# Synthetic Data and Social Science Research: Accuracy Assessments and Practical Considerations from the SIPP Synthetic Beta*

Jordan Stanley, U.S. Census Bureau

Evan Totty, U.S. Census Bureau

Synthetic microdata – data retaining the structure of original microdata while replacing original values with modeled values for the sake of privacy – presents an opportunity to increase access to useful microdata for data users while meeting the privacy and confidentiality requirements for data providers. Synthetic data could be sufficient for many purposes, but lingering accuracy concerns could be addressed with a validation system through which the data providers run the external researcher's code on the internal data and share cleared output with the researcher. The U.S. Census Bureau has experience running such systems. In this chapter, we first describe the role of synthetic data within a tiered data access system and the importance of synthetic data accuracy in achieving a viable synthetic data product. Next, we review results from a recent set of empirical analyses we conducted to assess accuracy in the Survey of Income & Program Participation (SIPP) Synthetic Beta (SSB), a Census Bureau product that made linked survey-administrative data publicly available. Given this analysis and our experience working on the SSB project, we conclude with thoughts and questions regarding future implementations of synthetic data with validation.

* This chapter is based in part on our presentation and paper from the 2023 NBER conference on Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences. We thank the organizers and attendees for their feedback and discussion. We would also like to thank Gary Benedetto and Caleb Floyd for their comments on this chapter. 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).

# I. INTRODUCTION

Researchers and statistical agencies are currently facing challenges related to data dissemination within the modern data landscape. The age of "big data" has provided new opportunities and potential benefits while raising concerns and potential costs. For researchers, administrative data is used for a growing share of empirical work (Currie et al., 2020; Einav and Levin, 2014a; Nagaraj and Tranchero, 2023) and is generally seen as higher quality than traditional survey data (e.g., Figlio et al., 2016; Jarmin, 2019; Meyer et al., 2022). However, access to administrative data is often restricted, which has raised concerns related to equitable data access and scientific progress (Card et al., 2010; Cole et al., 2020; Equitable Data Working Group, 2022; Nagaraj and Tranchero, 2023). For statistical agencies and other data providers, administrative data provide opportunities for new and improved data products based on combining administrative and survey data (National Academies of Sciences, Engineering, and Medicine, 2023a; 2023b). At the same time, increases in the breadth and depth of data combined with advancements in computing resources have presented new challenges for protecting the privacy and confidentiality of data subjects as required by law (National Academies of Sciences, Engineering, and Medicine, 2024b).[1]

Tiered data access can help address these hurdles. A tiered approach could provide new intermediate microdata access mechanisms falling between publicly available survey microdata and restricted-access administrative microdata (Abraham, 2019; National Academies of Sciences, Engineering, and Medicine, 2024a; 2024b). Restricted access microdata is only available in data enclaves such as the Federal Statistical Research Data Centers (FSRDCs). FSRDCs provide access to restricted data but come with non-trivial barriers to access as well as limits to what research output can be released publicly. Publicly available microdata comes with the fewest barriers but provides relatively low-quality data due to the substantial privacy protections necessary to protect the confidentiality of data subjects.[2]

One intermediate data access mechanism that could exist within a tiered system is synthesis with validation and/or verification (Benedetto et al., 2024). Synthesis replaces observed microdata values with modeled values based on models that are trained on the original microdata (Dreschler

[1] See Keller et al. (2016) and National Academies of Sciences, Engineering, and Medicine (2017) for discussion of the privacy laws impacting statistical agencies.

[2] Publicly available microdata is intentionally perturbed in numerous ways to protect the confidentiality of respondents so that the data can be made public. Examples include top- and bottom-coding, coarsening, rounding, suppression, sub-sampling, swapping, synthesis, and noise injection.

and Haensch, 2024). The synthetic microdata is made publicly available, and researchers can then have their code "validated" on the original data by the data provider. The validation involves running the user's code on the internal confidential data and sending the cleared results to the user. Synthetic microdata has stronger privacy protection than traditional publicly available microdata and users do not interact directly with the internal microdata. Because of this, the data provider may be able to provide more detail on the synthetic file *and the internal file used for validation* than on a traditional publicly available file. Therefore, while synthetic data with validation/verification carries more barriers to access than traditional publicly available files (e.g., the user must apply for access to the validation system, meet coding standards required for validation, await validation results, etc.), the validation/verification mechanism indirectly provides access to higher-quality data (e.g., the internal file used for validation can avoid many or all of the intentional perturbations applied to non-synthetic publicly available microdata files).

There are many open questions regarding the use of synthetic data and validation for social science research. Among them are (1) to what extent synthetic data and validation impact the development and execution of research ideas and (2) whether data providers can expand synthetic data and validation to data products with a large user base. The answer to both questions likely depends in part on the accuracy of the synthetic data. Greater similarity between statistical results generated on the original versus synthetic data would mitigate the impact of such a system on the development of research ideas by allowing more exploratory analysis and iterative methodology development based on the synthetic data alone. Greater similarity could also lower the demand for validations, which would further reduce the hurdles associated with adopting synthetic data for scientific research (e.g., by avoiding coding standards, validation queues, disclosure reviews, etc. for researchers) and would mitigate challenges of expanding such a system to data products with a large user base (e.g., by reducing the resources and staff needed for the data provider to perform validation analyses and reviewing output).[3]

There are few implementations of such a tiered access system, but the Census Bureau has experience in this area. The Survey of Income and Program Participation (SIPP) Synthetic Beta

[3] Note that there is a tradeoff between accuracy and privacy even with synthetic data: too much similarity in statistical results between the original and synthetic data could mean that the synthetic data are simply reproducing many of the original records and/or that inference attacks could still be successful. Balancing this tradeoff is a choice and requires careful evaluation. However, in our experience, synthetic data that is "too accurate" for surveys with hundreds of variables and countless use cases is not yet a binding constraint.

(SSB) is a publicly available synthetic dataset modeled on the SIPP Gold Standard File – an internal data product that links panels of the SIPP to data from the Internal Revenue Service and the Social Security Administration. External users could apply for access, develop and run their analysis on the SSB, and then submit their code as part of validation requests. Internal Census staff would run the validation, conduct a disclosure review of the statistical output, and, if appropriate, release the cleared output to the external researcher.

Given our experience using the SIPP Gold Standard File for empirical research and our past role as support staff for SSB validations, we recently undertook a series of analyses to assess accuracy in the SSB by comparing estimates derived from the GSF to those derived from the SSB (Stanley and Totty, 2024). We performed assorted econometric analyses and generated various sets of descriptive statistics to examine how the SSB performed under different empirical requests. In this chapter, we first summarize the results from our tests of SSB accuracy relative to the GSF. We then discuss the implications of the results for the feasibility of synthetic data with validation from the perspective of its impact on the scientific research process and its applicability to data products with a large user base. We also provide some additional practical considerations of developing, managing, and using such a system.

Overall, we found that the SSB performs quite well at replicating statistical results from the GSF. In the cases we tested, the SSB either appeared capable of standing alone without the need for validation or had shortcomings that were sensible given the SSB design and modeling. We believe the SSB shows the potential for synthetic data to achieve a level of accuracy that could ease its impact on the development and execution of research ideas and help reduce the challenge of expanding the system to a large user base. However, the specific design and management of the SSB would be difficult to expand to a much larger user base. We discuss some of the reasons why and possible solutions later in this chapter.

We will proceed with a brief description of the SSB data and validation system followed by an overview of our empirical results. We conclude with a discussion of the implications of synthetic data with validation for scientific research and considerations for future synthetic data applications.

## II. DATA AND METHODOLOGY

The main datasets we use are the SSB and the SIPP GSF (the internal reference file for the SSB).[4] The GSF consists of multiple panels of the SIPP (in the most recent version of the data, the panels were 1984, 1990 – 1993, 1996, 2004, and 2008). Additional variables from Internal Revenue Service (IRS) and Social Security Administration (SSA) data were added for SIPP respondents who could be linked to the administrative data sources. The SSB was created from the GSF using sequential regression multivariate imputation (SRMI). This methodology uses regression analysis to replace observed data values with modeled values.

The first version of the SSB was created in 2003. The most recent version of the SSB is version 7, which was released in 2018. We used version 7 for our analyses in Stanley and Totty (2024). Version 7 of the SSB is fully synthetic, meaning all data values were modeled. Note that only a subset of SIPP variables was included in the SSB. For full background on the SSB, see Benedetto, Stanley, and Totty (2018).

In Stanley and Totty (2024) we produced a series of estimates typical in empirical economics research. Each of the analyses was focused to some degree on earnings – a traditionally important outcome of interest in economics. We generated descriptive statistics, figures, and regression results. As the SSB was intended to support a wide range of unknown use cases, we also tried to cover an array of research topics and empirical methodologies.[5] In each case, we ran the same analysis on the SSB and on the GSF, thus mimicking how the process would work for external data users. We then compared the similarity of the SSB results to the GSF results.

The exact analytical samples for a given set of estimates depend on the analysis being run, but our main sample consists of person-year observations with annual measures of earnings from both the Detailed Earnings Record (DER) and the SIPP. The GSF and SSB have roughly 783,000 persons, and our person-year sample contains 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. The main subsample of interest consists 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. In many of our regression models, the natural logarithm of earnings is the outcome

[4] External researchers can request access to the SSB via a nominal application process, after which the SSB data files are made available for download. Researchers can then build their analysis using the SSB and, if desired, submit their code for validation on the GSF. See https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html for more details.

[5] In a similar paper to ours, Carr et al. (2023) provide a detailed analysis of the accuracy of the SSB for *single topic* (earnings dynamics).

of interest, meaning most of the subsamples are derived from this positive earner sample. Additional subsamples are created when accounting for missing data as well as age ranges of interest for the assorted analyses. See Stanley and Totty (2024) for more details.

## III. SIMILARITY BETWEEN SYNTHETIC AND INTERNAL RESULTS

The full set of results is included in Stanley and Totty (2024); here, we will reproduce some of the statistical summaries from our analysis to aid the discussion. Note that these same summary tables and figures appear in Stanley and Totty (2024).

Figure 1 is a scatter plot of all the GSF versus SSB results separated by descriptive versus model-based results. The SSB does not perfectly replicate all the results from the GSF, but the bunching around the 45-degree line signifies a high correlation between results derived from the GSF and those derived from the SSB – especially for descriptive estimates. The descriptive and model-based results can be further divided by type of estimate. In Figure 1, we see that the descriptive results show a steeper slope of association relative to the 45-degree line. When we group the descriptive results by type of statistic, medians showed the strongest performance by the SSB in replicating the GSF estimates, while the other sorts of descriptive estimates still showed a strong positive association.[6]

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 2 groups the model-based results by different types of models. The worst-performing types of models in our exercises were (1) the models incorporating merged external data that were not part of the synthesis process and (2) so-called "within-person" models that include individual-level fixed effects, interactive individual- and time-level fixed effects, or individual-level hazard models. The first type illustrated virtually no association between the SSB and GSF results, while the second type illustrated a significantly weakened association. The remaining regression results based on other models illustrate a very

[6] See Figure 2 in Stanley and Totty (2024) for the scatterplot of descriptive results broken out by type of descriptive statistic.

strong association between the SSB and GSF results (see the right-most panel of Figure 2). These other types of models include ordinary least squares regressions, two-stage least squares regressions, and regression discontinuity designs.

We further assessed the differences in estimate magnitudes by computing the absolute relative error in estimates derived from the SSB compared to their counterparts in the GSF (see Figure 3). As seen qualitatively in Figure 1, the SSB is more similar to the GSF for our descriptive analyses than for our model-based analyses. This is supported quantitatively in Figure 3 as the median absolute relative error is 0.08 for our descriptive results and 0.24 for our model-based results. More than half of our descriptive results derived from the SSB are within 10% of their GSF counterparts. For the model-based SSB results, the majority are within 25% of their GSF counterparts.[7]

Finally, Table 1 summarizes differences in the inferences drawn from the GSF results versus those based on the SSB results. We consider multiple inference benchmarks. First, we compare confidence intervals between the two datasets. As shown in Panel A of Table 1, 95% confidence intervals in the SSB were approximately twice as wide as the GSF confidence intervals on average. The wider confidence intervals are due in part to the confidence interval construction methodology from Reiter (2004) that accounts for the additional uncertainty introduced by the synthesis process. 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. An alternative benchmark is to compare sign and statistical significance between the GSF and SSB. Replicating sign and statistical significance can be useful in multiple ways. For certain analyses, the sign and statistical significance (relative to a null hypothesis of zero effect) of a particular estimate is the key statistical conclusion. Further, with the validation option, accuracy of sign and statistical significance may be sufficient for research goals until a validation is performed to acquire results from the internal reference file. 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

[7] 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 in Stanley and Totty (2024). Excluding those results reduces the median absolute relative error for our model-based results from 0.24 to 0.12.

of statistical conclusions based on hypothesis testing. Relative to the corresponding GSF results, the SSB results produce the same statistical conclusion for approximately 63% of the coefficient estimates, fail to replicate a statistically significant relationship for approximately 33% of estimates, imply a spurious relationship for approximately 2% of estimates, and imply a significant relationship in the opposite direction for approximately 2% of estimates. It is reassuring that spurious and opposite-sign statistical relationships in the SSB (relative to the GSF findings) are exceedingly rare (4% of the time in our models).

In looking across all the estimates in Stanley and Totty (2024), differences between the SSB and GSF are often consistent with interpretable and expected patterns. Statistics that are sensitive to outliers (e.g., means) may be less likely to be replicated in synthetic data than statistics that are not sensitive to outliers (e.g., medians) because synthetic data inherently attempt to mask sensitive values such as outliers. This is observable in Figure 2 from Stanley and Totty (2024) as discussed earlier in this section. The findings from regressions that rely solely on variables already in the data (i.e., variables used in the synthetic data models) could be more replicable than regressions involving variables merged onto synthetic data after synthesis because the latter are not used in the synthetic data modeling process which can thus obfuscate relationships with the synthetic variables. This scenario is seen in our results and observable in Figure 2 for this chapter. Synthetic data modeling decisions may also explain some of the differences we saw comparing GSF results to SSB results. For example, results relying on within-person earnings dynamics tended to yield poorer replications in our analysis. This pattern may be due to a modeling decision – the synthetic data models for the SSB were primarily based on modeling variable *levels* rather than *year-over-year changes*. It is possible that the SSB could have performed better replicating such results if the SSB explicitly modeled within-person changes in earnings over time.

Our results also point to some inherent challenges of synthetic data. The relatively poor performance for the accuracy of modeled statistics compared to descriptive statistics (see Figure 1) and for confidence interval coverage (see Table 1) demonstrate the challenge of generating unbiased synthetic data for covariate relationships. This is especially challenging for data products with many variables and a large number of possible (and unknown) use cases. While synthetic data confidence interval methods such as those from Reiter (2004) can account for the additional *uncertainty* introduced by the modeling process, they cannot account for *bias* that arises due to synthetic data models that fail to account for all possible covariate relationships.

## IV. IMPLICATIONS FOR SCIENTIFIC RESEARCH AND SCALING VALIDATION

Our results demonstrate that the SSB provided a high level of accuracy for descriptive statistics (such as means and medians) and for modeled relationships that are congenial with the synthetic data models.[8] This shows the potential for synthetic data to stand alone for some applications, which would ease the integration of synthetic data with validation into the scientific research process and help agencies expand such systems to large user bases.

Despite this potential, there are several remaining challenges. One challenge is the amount of manual labor that went into validations for the SSB. Code was manually vetted and run by internal staff. Code would often break due to missing dependencies in the researcher's code and/or conflicting software versions. Disclosure avoidance and output review was also done manually by internal staff. Several of these steps are capable of being automated to some degree, particularly given recent advances in artificial intelligence. Examples include using static and dynamic analysis for vetting of code, using containers for packaging and executing all code and data dependencies, and implementing automated disclosure avoidance application and/or review. See Benedetto et al. (2024) for additional discussion related to the practical challenges associated with developing and managing a validation service, including opportunities for automation.

Another challenge is that researchers may be unlikely to trust that the synthetic data are sufficient for their purposes, even when the synthetic data can be expected to provide a high level of accuracy (e.g., for descriptive statistics or relatively simple modeled statistics). Researchers and data providers could therefore benefit from ways to assess the "trustworthiness" of a statistic generated from the synthetic data before requesting a validation. This could be achieved with a verification option, which provides a summary measure of the similarity between the synthetic and internal statistic without releasing the internal statistic. However, verification is still costly with respect to coding standards and validation queue for the researchers as well as resources and privacy leakage for the data provider. Another option would be a theoretical and/or empirical assessment by the researcher that attempts to gauge the trustworthiness of a synthetic statistic without access to the internal data. More research is needed on whether such an option could exist.

[8] Synthetic data models and data user models are said to be "congenial" if they are based on the same assumptions. Congeniality is required for valid statistical inference (Abowd and Schmutte, 2015; Dreschler and Haensch, 2024; Meng, 1994).

At the very least, releasing information about the synthetic data models would allow researchers to assess the congeniality of their planned empirical models with what was done in the synthetic data models.

Finally, there are several interesting opportunities for synthetic data with validation to impact scientific progress besides providing easier access to high quality microdata. To begin with, there are opportunities for indirect effects of such a system on researcher behavior. Such a system could improve reproducibility/replicability, increase the usage of pre-analysis plans, and reduce p-hacking – scientific ideals that have received heightened attention in economics (Brodeur et al., 2016; Brodeur et al., 2020; Brodeur et al., 2023; Coffman and Niederle, 2015; Olken, 2015; Vilhuber et al., 2023; Whited, 2023). Expanding access to administrative data via synthesis and validation would allow for easier reproducibility/replicability by data editors and other researchers. Additionally, developing an analysis plan on synthetic data without yet knowing what results the validation will show is akin in some ways to the use of pre-analysis plans. Furthermore, this uncertainty regarding validation results combined with a limit on the number and/or size of validation requests could also make it difficult for researchers to engage in p-hacking.

There are also opportunities for the research profession (e.g., journals, conferences, etc.) to adjust their processes to better accommodate tiered access projects. For example, conference presentations and even initial journal submissions could allow for synthetic results with the knowledge that validated results will follow. This would be similar to how some journals now allow initial review and in-principle acceptance based solely on pre-analysis plans before the analysis is actually conducted (Arpinon and Espinosa, 2023).[9]

## V. CONCLUSION

Data providers and researchers must contend with many modern challenges. Threats to privacy and confidentiality are increasing in prevalence and complexity. Current dissemination methods may be insufficient, and data providers need to determine new and improved ways to provide useful data and statistics while protecting privacy. Researchers have valid concerns about the

[9] Such submissions are referred to as "registered reports." The first economics journal to incorporate registered reports was the Journal of Development Economics in 2018.

appropriate balance of data privacy and data accuracy (and how that is determined), ramifications for empirical research, and equitable data access.

A tiered access approach such as synthetic data with validation/verification offers one potential solution to these concerns. Synthetic data can offer sufficient privacy protections and democratize data access by making previously restricted data publicly available. A standalone synthetic data product could be sufficiently accurate for many purposes, and lingering accuracy concerns could be addressed through validation and/or verification. As such, this sort of setup could be beneficial to both researchers and data providers.

Pareto improvements of course sound great on paper, but there are several open questions and practical considerations for producing synthetic data and offering a validation and/or verification option. Developing research ideas and analysis plans with synthetic data would be a substantial change for researchers who are accustomed to adjusting their analyses and research questions based on what they discover in the data. Meeting the requirements for coding standards and disclosure avoidance review in order to receive a validation could also prove to be a challenge for some users. Furthermore, user demand for validations could impact how exactly validation/verification services can be offered and how quickly results could be provided if the validation process requires an abundance of manual labor.

Our research endeavor discussed in this chapter focused on the accuracy of empirical results comparing one synthetic product (the SIPP Synthetic Beta) to its corresponding internal reference file (the SIPP GSF). Greater similarity between statistics generated on the synthetic and internal data can provide benefits for agencies providing the product and for researchers using the product. Our overall findings point to an imperfect but strong association between the GSF and SSB results. Specifically, we found that the SSB did a good job replicating many "basic" analyses (e.g., descriptive statistics and simple regression analysis) while struggling with other applications.

Our test cases attempted to cover numerous statistical methods and research topics but obviously only represent a tiny subset of the possible analyses. Furthermore, the GSF is only one dataset and the SSB was created using one particular method for generating synthetic data (SRMI). Finally, our analysis treated all differences between the original and synthetic data as reductions in accuracy, which is an imperfect assumption when the original data already contain errors (Totty and Watson, 2024). For many reasons discussed here and in our companion paper [see Stanley and Totty (2024)], we view our results for accuracy of the SSB to be a floor for what synthetic data

can accomplish. For example, modern synthesis methods like non-parametric CART and machine learning are easier to implement and can generate more accurate synthetic data (Drechsler and Reiter, 2011; Reiter, 2005; Reiter and Kinney, 2012). Recent developments in deep learning, such as Generative Adversarial Networks and Large Language Models, also show promise for delivering high quality synthetic data while requiring minimal human input, although to-date neither clearly surpasses CART performance for moderately-sized sample surveys (Akiya, Ishihara, and Yamamoto, 2024; Lautrup et al., 2024; Miletic and Sariyar, 2025). Nonetheless, the SSB in its original form provides important insights into the opportunities and challenges of synthetic data with validation as a data access tier.

References


Abraham, K. G. (2019). Reconciling data access and privacy: Building a sustainable model for the future. In *AEA Papers and Proceedings* (Vol. 109, pp. 409-413). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Akiya, I., Ishihara, T., Yamamoto, K. (2024). Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study. *JMIR Medical Informatics, June 18*.

Arpinon, T., & Espinosa, R. (2023). A practical guide to registered reports for economists. *Journal of the Economic Science Association*, *9*(1), 90-122.

Benedetto, G., Rodríguez, R. A., Stanley, J. C., & Totty, E.. (2024). Validation Services for Confidential Data. In Jörg Drechsler, Daniel Kifer, Jerome Reiter, Aleksandra Slavković (Eds.), *Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data*. Routledge.

Benedetto, G., Stanley, J. C., & Totty, E. (2018). The Creation and Use of the SIPP Synthetic Beta v7.0. CES Technical Notes Series 18-03, Center for Economic Studies, U.S. Census Bureau. Retrieved from https://ideas.repec.org/p/cen/tnotes/18-03.html.

Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). Unpacking p-hacking and publication bias. *American Economic Review*, 113(11), 2974-3002.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11), 3634-3660.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1), 1-32.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *American economic association, ten years and beyond: Economists answer NSF's call for long-term research agendas*.

Carr, M., Wiemers, E., & Moffitt, R. A. (2023). Using Synthetic Data to Estimate Earnings Dynamics: Evidence from the SIPP GSF and SIPP SSB. *Available at SSRN 4496224*.

Coffman, L. C., & Niederle, M. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives*, 29 (3): 81–98.

Cole, S., Dhaliwal, I., Sautmann, A., & Vilhuber, L. (2020). Handbook on using administrative data for research and evidence-based policy. *https://admindatahandbook. mit. edu/book/v1. 0-rc5/index.html*.

Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings* (Vol. 110, pp. 42-48). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Drechsler, J., & Haensch, A. C. (2024). 30 years of synthetic data. *Statistical Science*, 39(2), 221-242.

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

Einav, L., & Levin, J. (2014a). Economics in the age of big data. *Science*, *346* (6210).

Equitable Data Working Group. (2022). A vision for equitable data: Recommendations from the Equitable Data Working Group. https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf.

Figlio, D., Karbownik, K., & Salvanes, K. G. (2016). Education research and administrative data. In *Handbook of the economics of education* (Vol. 5, pp. 75-138). Elsevier.

Jarmin, R. S. (2019). Evolving measurement for an evolving economy: thoughts on 21st century US economic statistics. *Journal of Economic Perspectives*, *33*(1), 165-184.

Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? A review of the issues. Annual Review of Statistics and Its Application, 3(1), 161-180.

Lautrup, A. D., Hyrup, T., Zimek, A., Schneider-Kamp, P. (2024). Systematic Review of Generative Modelling Tools and Utility Metrics for Fully Synthetic Tabular Data. *ACM Computing Surveys, 57(4)*, 1-38.

Miletic, M., Sariyar, M. (2025). Utility-based Analysis of Statistical Approaches and Deep Learning Models for Synthetic Data Generation With Focus on Correlation Structures: Algorithm Development and Validation. *JMIR AI, March 20*.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, 538-558.

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.

Nagaraj, A., & Tranchero, M. (2023). *How does data access shape science? Evidence from the impact of US census's research data centers on economics research*. National Bureau of Economic Research No. w31372. https://doi.org/ 10.3386/w31372.

National Academies of Sciences, Engineering, and Medicine. (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine. (2023a). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26804.

National Academies of Sciences, Engineering, and Medicine. (2023b). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26688.

National Academies of Sciences, Engineering, and Medicine. (2024a). *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27169.

National Academies of Sciences, Engineering, and Medicine. (2024b). *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27335.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3), 61-80.

Reiter, J. P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology*, 30, 235-242.

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

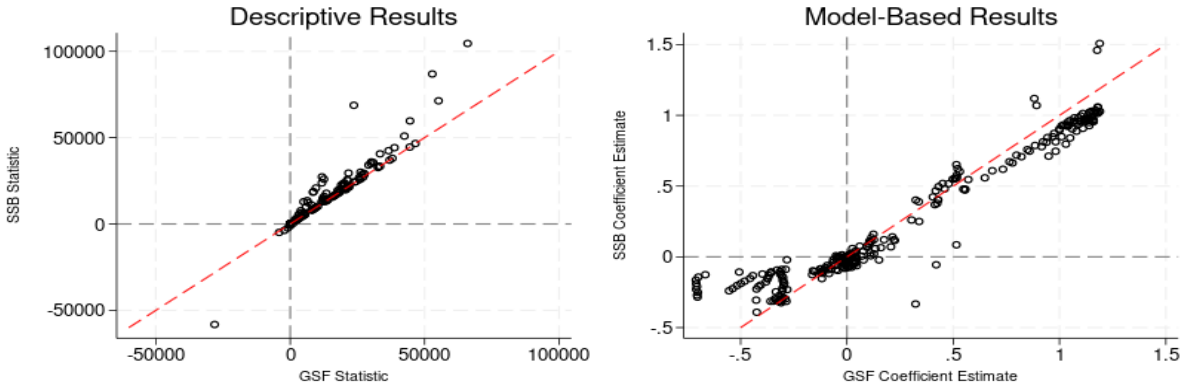
Stanley, J., & Totty, E. (2024). A Penny Synthesized is a Penny Earned? An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta. *Harvard Data Science Review,* forthcoming. https://doi.org/10.1162/99608f92.c168891c.

Totty, E., & Watson, T.. (2024). Total Survey Error and Statistical Disclosure Limitation. Census Bureau Working Paper Number ced-wp-2024-001. https://www.census.gov/library/working-papers/2024/adrm/ced-wp-2024-001.html.

Vilhuber, L., Schmutte, I., Michuda, A., & Connolly, M. (2023). Reinforcing Reproducibility and Replicability: An Introduction. *Harvard Data Science Review*, *5*(3).

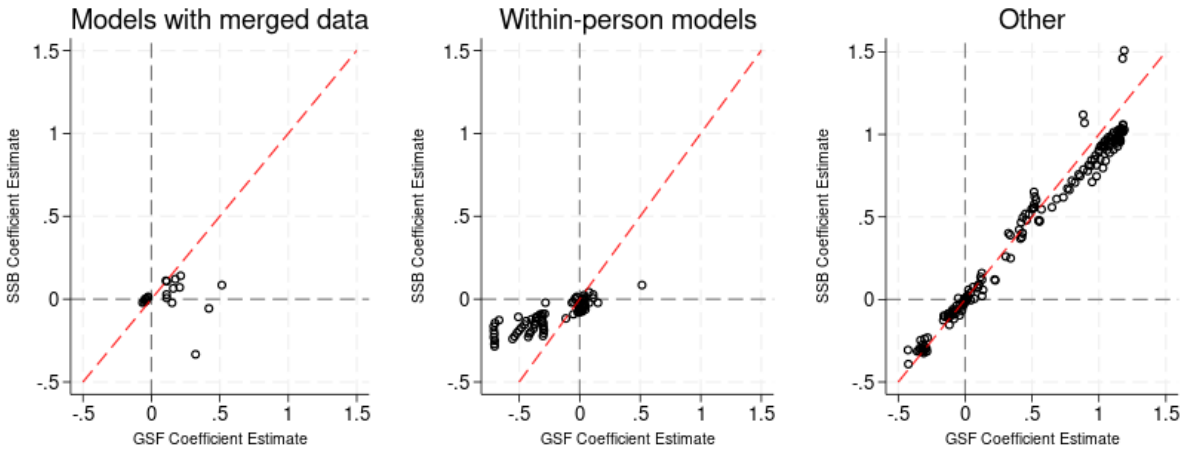
Whited, T. (2023). Costs and Benefits of Reproducibility in Finance and Economics. *Harvard Data Science Review*, *5*(3).

**Figure 1. Scatter Plot of GSF and SSB Results**



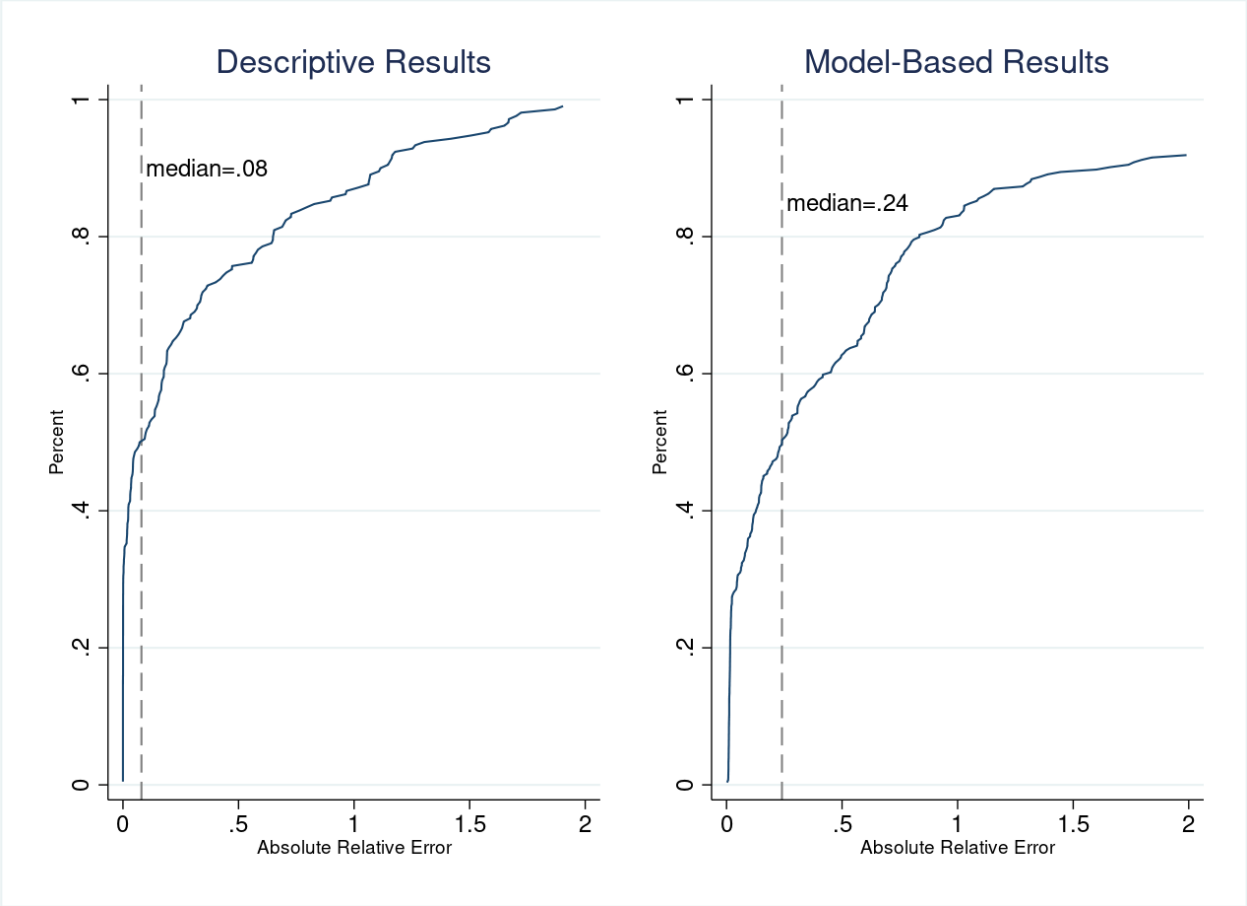

The right figure plots the GSF versus SSB results for the regression-based results shown in the Appendix. The left figure plots the remaining statistics in the paper (e.g., means, medians, ratios, and counts). In each, the X axis is the estimate using the internal GSF, the Y axis is the estimate using the SSB, and the red line is the 45-degree line. See Stanley and Totty (2024) for additional details.
Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001. Figure also appears in Stanley and Totty (2024).

**Figure 2. Scatter Plot 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. See Stanley and Totty (2024) for additional details.
Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001. Figure also appears in Stanley and Totty (2024).

**Figure 3. Distribution of Absolute Relative Errors**




The right figure plots the GSF versus SSB results for the regression-based results. The left figure plots the remaining statistics in the paper (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. See Stanley and Totty (2024) for additional details.
Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001. Figure also appears in Stanley and Totty (2024).

**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% |

The comparison includes all regression-based results except for those that do not report a standard error or confidence interval. See Stanley and Totty (2024) for additional details.
Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001. Table also appears in Stanley and Totty (2024).