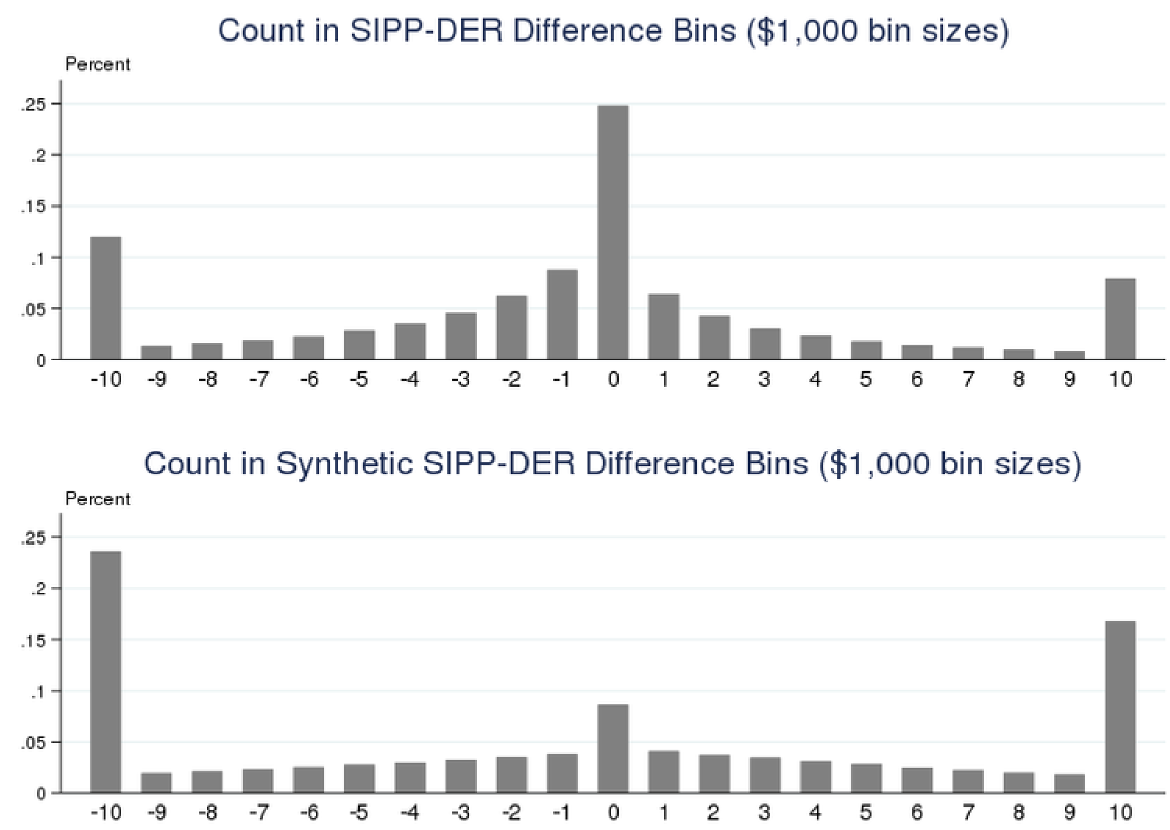

Figure A3. Distribution of earnings differences across data sets and sources. The figure shows a histogram of the difference between the Survey of Income and Program Participation (SIPP) and Detailed Earnings Record (DER) earnings (SIPP minus DER) for the original (GSF) and synthetic (SSB) data. The 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. Additional details in Section 2.1.1. From 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.

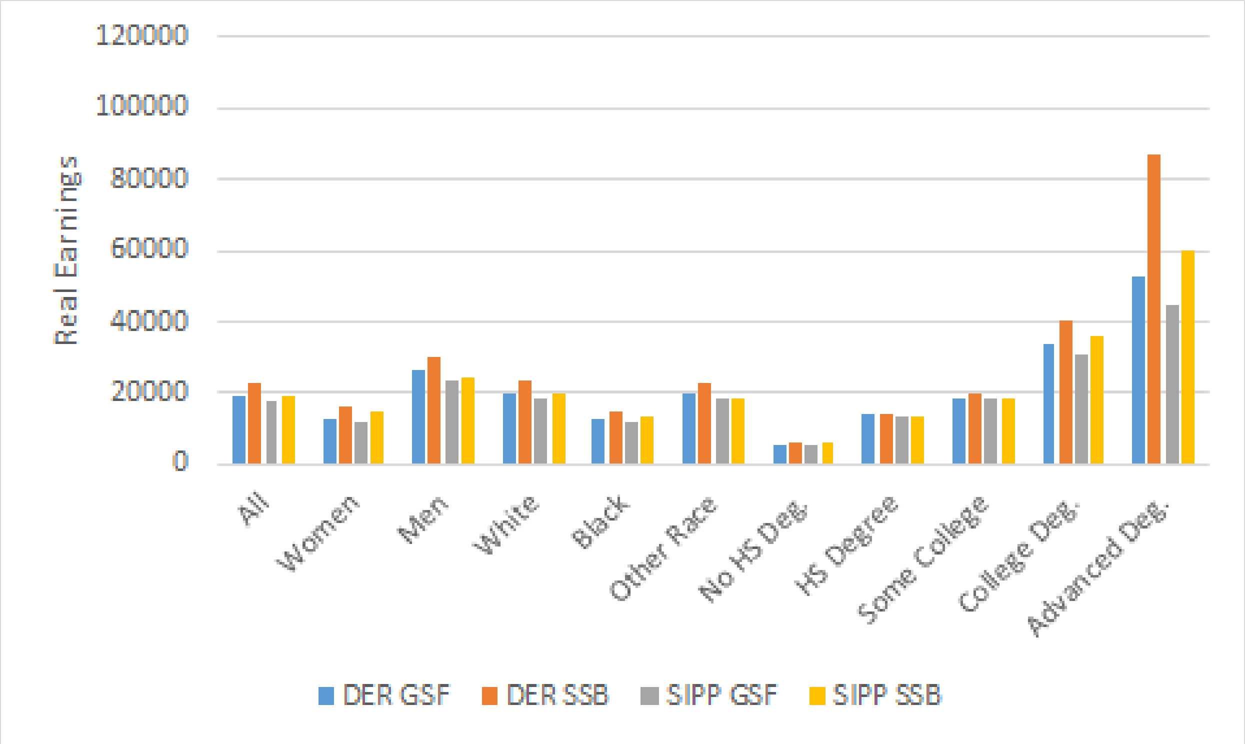

Figure A4. Mean earnings by data source and data set—Full sample. The figure shows mean real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]). “All” consists of individuals in the respective data set who have nonmissing earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

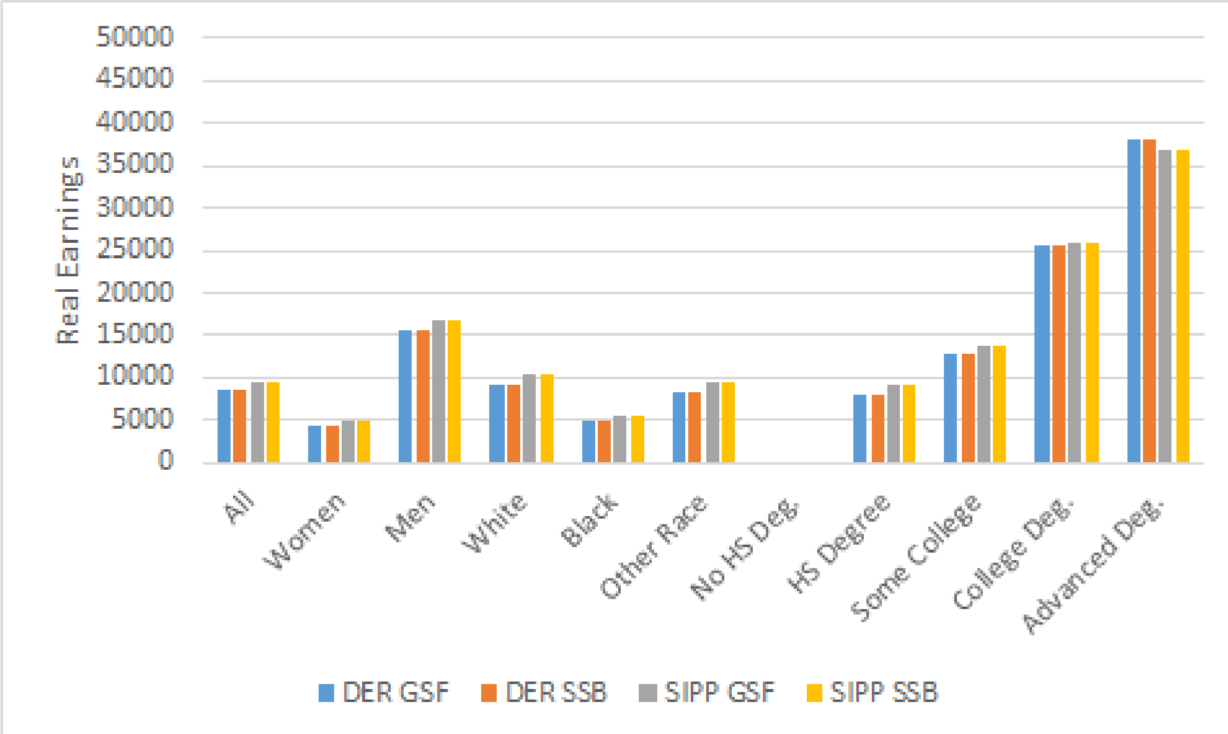

Figure A5. Median earnings by data source and data set—Full sample. The figure shows median real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]). “All” consists of individuals in the respective data set who have nonmissing earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

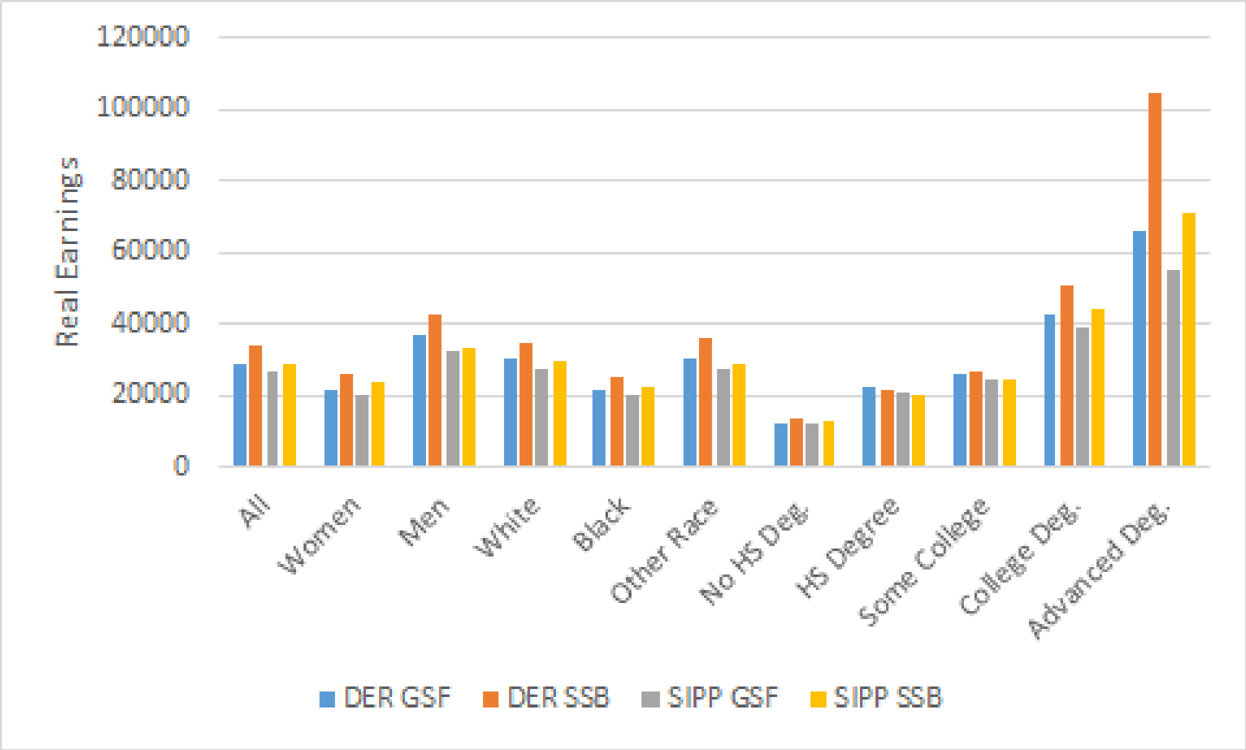

Figure A6. Mean earnings by data source and data set—Positive earners. The figure shows mean real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]). “All” consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

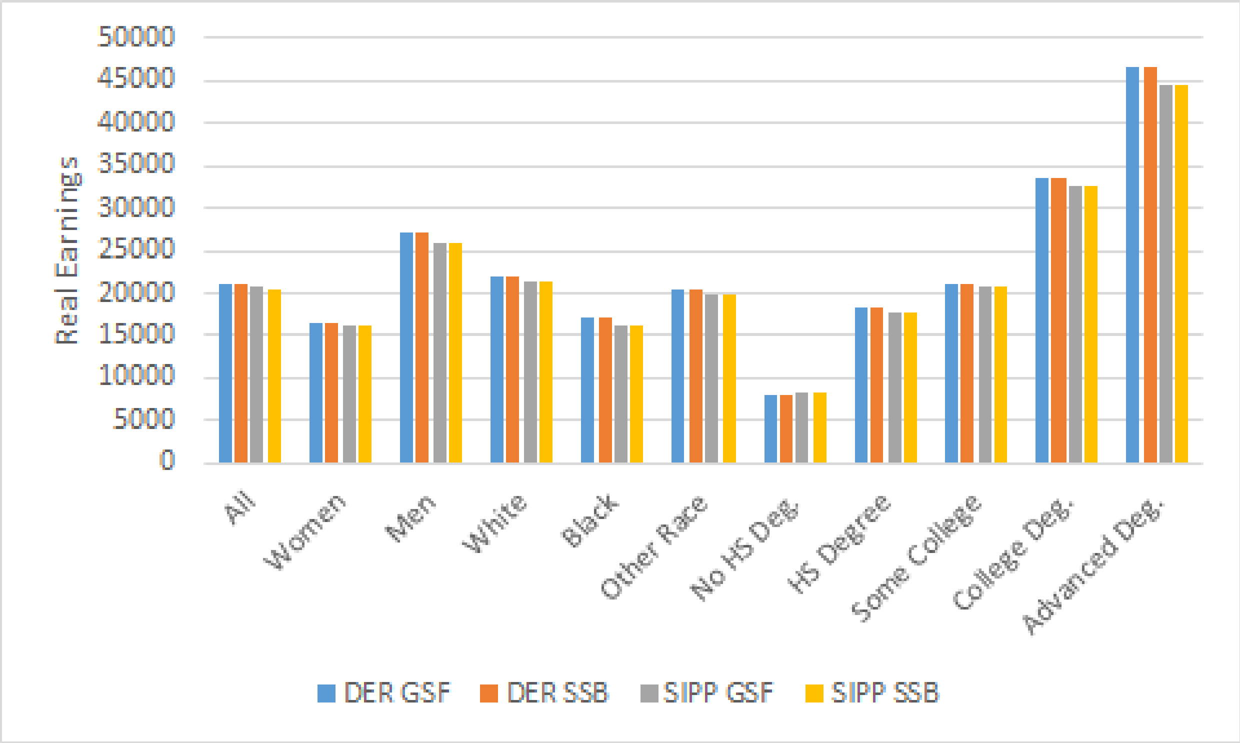

Figure A7. Median earnings by data source and data set—Positive earners. The figure shows median real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]). “All” consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

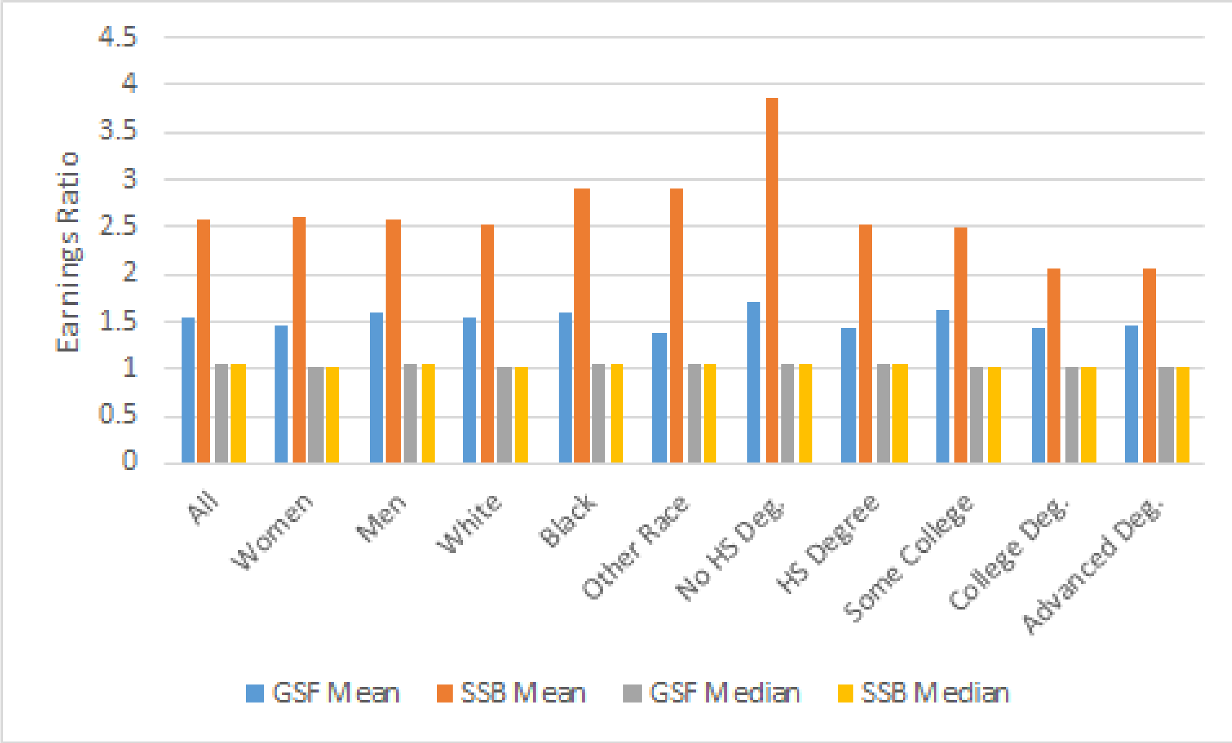

Figure A8. Earnings ratios between administrative records and survey responses. The figure shows average earnings ratios (i.e., quotients) between Detailed Earnings Record (DER) earnings values and Survey of Income and Program Participation (SIPP) earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and statistics (mean or median). “All” consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

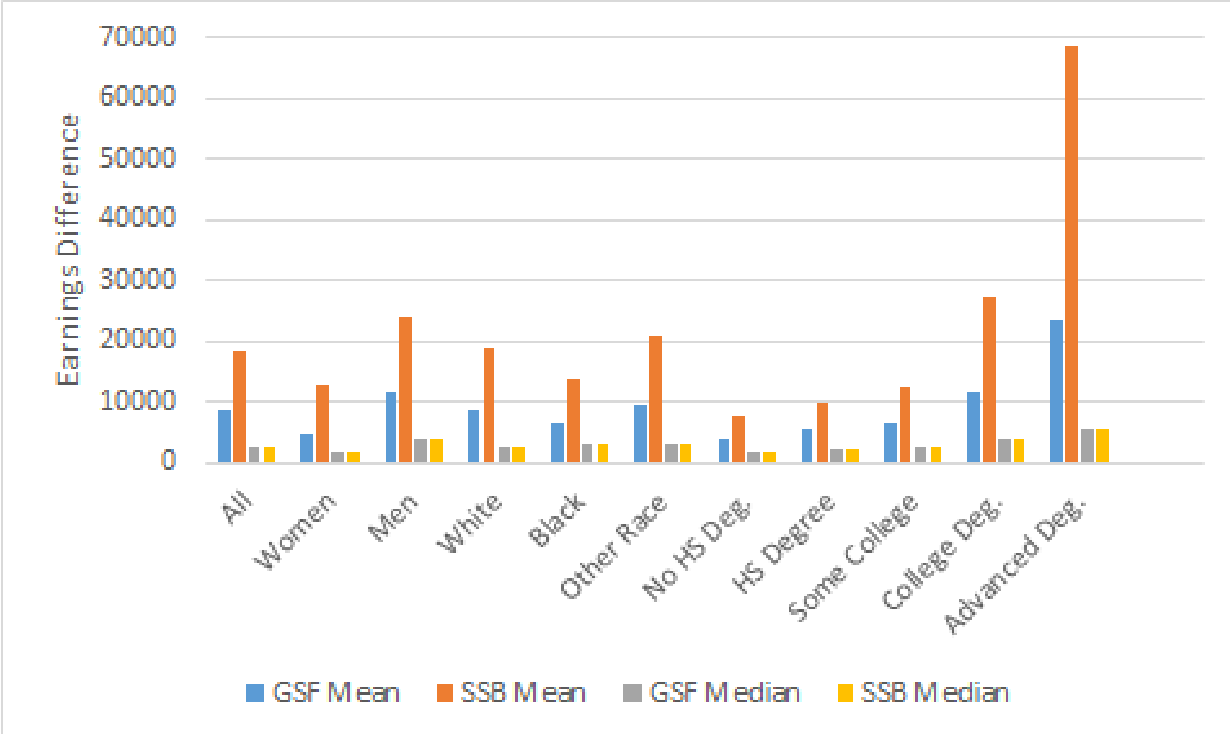

Figure A9. Earnings differences between administrative records and survey responses. The figure shows average or median absolute earnings differences between Detailed Earnings Record (DER) earnings values and Survey of Income and Program Participation (SIPP) earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and statistics (mean or median). “All” consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.1. From 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.

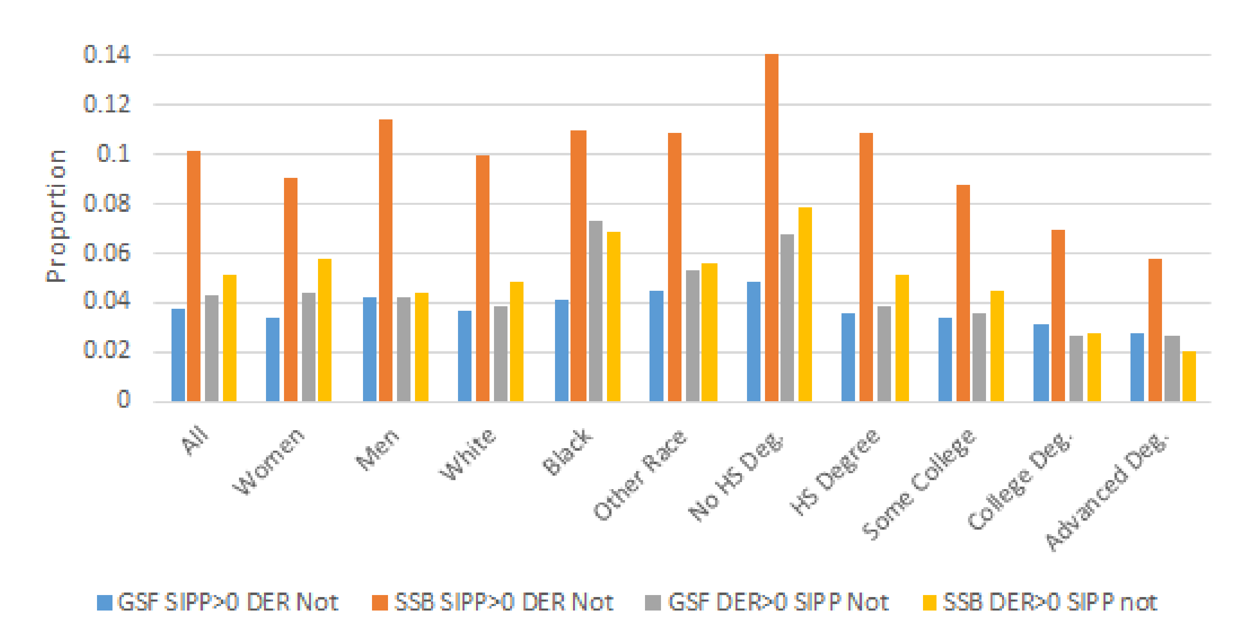

Figure A10. Comparison of positive earnings values across data sets and sources. The figure shows comparisons of proportions of records with positive or nonpositive and nonmissing earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings value source (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]) combinations. “All” consists of individuals in the respective data set who have nonmissing earnings values in both the DER and SIPP, and the other categories represent subsets of this main sample. Additional details in Section 2.1.2. From 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.

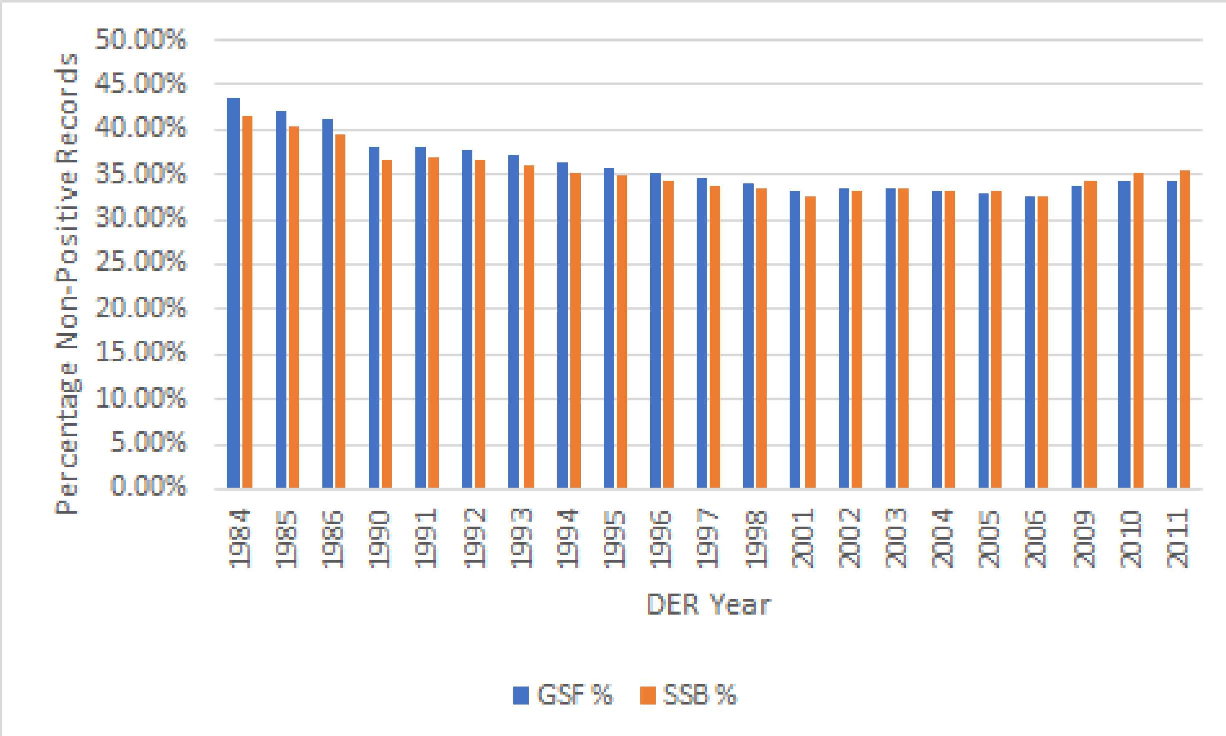

Figure A11. Comparison of records with nonpositive earnings across data sources. The figure shows the proportion of records with nonpositive and nonmissing Detailed Earnings Record (DER) earnings in assorted years. The sample used here consists of all individuals in the respective data set. Additional details in [Section 2.1.2.](#) From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-285.

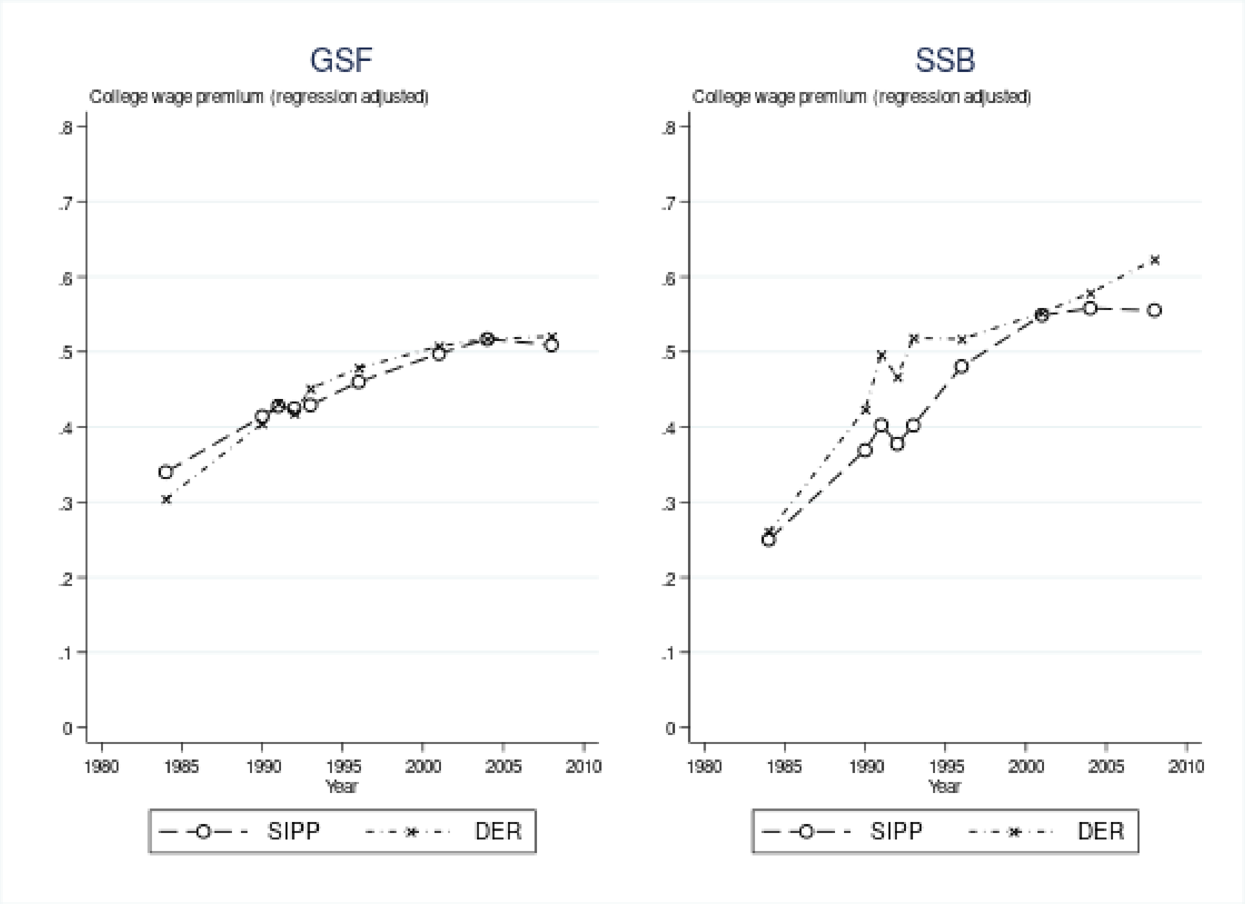

Figure A12. College wage premium estimates comparing data sets. The figure plots the regression-adjusted college wage premium across Survey of Income and Program Participation (SIPP) panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP and Detailed Earnings Record [DER]) and data source (GSF or 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 the main text. Additional details on this analysis are in Section 2.2.3. From 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.

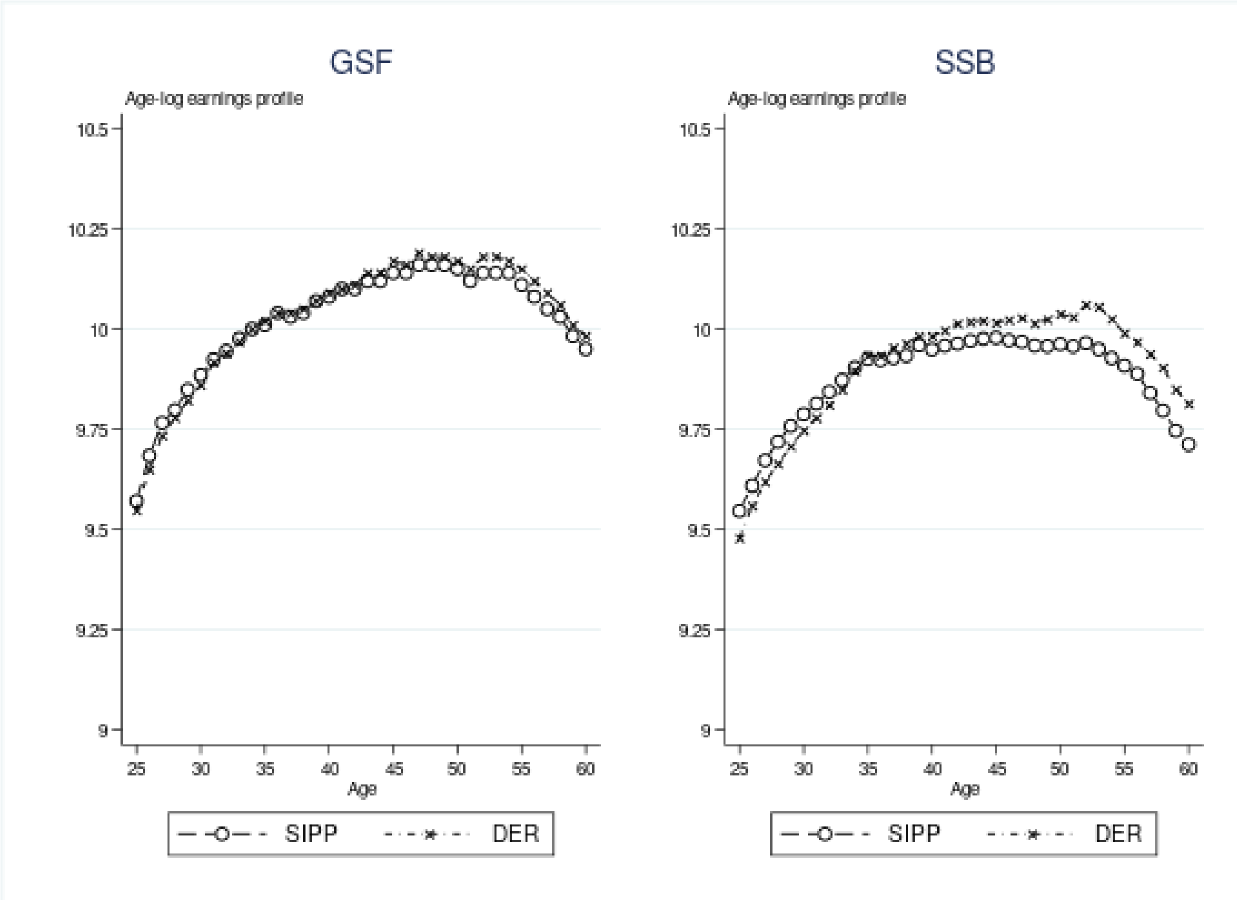

Figure A13. Lifecycle earnings trends comparing data sets. The figure plots the lifecycle-earnings profile. Four different versions of the regression are estimated based on the different earnings measure (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]) and data source (GSF or 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 the main text. Additional details on this analysis are in Section 2.2.3. From 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.

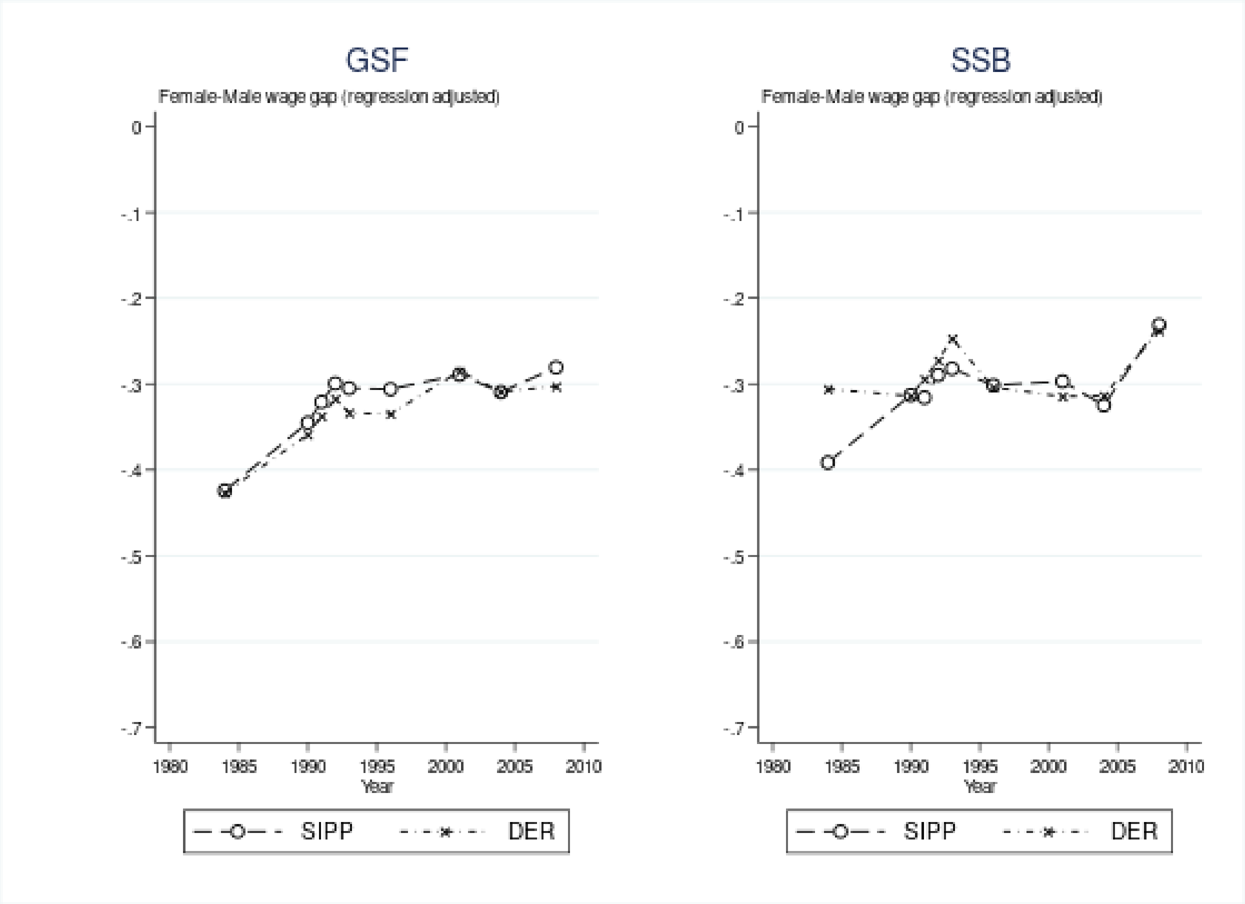

Figure A14. Gender wage gap trends comparing data sets. The figure plots the regression-adjusted gender wage gap across Survey of Income and Program Participation (SIPP) panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP and Detailed Earnings Record [DER]) and data source (GSF or 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 the main text. Additional details on this analysis are in Section 2.2.3. From 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.

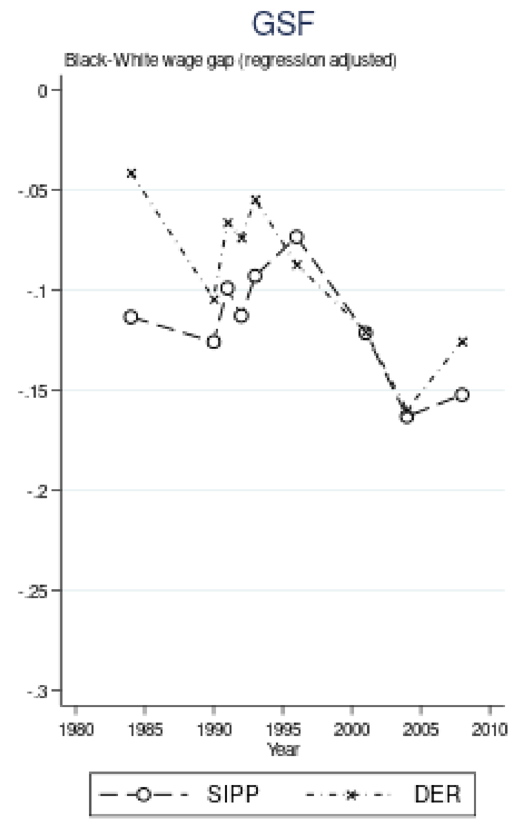

Figure A15. Black–White wage gap trends comparing data sets. The figure plots the regression-adjusted Black–White wage gap across Survey of Income and Program Participation (SIPP) panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP or Detailed Earnings Record [DER]) and data source (GSF or 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 the main text. Additional details on this analysis are in Section 2.2.3. From 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.

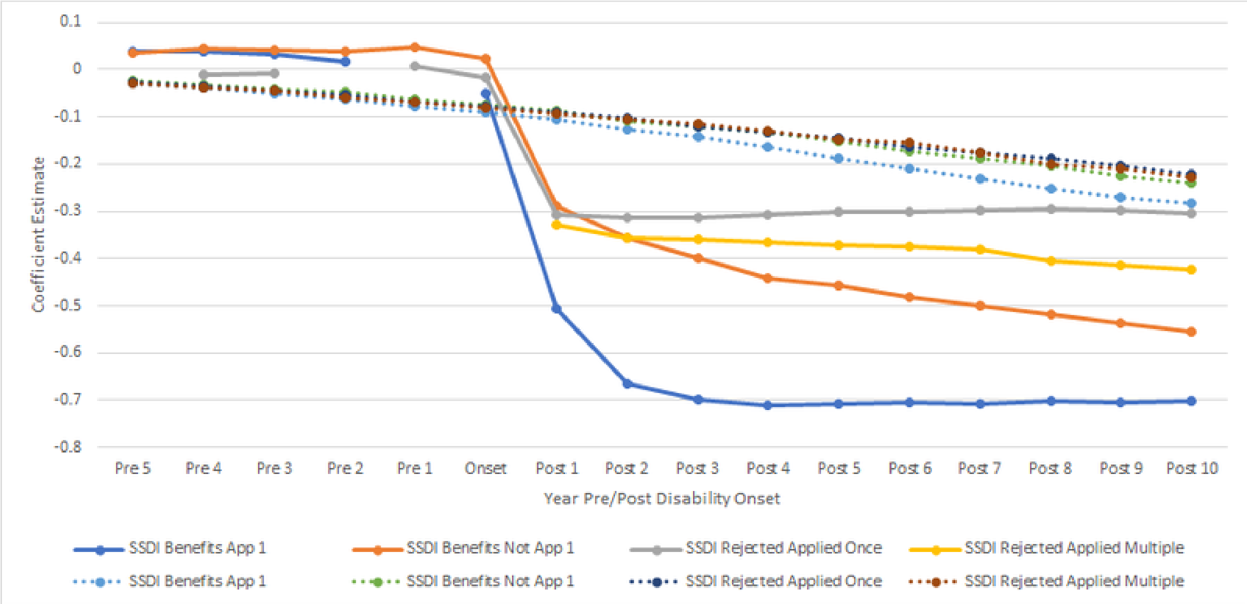

Figure A16. Estimated effects of SSDI application status on likelihood of positive earnings. The solid lines represent Gold Standard File (GSF) estimates, and the dotted lines represent Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) estimates. *N* for SSB is 4,167,000 and *N* for GSF is 4,740,000. Analytical sample includes individuals aged 30 through 61 in the SIPP GSF who applied for Social Security Disability Insurance (SSDI) benefits or never applied for SSDI benefits and had nonmissing 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 Detailed Earnings Record (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 nondisabled individuals, defined here as persons in the SIPP GSF who never applied for SSDI and have nonmissing 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 at 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 2.2.6 for additional details. From 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.

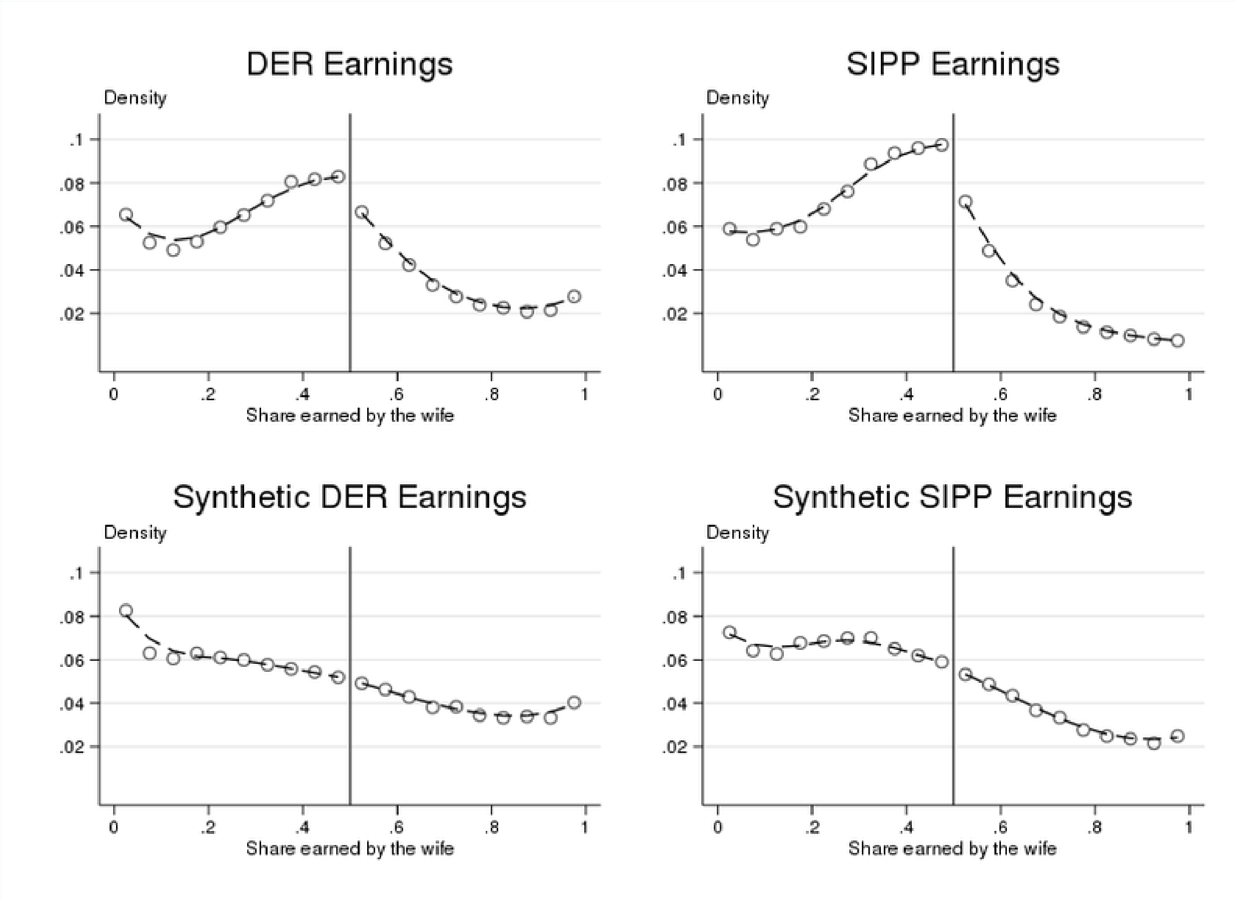

Figure A17. Distributions of wife's earnings share across data sets and sources. The figure plots the share of household income that is earned by the wife. Four different versions of the regression are estimated based on the different earnings measure (Survey of Income and Program Participation [SIPP] and Detailed Earnings Record [DER]) and data source (GSF or 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 2.2.8. From 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.

Table A1. Moments for the standard deviation of within-person earnings.

	(1)	(2)	(3)	(4)	(5)	(6)
	Mean	**P10**	**P25**	**P50**	**P75**	**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

Note. 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 (Survey of Income and Program Participation) panel was excluded. There were 390,000 individuals in the SIPP Synthetic Beta (SSB) sample and 378,000 individuals in the Gold Standard File (GSF) sample. Additional details in Section 2.2.1. From CBDRB-FY23-CED009-0001.

Table A2. Predictors of missing earnings.

	(1)	(2)	(3)	(4)
	GSF		SSB	
	Missing SIPP Earnings	Missing DER Earnings	Missing SIPP Earnings	Missing DER Earnings
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

Note. 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 (Survey of Income and Program Participation) columns is all person-month observations for individuals age 15 and older. The sample for the Detailed Earnings Record (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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.1. From 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.

Table A3. Mincer model regression results.

	(1)	**(2)**	**(3)**	**(4)**
	GSF		**SSB**	
	Log SIPP Earnings	**Log DER Earnings**	**Log SIPP Earnings**	**Log DER Earnings**
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

Note. 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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.2. DER = Detailed Earnings Record. From U.S.

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

Table A4. 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

Note. 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 & 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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.4. OLS = ordinary least squares; DER = Detailed Earnings Record. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195, CBDRB-FY23-CED009-0001.

Table A5. Kejriwal et al. (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

Note. 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. Two-stage least squares (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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.4. OLS = ordinary least squares. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Table A6. Kejriwal et al. (2020) Table 4 replication.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	IFE	**IFE**	**CCEP**	**CCEP**	**CCEP-2**	**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

Note. 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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.4. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Table A7. Kejriwal et al. (2020) Table 5 replication.

	(1)	(2)	(3)	(4)
	OLSMG	**IFEMG**	**CCEMG**	**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

Note. 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 (ordinary least squares), IFE, CCE, and CCE-2 estimators (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). 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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.4. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Table A8. Vietnam War draft lottery and civilian earnings.

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

Note. 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 Social Security Administration's Summary Earnings Record (SER) from 1960 to 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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.5. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY23-CED009-0001.

Table A9. Estimated effect of SSDI benefits receipt on likelihood of positive earnings.

	(1)	(2)
	GSF	**SSB**
SSDI Benefits x Post Disability Onset	−0.2822*** (0.004223)	−0.02096*** (0.004908)

Observations	917,000	1,372,000

Note. Analytical sample includes individuals aged 30 through 61 in the SIPP (Survey of Income and Program Participation) GSF who applied for Social Security Disability Insurance (SSDI) benefits. Further, the 1984 panel was dropped in the interest of having sufficient pre-SIPP Detailed Earnings Record (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: 1%***, 5%**, and 10%*. Additional details in Section 2.2.6. From 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.

Table A10. 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	33000	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	29000	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

Note. 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 main text, further limited to teens ages 16–19 without any missing covariates. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: 1%***, 5%**, and 10%*. Additional details in Section 2.2.7. DER = Detailed Earnings Record. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY21-195.

Table A11. 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

Note. 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 Detailed Earnings Record (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 full details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: 1%***, 5%**, and 10%*. Additional details in Section 2.2.7. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY20-CED001-B0003.

Table A12. 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

Note. 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 full details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: 1%***, 5%**, and 10%*. Additional details in Section 2.2.7. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY201-CED002-B0003.

Table A13. Hampton and Totty (2021) Table 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

Note. 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 full details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: 1%***, 5%**, and 10%*. Additional details in Section 2.2.7. FRA = Full Retirement Age. From U.S. Census Bureau Gold Standard File (GSF) and SIPP (Survey of Income and Program Participation) Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY20-CED001-B0003

Footnotes

1. As an example, according to the Pew Research Center, 61% of surveyed individuals who were not sure that they would respond to the 2020 Census cited “the census asks for too much personal information” as a reason that they may not respond (Cohn et al., 2020). This was the most common reason, exceeding “don't trust the government to use the information properly,” “don't know enough about it,” and “it will take up too much time.” ↩
2. The idea of using model-based imputation as a form of privacy protection and to create synthetic microdata originates from Little (1993) and Rubin (1993). ↩
3. 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. ↩
4. 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 data set (Federal Committee on Statistical Methodology, 2005). ↩
5. The SSB does not satisfy formal privacy, including the most well-known variant—differential privacy (Dwork et al., 2006). ↩
6. Approval only depends upon a user’s research plan being feasible with the SSB/GSF given the existing variables in the data and the existing disclosure avoidance rules. Approval is usually granted within 1 to 2 business days unless an issue is spotted with the application. Historically, approved external researchers had

access to the synthetic data through a synthetic data server (SDS) hosted by Cornell University. At the time of writing this article, the Cornell SDS was recently decommissioned because the external funding for the server ended. The Census Bureau is actively working on alternative ways to provide access to the SSB. ↩

7. Among the few examples is Benedetto et al. (2010), who studied the effects of graduating during a recession on earnings and reported results from both the GSF and SSB. Kinney et al. (2011, 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. ↩

8. 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 to make clear that all such statements about replication or similarity/differences are based on sign and statistical significance comparisons and/or directly comparing magnitudes rather than formal statistical testing. ↩

9. The statistics in this paragraph were cleared for release: CBDRB-FY21-285. ↩

10. The statistics in this paragraph were cleared for release: CBDRB-FY21-285. ↩

11. 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 & DiNardo, 2002). This term refers to the rise in the use of technology that generally complements 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. ↩

12. 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. ↩

13. 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. ↩

14. 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 is typically included in 'Mincer regressions' in either a quadratic or quartic functional form (Murphy & Welch, 1990). ↩

15. 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. ↩
16. Tables A5–A7 in our article replicate the main results in Tables 3–5 of Kejriwal et al. (2020), respectively. ↩
17. There were also lotteries in 1973–1975, but nobody was actually drafted after 1972. ↩
18. The first draft, occurring in 1970, also included individuals who turned 21–26 in that year. ↩
19. 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. ↩
20. Teenagers and restaurant workers are often the focus of minimum wage studies on employment (Allegretto et al., 2011, 2017; Dube et al., 2010; Neumark et al., 2014b, a). ↩
21. SSB users were allowed to merge external data to the SSB and GSF, as long as it was publicly available. ↩
22. 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. ↩
23. We cite and replicate their working paper here rather than the published version of the paper referenced in Section 1 (Hampton & Totty, 2023) because we replicated the results in the working paper before it was published. The published version contains most of the results that we included from the working paper, but some of the results were removed in the published version of the paper. ↩
24. 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. ↩
25. Full-time and part-time employment are based on the amount of a person's earnings in a given year relative to their lifetime highest amount. 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. ↩

26. 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. ↩
27. Notably, the relationship between reaching FRA and claiming does not depend on the merged external minimum wage data. One other result of note from Tables A11–A13 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 A1, which shows a larger density of individuals with low earnings in the SSB than in the GSF. ↩
28. Much of the difference in accuracy between the descriptive and model-based results is driven by one particular model-based use case: The Social Security Disability Insurance event study from Section 2.2.6 and Figure A16 of the Appendix. Excluding those results reduces the median absolute relative error for our model-based results from 0.24 to 0.12. ↩
29. Here we refer to synthetic data *bias*, as opposed to *uncertainty* due to variance or noise, as the tendency of a statistic from the synthetic data to consistently differ from that of the original data in the same direction across multiple implicates of synthetic data, no matter the number of implicates. Confidence interval construction methods such as those from Reiter (2004) account for synthesis uncertainty but not bias. The failure of multiple imputation methods to provide valid confidence intervals due to bias introduced by the synthetic data models is related to the concept of 'congeniality' between the synthesis/imputation model and the researcher's model (Xie & Meng, 2017). ↩
30. The median number of observations (rounded according to Census Bureau rounding rules) across the set of model-based results summarized in Table 1 is 382,000. The mean is 3,654,449. ↩
31. 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). ↩
32. For example, the annual SIPP earnings used in our present article relied on having nonmissing monthly earnings values for all 12 months of the calendar year, and having a nonpositive annual value (i.e., summing all 12 nonmissing 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 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. ↩

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. https://doi.org/10.1162/REST_a_00352

  ↩
- Abowd, J. M., Benedetto, G., Garfinkel, S. L., Dahl, S. A., Dajani, A. N., Graham, M., Hawes, M., Karwa, V., Kifer, D., Kim, H., Leclerc, P., Machanvajjhala, A., Reiter, J. P., Rodriquez, R., Schmutte, I. M., Sexton, W.N., Singer, P. E., & Villhuber, L. (2020). *The modernization of statistical disclosure limitation at the U.S. Census Bureau*. U.S. Census Bureau Working Paper. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0019793917692788

  ↩
- Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security. *American Economic Review*, *80*(3), 313–336. https://www.jstor.org/stable/2006669

  ↩

- Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics, 106*(4), 979–1014. https://doi.org/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. https://doi.org/10.1016/j.econedurev.2019.02.003

  ↩
- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279. https://doi.org/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. https://doi.org/10.1257/app.6.2.178

  ↩
- Bee, A., & Mitchell, J. (2017). Do older Americans have more income than we think? In *Proceedings of the Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association, 110*, 1–85. https://www.jstor.org/stable/26794437

  ↩
- Benedetto, G., Stanley, J. C., & Totty, E. (2018). *The creation and use of the SIPP Synthetic Beta v7.0*. U.S. Census Bureau. https://www2.census.gov/adrm/CED/Papers/CY18/2018-11-BenedettoStanleyTotty-Creation%20SIPP.pdf

  ↩
- Bertrand, M., Kamenica, E., & Pan, J. (2015). Gender identity and relative income within households. *The Quarterly Journal of Economics, 130*(2), 571–614. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/01621459.1995.10476536

  ↩

- Card, D. (2003). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127–1160. https://doi.org/10.1111/1468-0262.00237

  ↩

- 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. https://doi.org/10.1162/00335530151144140

  ↩

- Carr, M. D., & Wiemers, E. E. (2022). The decline in long-term earnings mobility in the US: Evidence from survey-linked administrative data. *Labour Economics, 78*, Article 102170. https://doi.org/10.1016/j.labeco.2022.102170

  ↩

- Carr, M. D., Moffitt, R. A., & Wiemers, E. E. (2022). Reconciling trends in male earnings volatility: Evidence from the SIPP survey and administrative data. *Journal of Business & Economic Statistics*, *41*(1), 26–32. https://doi.org/10.1080/07350015.2022.2126845

  ↩

- Carr, M., Wiemers, E., & Moffitt, R. A. (2023). *Using synthetic data to estimate earnings dynamics: Evidence from the SIPP GSF and SIPP SSB*. SSRN. https://dx.doi.org/10.2139/ssrn.4496224

  ↩

- 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. https://doi.org/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. https://doi.org/10.3368/jhr.XXXVIII.3.618

  ↩

- Cohn, D., Brown, A. & Keeter, S., (2020). *Most adults aware of 2020 Census and ready to respond, but don't know key details*. Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/wp-content/uploads/sites/3/2020/02/PSDT_02.20.20_census_full_report.pdf

  ↩

- 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. 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. 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. https://doi.org/10.48550/arXiv.2304.02107

↩

- Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis, 55*(12), 3232–3243. https://doi.org/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. https://doi.org/10.1162/REST_a_00039

↩

- Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica, 45*(1), 1–22. https://doi.org/10.2307/1913285

↩

- Hampton, M., & Totty, E. (2021). *Minimum wages, retirement timing, and labor supply*. U.S. Census Bureau working paper. https://doi.org/10.13140/RG.2.2.32952.67840

↩

- Hampton, M., & Totty, E. (2023). Minimum wages, retirement timing, and labor supply. *Journal of Public Economics, 224*, Article 104924. https://doi.org/10.1016/j.jpubeco.2023.104924

↩

- 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.), *Handbook of economics of education* (Vol. 1, pp. 307–458). Elsevier. https://doi.org/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. https://doi.org/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. https://doi.org/10.1257/aer.p20161120

↩

- Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*(3), 1–9. https://doi.org/10.1198/000313006X124640

↩

- 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. https://doi.org/10.1002/jae.2759

↩

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

↩

- 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. https://doi.org/10.1257/app.20170478

↩

- Meyer, B. D., & Mittag, N. (2021). An empirical total survey error decomposition using data combination. *Journal of Econometrics, 224*(2), 286–305. https://doi.org/10.1016/j.jeconom.2020.03.026

↩

- Meyer, B. D., Mittag, N., & George, R. M. (2022). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *Journal of Human Resources*, *57*(5), 1605–1644. https://doi.org/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. https://doi.org/10.1257/jep.29.4.199

↩

- Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1086/298220

  ↩

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

  ↩

- Neumark, D., Salas, J. M., & Wascher, W. (2014b). Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater? *ILR Review*, *67*(3), 608–648. https://doi.org/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. https://doi.org/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. https://doi.org/10.1111/j.1468-0262.2006.00692.x

  ↩

- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, *30*(2), 235–242. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7755-eng.pdf?st=8NPjUqX9

  ↩

- Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, *21*(3), 441–462. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-

↩

- Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, *28*(4), 583–590. https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/inferentially-valid-partially-synthetic-data-generating-from-posterior-predictive-distributions-not-necessary.pdf

  ↩

- Totty, E. (2017). The effect of minimum wages and employment: A factor model approach. *Economic Inquiry*, *55*(4), 1712–1737. https://doi.org/10.1111/ecin.12472

  ↩