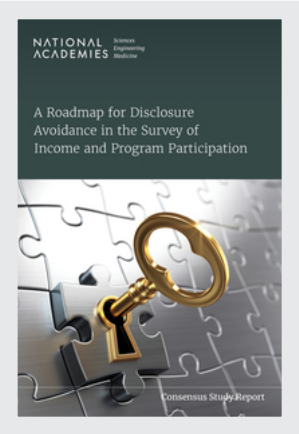
NATIONAL ACADEMIES Sciences Engineering Medicine

NATIONAL ACADEMIES PRESS Washington, DC

This PDF is available at http://nap.nationalacademies.org/27169



BUY THIS BOOK

FIND RELATED TITLES

A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (2023)

DETAILS

330 pages | 6 x 9 | PAPERBACK

ISBN 978-0-309-70710-7 | DOI 10.17226/27169

CONTRIBUTORS

Trivellore Raghunathan and Bradford Chaney, Editors; Panel to Create a Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine. 2023. *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.

Visit the National Academies Press at nap.edu and login or register to get:

- Access to free PDF downloads of thousands of publications
- 10% off the price of print publications
- Email or social media notifications of new titles related to your interests
- Special offers and discounts

All downloadable National Academies titles are free to be used for personal and/or non-commercial academic use. Users may also freely post links to our titles on this website; non-commercial academic users are encouraged to link to the version on this website rather than distribute a downloaded PDF to ensure that all users are accessing the latest authoritative version of the work. All other uses require written permission. (Request Permission)

This PDF is protected by copyright and owned by the National Academy of Sciences; unless otherwise indicated, the National Academy of Sciences retains copyright to all materials in this PDF with all rights reserved.

NATIONAL ACADEMIES *Sciences Engineering Medicine*

NATIONAL ACADEMIES PRESS
Washington, DC

A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

Trivellore Raghunathan and Bradford Chaney, *Editors*

Panel to Create a Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

Committee on National Statistics

Division of Behavioral and Social Science and Education

Consensus Study Report

Prepublication copy, uncorrected proofs.
Copyright National Academy of Sciences. All rights reserved.

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by a contract between the National Academy of Sciences and the U.S. Census Bureau (#1333LB21D0000003/1333LB21F00000248). Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (No. 1560294) and several individual contracts. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-XXXXX-X
International Standard Book Number-10: 0-309-XXXXX-X
Digital Object Identifier: https://doi.org/10.17226/27169

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Copyright 2023 by the National Academy of Sciences. National Academies of Sciences, Engineering, and Medicine and National Academies Press and the graphical logos for each are all trademarks of the National Academy of Sciences. All rights reserved.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2023. *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.

Copyright National Academy of Sciences. All rights reserved.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

Copyright National Academy of Sciences. All rights reserved.

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

Copyright National Academy of Sciences. All rights reserved.

PANEL TO CREATE A ROADMAP FOR DISCLOSURE AVOIDANCE IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

TRIVELLORE RAGHUNATHAN (*Chair*), University of Michigan
SCOTT H. HOLAN, University of Missouri
V. JOSEPH HOTZ, Duke University
THOMAS KRENZKE, Westat
FANG LIU, University of Notre Dame
ROBERT A. MOFFITT, Johns Hopkins University
AMY PIENTA, Inter-university Consortium for Political and Social Research
NATALIE SHLOMO, University of Manchester
ALEKSANDRA (SEŠA) SLAVKOVIĆ, Pennsylvania State University
HEEJU SOHN, Emory University
SALIL VADHAN, Harvard School of Engineering and Applied Sciences
JENNIFER VAN HOOK, Pennsylvania State University

BRADFORD CHANEY, *Study Director*
DAVID JOHNSON, *Senior Program Officer*
NANCY KIRKENDALL, *Senior Program Officer*
MADELEINE GOEDICKE, *Senior Program Assistant*
JOSHUA LANG, *Senior Program Assistant*

Copyright National Academy of Sciences. All rights reserved.

COMMITTEE ON NATIONAL STATISTICS

KATHARINE ABRAHAM *(Chair)*, Department of Economics, University of Maryland, College Park
MICK P. COUPER, Institute for Social Research, University of Michigan
DIANA FARRELL, JPMorgan Chase Institute, Washington, DC
ROBERT GOERGE, Chapin Hall at the University of Chicago
ERICA L. GROSHEN, School of Industrial and Labor Relations, Cornell University
DANIEL E. HO, Stanford Law School, Stanford University
HILARY HOYNES, Goldman School of Public Policy, University of California-Berkeley
DANIEL KIFER, Department of Computer Science and Engineering, The Pennsylvania State University
SHARON LOHR, School of Mathematical and Statistical Sciences, Arizona State University, Emerita
NELA RICHARDSON, ADP Research Institute, Roseland, NJ
C. MATTHEW SNIPP, School of the Humanities and Sciences, Stanford University
ELIZABETH A. STUART, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

MELISSA CHIU, *Director*
BRIAN HARRIS-KOJETIN, *Senior Scholar*
CONSTANCE F. CITRO, *Senior Scholar*

Prepublication copy, uncorrected proofs

Copyright National Academy of Sciences. All rights reserved.

Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

CLAIRE MCKAY BOWEN, Urban Institute
SAKI KINNEY, RTI International
DAVID VAN RIPER, University of Minnesota
KENNETH W. WACHTER, University of California, Berkeley
LANCE A. WALLER, Emory University
EMILY E. WIEMERS, Syracuse University

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by **John L. Czajka**, Independent consultant, Bethesda, MD, and **William W. Stead**, Vanderbilt University Medical Center. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Copyright National Academy of Sciences. All rights reserved.

Acknowledgments

This Consensus Study Report reflects the invaluable contributions of many colleagues, whom we thank for their generous time, effort, and expert guidance. On behalf of the panel, I extend our deepest appreciation to the sponsor of this work: the Census Bureau within the U.S. Department of Commerce. Without the Census Bureau's support, including through briefings and responses to our information requests, this study would not have been completed. In particular, we thank David Waddington, Division Chief of the Social, Economic, and Housing Statistics Division; Jason Fields, Senior Researcher for Demographic Programs and the SIPP; and Holly Fee of the Social, Economic, and Housing Statistics Division. We also thank all of those who provided briefings on key issues to the panel; these include Gary Benedetto, Steve Clark, Aref Dajani, Holly Fee, Jason Fields, Benjamin Gurrentz, Adriana Hernández-Viver, Yerís H. Mayol-García, Robert Munk, Rolando Rodriguez, Rachel Shattuck, Phyllis Singer, Jordan Stanley, Sam Szelepka, Evan Totty, and Ashley Westra, all of the Census Bureau; Jerry Reiter, Duke University, dana boyd, Microsoft Research and Georgetown University; and Lars Vilhuber, Cornell University.

The committee also extends its gratitude to members of the staff of the National Academies for their significant contributions to this report. Kirsten Sampson Snyder and Bea Porter masterfully shepherded the report through the review and production process, and Marc DeFrancis provided useful editorial advice that streamlined the report. Joshua Lang and Madeleine Goedicke provided administrative and logistical support for numerous panel meetings.

Brian Harris-Kojetin, senior scholar and former director of the Committee on National Statistics, and Melissa Chiu, current director of the Committee on National Statistics, had key roles in the original study design and selection and recruitment of the study panel, along with ongoing support of the panel and the preparation of the report. Bradford Chaney, study director and senior program officer, assisted in leading the panel and acquiring needed resources. Nancy

Copyright National Academy of Sciences. All rights reserved.

Kirkendall and David Johnson, both senior program officers, provided valuable assistance based on their past experience with SIPP and the Census Bureau.

To my colleagues on the panel, I appreciate your diligence and expertise in examining the difficult issues raised in this study, and your spirit of cooperation in coming together to reach a consensus. Your shared wisdom from across a wide range of expertise areas, team spirit, and generosity of time brought innovative ideas to the discussions and produced this report. It was a great pleasure to work with you all. Thank you.

Trivellore Raghunathan, *Chair*

Consensus Panel on a Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

Copyright National Academy of Sciences. All rights reserved.

Contents

Acronyms **xvi**

Glossary **xvii**

Summary **1**

1 Introduction **16**

Statement of Task 20
How the Panel Did Its Work 20
Outline of the Report 25
Previous CNSTAT Studies Relating to SIPP 26
An Interim Assessment of SIPP (1989) 26
The Future of the Survey of Income and Program Participation (1993) 27
Reengineering the Survey of Income and Program Participation (2009) 28
The 2014 Redesign of the Survey of Income and Program Participation: An Assessment (2018) 29

2 Overview: SIPP and Its Uses **31**

Formative and Early Years of SIPP 31
What Makes Sipp Unique and Valuable 33
Sampling Design and Methodology for SIPP 35
The Content Collected Through SIPP 37
Topics Covered 38
Selected Studies Show the Range and Depth of SIPP 42
Data Products from SIPP 46
Census Bureau Reports 46
Microdata for Data Analysis 47
disclosure Avoidance in SIPP from the 1990s to the Present 51
Administrative Data Added to SIPP 53
How SIPP Data Are Used 55
Topics Covered in Research Using SIPP 55
Conclusions 56

3 Measuring of Disclosure Risk and Ways of Assessing It **58**

Types of Disclosure Risk 58
Measures of Disclosure Risk 60
Differential Privacy and Disclosure Risk 61
Comparing the Measures of Disclosure Risk 63

Copyright National Academy of Sciences. All rights reserved.

Assessing Disclosure Risk Through Analytic or Empirical Approaches 65
Quantifying Risks of Re-identification 66
Examining Disclosure Risks Faced by SIPP 70
The Level of Detail in SIPP Presents a Challenge 71
Choosing Which External Databases to Compare With SIPP 76
The Census Bureau's Measurement of Risk Through a Re-identification Study 78
Recommendations to Census for On-Going Disclosure Risk Assessment Strategy for the SIPP 80
Limitations of the Census Bureau's Re-Identification Study 81
Longitudinal Data Increases the Risk of Disclosure 86
Making Disclosure Analysis an Ongoing Activity 87
Conclusions and Recommendations 88

4 Overview of Disclosure Limitation Approaches 92

Protecting Confidentiality Through Offering Multiple Modes of Access 92
Methods of Disclosure Limitation Involving Changes to the File 96
Designing Disclosure Avoidance to Maximize Usability 102
Conclusions and Recommendations 104

5 Disclosure Limitation Approaches: Secure Online Data Access (SODA) 106

Privacy Protections in the SODA 108
Enhancing Equity in the SODA 109
SODA Capabilities 110
Requirements for Organizations Managing a Census Bureau SODA 111
Conclusions and Recommendations 113

6 Disclosure Limitation Approaches: Synthetic Data 114

Overview of Data Synthesis 114
Data Synthesis Approaches 118
Disclosure Risk 120
Statistical Validity 125
Statistical Evaluation of Data Utility 126
Validation and Verification Systems 128
SIPP Synthetic Beta (SIPP SSB) 131
Challenges in SIPP Data Synthesis and Utility Evaluation 133
Conclusions and Recommendations 135

7 Disclosure Limitation Approaches: Flexible Table Generator and Remote Analysis Platforms 138

Introduction 138
Table Generator Platform 139
Conclusion and Recommendation 141

Copyright National Academy of Sciences. All rights reserved.

8 Disclosure Limitation Approaches: Geography Variables **142**

Use 1: Identifying Specific Geographies Is Sometimes Unnecessary 143
Use 2: Making Subnational Estimates 144
Spatio-temporal Models 146
Modelling at the Unit Level 147
Incorporating Spatial Dependence and Correlated Random Effects 149
Leveraging Auxiliary Data and Applying Disclosure Avoidance 150
Use 3: Identifying Specific States or Localities 151

9 Maintaining Usability While Preserving Confidentiality: Potential Strategies **153**

Accuracy 154
Feasibility 156
Use 1: Analyses Relying on Unique SIPP Content 158
Use 2: Longitudinal Analysis 161
Use 3: Analyses Relying on Granular Data and Complex Recoded Data 162
Use 4: Causal Effects of Public Policies 162
Use 5: Analyses Relying on Administrative Record Linkages 163
Accessibility 168
How the Introduction of SODA and Reinterpretation of Title 13 Would Affect Accessibility 171
Title 13 and Data Access 175
Conclusions and Recommendations 181

10 Conclusions and Recommendations **184**

SIPP and the Risk of Disclosure Avoidance 184
SIPP Use and Usability 185
Current SIPP Approaches to Disclosure Avoidance 188
Tiers of Access as a Tool for Protecting Privacy 189
Additional Tools to Support Disclosure Avoidance 191
Title 13 and Its Requirements for Disclosure Avoidance 194
The Benefits of Enhanced Communication 195
Linking SIPP Data With Alternative Data Sources 199
Future Research to Address Current Knowledge Gaps 200
Census Bureau Resources for Addressing Disclosure Avoidance 201
A Roadmap for Addressing Disclosure Avoidance in SIPP 202

References **206**

Appendix A Technical Details on Measuring Disclosure Risk **223**

Appendix B Inferences Based on Multiple Synthetic Data **225**

Appendix C Technical Details for Differential Privacy Table Builder **226**

Copyright National Academy of Sciences. All rights reserved.

Appendix D Technical Details for Geography Variables **229**

Appendix E Data Collection Report **231**

Appendix F Biographical Sketches of the Committee **246**

Copyright National Academy of Sciences. All rights reserved.

Boxes, Figures, and Tables

Boxes

S-1	Statement of Task	2
S-2	Methods of Adjusting the Data to Protect Confidentiality	5
1-1	Statement of Task	20
2-1	Relationship Categories Used in SIPP	41
2-2	SIPP Disclosure Avoidance Procedures	51
4-1	Sample Data Usage Agreement	94
9-1	Illustration of Feasibility: Descriptive Analysis Using Unique SIPP Content	166
9-2	Illustration of Feasibility: Longitudinal Analysis with Household Relational Data	167

Figures

S-1a	Stages of disclosure avoidance	14
S-1b	Disclosure avoidance approaches and tiers of access	15
6-1	Fully synthetic data	116
6-2	Selected variables are synthetic	117
6-3	Variables are synthesized for selected respondents	117
6-4	How a validation server works	129
9-1	Uses of SIPP data in the most cited and recent studies (percentage)	158
9-2	Unique SIPP content used in the most cited and recent studies (percentage)	159
9-3	Number of respondents reporting use of various SIPP modules	160
9-4	Share of first authors at institutions or in cities with Federal Statistical Research Data Centers (FSRDCs)	169
9-5	Determinants and barriers to accessing restricted Census data through the current FSRDC system	172
E-1	Characteristics of respondents to call for information	232
E-2	Number of different SIPP data sources used by respondents to the call for information	233
E-3	Types of respondents to the call for information that used each format of SIPP data file	233
E-5	Number of respondents to the call for information who used each module or topic area within SIPP	234
E-5	Number of modules or topic areas used by respondents to the call for information	235
E-6	Types of analysis performed by respondents to the call for information	235
E-7	Difficulties experienced by respondents to the call for information wıhen using the public-use files	236
E-8	Difficulties experienced by respondents to the call for information when using the synthetic data	236
E-9	Impact of encountering difficulties with accessing SIPP data	240
E-10	How the results from SIPP data were used	241
E-11	Fields in which SIPP data findings were published	242
E-12	Degree to which SIPP findings could be met by standardized tables	243

Copyright National Academy of Sciences. All rights reserved.

Tables

1-1	List of Briefings Provided to the Panel	23
2-1	Data Collected in SIPP 2020, by Broad Category	38
2-2	Survey of Income and Program Participation Bibliographic References, 2000–2014, by Topic	55
3-1	Percentage of SIPP Households That Are Unique in Wave 1, Based on the Number of Types of Information Included	73
3-2	Percentage of SIPP Households That Are Unique Across Waves, Based on the Number of Types of Information Included	74
3-3	Percentage of SIPP Households That Are Unique, Based on the Number of Types of Information Included, and Replacing Occupation with Highest Level of Education	75
3-4	Key Areas in Which Three Commercial Databases Have Data That Correspond to SIPP Data	77
9-1	Matrix for Evaluating Feasibility with the Context of Various Modes of Access	165
9-2	Example of Evaluation of Accessibility by Mode of Access and User Type (1=low to 4 = high)	170

Copyright National Academy of Sciences. All rights reserved.

Acronyms

ACS	American Community Survey
CPS	Current Population Survey
DRB	Disclosure Review Board
FSRDC	Federal Statistical Research Data Center
id	identifier
IRB	institutional review board
IRS	Internal Revenue Service
PIK	Protected Identification Key
SAE	small area estimation
SDL	statistical disclosure limitation
SODA	secure online data access
SIPP	Survey on Income and Program Participation
SSA	Social Security Administration
SSB	SIPP Synthetic Beta
VDE	virtual data enclave

Copyright National Academy of Sciences. All rights reserved.

Glossary

Added noise for privacy protection: The altering of survey responses (e.g., by adding or subtracting some amount), which may vary across responses, and may be randomized both in terms of which data are altered and in how much the data are altered.

Bottom-coding: Setting a minimum value that may be released; for example, all values at or below $1,000 are set to $1,000.

Data perturbation: Changes to the data to protect confidentiality; these include adding noise and data swapping.

Data suppression: Reducing the amount of data that are released, such as by completely eliminating some measures, modifying the measures to make them less specific (e.g., top-coding, bottom-coding, and collapsing a continuous variable to become a categorical variable), and modifying what data are released (e.g., suppressed table cells based on fewer than three observations).

Data swapping: Data items are swapped between two or more comparable respondents in order to protect confidentiality and to provide deniability if someone claims to have identified a respondent; for example, swapping the state of residence for two respondents. The swapping may be *directed* (i.e., designed to address a particular disclosure risk for a respondent) or *random*. The purpose is to maintain the same overall totals (and hopefully similar statistical relationships) while protecting the confidentiality of who gave what response.

Data Use Agreement (DUA): Specifies limitations on how data may be used and publicly released; for example, that the data must be used only for statistical purposes and not to identify individuals.

Differential privacy (DP): Differential privacy is the leading form of formal privacy used by government agencies and by researchers on privacy methods. It is a framework for both quantifying the level of disclosure risk (under several related metrics) and developing

Copyright National Academy of Sciences. All rights reserved.

disclosure limitation methods that control the risk of single and multiple releases under those metrics.

Disclosure Review Board (DRB): A committee that sets limits on what data may be released; for example, by limiting what variables may be included in a public-use file, or restricting what tables may be published.

Federal Statistical Research Data Center (FSRDC): These are created through partnerships between federal statistical agencies and research leader institutions. They provide secure access to restricted data, either on-site or virtually. There is an approval process for allowing access, both for individuals seeking access and for the project to be performed. There are also financial costs involved.

Formal privacy: Formal privacy refers to any rigorous and unambiguous framework for quantifying disclosure risk in an internally consistent statistical framework that bounds the success probability of a wide class of potential attacks on privacy.

Gold Standard File: A file containing original (nonsynthesized) data created by the Census Bureau as a step toward producing the synthetic data. It is also used to verify whether statistics based on the synthetic data are consistent with those using the original data. It is not a master file of all SIPP original data, but rather was created specifically for the SIPP synthetic data.

Institutional Review Board (IRB): A committee that reviews potential research studies on human subjects and monitors them to ensure that they comply with applicable regulations, meet commonly accepted ethical standards, follow institutional policies, and adequately protect research participants.

Microdata: Data at the level of individual persons or respondents (as differentiated from summary statistical data).

National Secure Data Service (NSDS): The creation of an NSDS has been proposed by the Commission on Evidence-Based Policymaking to support statistical evidence building through data sharing and linking, providing a pathway for those desiring data access and expertise. Currently the National Science Foundation is carrying out a demonstration project to inform whether and how an NSDS will be established in the future.[1]

[1] https://ncses.nsf.gov/about/national-secure-data-service-demo

Copyright National Academy of Sciences. All rights reserved.

Privacy budget: A limit on the total amount of data that may be released within the context of formal privacy; each published statistic draws from this budget, and at some point either no additional statistics may be released or the privacy budget must be changed.

Public-use file: A file containing microdata that can be downloaded by anyone and may be analyzed and reported on without limitations.

Quasi-identifier: A data value that doesn't directly identify a person but that might be used to identify a person. For example, while a name or address would identify a person directly and be an identifier, a zip code would provide highly specific information that might *help* to identify a person and would therefore be a quasi-identifier.

Recoding: Often used in disclosure avoidance to reduce the number of discrete values that appear. For example, a continuous variable such as household income might be converted into a categorical measure with only a few categories, or a categorical variable such as the state of residence might be converted to a measure of geographic region. Recoding is also used to make two different databases more consistent with each other.

Restricted-use file: A file in which there are limitations in how the data may be analyzed and reported on. The restrictions may range from clicking on a user agreement concerning how the data will be used (with the data remaining available to anyone consenting to the user agreement) to a file in which an application process is designed to control who is allowed access, and in which strong controls may in place on what data can be accessed and what can be reported.

Secure online data access (SODA): A mechanism through which data may be accessed virtually (online) with controls to protect respondents' privacy, such as a process to gain permission to work with the data, controls on which data are accessible, and controls on what data may be publicly released. Also called a *virtual data enclave*, the term is used here to differentiate it from virtual data access through FSRDCs, with less stringent controls on access and potentially less complete access than would be available through an FSRDC.

SIPP Synthetic Beta (SSB): A synthetic data product created by the Census Bureau that combines selected data from SIPP with administrative tax and benefit data. Early versions were partially synthetic; the latest version is fully synthetic.

Synthetic data: Data that are created through a statistical modeling process to have the same statistical properties as the original data. The intention is to allow researchers to perform the

Copyright National Academy of Sciences. All rights reserved.

kinds of statistical calculations and get results that are similar to what would be produced from the original data without allowing access to the original data. The data may be fully synthetic (all of the records are generated from the model) or partially synthetic (some records or variables are generated from the model, while others are identical to the original data).

Top-coding: Setting a maximum value that may be released, for example by setting all values at or above $10,000 to $10,000.

Verification and validation system: A system designed to measure whether the results using altered data (due to disclosure avoidance procedures, particularly as applied to synthetic data) are comparable to those from the unaltered data and that provides researchers with validated results that may include added noise to protect confidentiality.

Copyright National Academy of Sciences. All rights reserved.

Summary

The Survey of Income and Program Participation (SIPP) is one of the Census Bureau's major regularly repeating surveys, with the first interviews conducted in 1983. It is designed to collect information about the costs, impacts, and effectiveness of government social programs designed to help the poor, and to answer questions about short-term financial and family dynamics. While largely focused on collecting financial data on income, assets, and expenditures, it also collects detailed information about the composition of households. It is unique in combining three features: an extremely large number of variables across multiple domains, detailed information on participation in social assistance programs, and the capacity to track monthly changes over a four-year period. These features make it a uniquely valuable resource for researchers and policy analysts, across a wide variety of domains.

While the Census Bureau has a general responsibility to protect the confidentiality of its survey respondents, SIPP presents special challenges. Because of the types of financial and personal data collected, the data can be considered highly sensitive, making confidentiality even more important. At the same time, SIPP collects so much data about multiple people in a household there is a heightened possibility, relative to most surveys, that those data can be combined in a way that will uniquely identify an individual. For example, the SIPP 2020 data include a household in Florida with a black male born in 1946 married to and sharing the household with an Asian female born in 1941, also with a child born in 1968 in the household; this combination of characteristics is sufficient to make the household unique within the SIPP sample, and one could add additional data from SIPP such as occupation, educational level, and homeowner/renter status that might make the household both unique and identifiable in the entire nation.

Further, the amount of data available from external sources that might be used to identify a SIPP respondent continues to increase. Many types of data are available on the public record, and while those data once were highly localized, the development of internet search tools and

Copyright National Academy of Sciences. All rights reserved.

powerful software allows the data to be aggregated into comprehensive databases. These data appear in commercial databases for tasks such as credit-checking, marketing, and house purchasing. As another example, as part of a policy of transparency some states have publish detailed information on the assistance they provide to needy individuals. Thus, databases that once might have been secure are becoming less secure as new data tools arise. Traditional practices used to protect the privacy of survey respondents are becoming less sufficient.

To determine the most appropriate disclosure avoidance procedures to use for SIPP, the Census Bureau asked the National Academies of Sciences, Engineering, and Medicine to convene an expert panel to address the following charge (Box S-1).

BOX S-1

Statement of Task

The National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc panel to consider the challenges facing the protection and release of public-use data from the Survey of Income and Program Participation (SIPP). As part of its fact gathering, the panel will consider:

- the evolving privacy risks to releasing survey data;
- developments in methods for protecting privacy and reducing risks of disclosure, including formal privacy methods being implemented at the Census Bureau;
- the dimensionality and longitudinal nature of the SIPP data;
- the linking of SIPP data with administrative data;
- existing SIPP data products and the utility of detailed public-use microdata that enable scientific discovery;
- selected other SIPP data products, such as a small area estimates program for key SIPP measures; and
- the need for protecting the confidentiality of the SIPP data, potentially across multiple data releases, while providing timely access for the many research uses of SIPP.

The panel will produce a report with conclusions and recommendations for disclosure protection and data provision from the SIPP program.

Copyright National Academy of Sciences. All rights reserved.

In response, the National Academies formed a panel consisting of experts in statistics, survey methodology, economics, computer science, data science, policy evaluation, and sociology.

The panel organized the statement of task into four broad questions:

- What are the needs of researchers and the general public in terms of making use of the SIPP data (Chapter 2)?
- What are the disclosure risks facing SIPP, and how can they best be assessed (Chapter 3)?
- What tools are available for addressing disclosure risks in SIPP, and how do they compare (Chapters 4 through 8)?
- What is the proper balance between protecting the confidentiality of respondents while supporting the best way to use data for valid inferences (Chapter 9)?

To address these four questions the panel conducted literature reviews; invited outside experts to provide information on SIPP, disclosure avoidance approaches, and data ethics; examined the 2020 SIPP public-use file and its documentation; and conducted a data collection asking SIPP data users to complete a short online questionnaire about their experiences in working with the SIPP data.

THE PANEL'S CONCLUSIONS

SIPP Is a Unique Resource and Continuing Access Is Important

The panel found the SIPP data to be a unique resource, offering data that are useful across a wide variety of research topics and academic disciplines. They are a major source for policy analysis, both by the federal government and by academic researchers. SIPP is used to examine multiple aspects of the U.S. economy, monitor the use of federal programs, examine the need for and impact of policies and programs, and conduct sociological research on topics such as parental presence and children in shared households. For example, by providing information on income, demographic characteristics, and program participation, SIPP can be used to model the impact of changing eligibility requirements.

Copyright National Academy of Sciences. All rights reserved.

The public-use file is the version of the data that has received the greatest use; therefore, maintaining such access should be a top priority with steps taken to protect confidentiality by limiting statistical disclosure. Otherwise, the value of SIPP can be expected to diminish, the number of users is likely to drop, and support for continuing SIPP as an ongoing survey is likely to decline. Researchers' use of SIPP was highly varied, using a mixture of the various modules within SIPP, and often used complex modelling. As a result, producing a dataset consisting of a limited set of key variables would not meet the needs of most current SIPP users.

SIPP's Current Level of Risk of Disclosure Remains Undetermined

Key to developing disclosure limitation strategies is to fully understand and measure the risk of disclosure. The actual risk of disclosure in SIPP data products has long been relatively unknown. The panel commends the Census Bureau on recently conducting a re-identification study to measure the likelihood that a data intruder could identify survey respondents by matching data in the SIPP public-use file with data contained in income tax data and Social Security records. The results were largely encouraging; some strata had no confirmed links at all, and overall less than 3 percent of putative links were confirmed as being correct (i.e., most putative links would be false positives); however, some strata had rates between 10 and 25 percent. Age was the measure that was most disclosive, though geography was used as a stratification variable rather than as a linking variable, so its impact relative to age cannot be measured. The study concluded there were no vulnerabilities requiring an immediate change to disclosure avoidance strategy.

The panel found the re-identification study to be highly informative but insufficient to properly evaluate the disclosure risks present in the public-use file. The report itself was very straightforward in acknowledging its limitations, and the research literature has developed many sophisticated methodologies for re-identification and other privacy attacks, making it difficult to draw confident conclusions from experiments with just one approach. For example, the report did not provide adequate information on how the presence of measures of changes over time affected the risk of disclosure, or of how looking at households rather than individuals might present a risk. The study was also too limited in examining what external databases (and consequently what types of data) might be used to identify survey respondents.

Copyright National Academy of Sciences. All rights reserved.

The Tools Available for Disclosure Avoidance

A variety of tools are available to support disclosure avoidance, and these might broadly be classified as falling within two categories: changing the data in ways that would make the data safe from disclosure and changing access to the data to prevent improper use. These two types of approaches can be used alone or in combination.

Box S-2 summarizes the ways in which data can be adjusted to protect confidentiality.

BOX S-2

Methods of Adjusting the Data to Protect Confidentiality

Traditional methods include the following:

- *Data suppression*: removing disclosive measures from the data file
- *Lessening the level of detail* contained in the data:
  - Rounding the data (e.g., to the nearest hundred)
  - Bottom-coding (e.g., using a single category for all values less than $1,000)
  - Top-coding (e.g., using a single category for all values greater than $10,000)
  - Recoding (e.g., converting the respondents' ages to three categories)
- *Perturbing the data:*
  - Data swapping (e.g., swapping the state for two individuals with similar characteristics)
  - Adding noise (e.g., adjusting the income by a randomized amount)

Using synthetic data: This is a process by which the actual values are replaced by estimates of those values based on other information about the respondent; one might replace only selected variables or the entire record.

Formal privacy: Differential privacy, absolute disclosure risk, and relative disclosure risk all provide metrics for determining how much data can be safely released, and could potentially be controlled by adding noise to the data and/or creating synthetic data. This report uses the term "formal privacy" to refer broadly to all such precise approaches to disclosure risk, of which

Copyright National Academy of Sciences. All rights reserved.

differential privacy is the most well-developed and widely adopted. Potentially, when data are altered there may be a validation process in which the statistical outputs are compared with results using the original data as a way of verifying that the results are consistent with the original data.

Currently the Census Bureau's options are limited in the sense that there are no off-the-shelf tools for developing either a fully synthetic data file or a differentially private data file. It would be viable to synthesize selected variables or to produce a table generator that restricts what tables can be produced and that adds noise to produce differentially private output, but the large number of variables and the presence of multiple observations over time both are complicating factors that could not be addressed at a full-file level in a reasonable amount of time.

A different approach is to change access to the data. Multiple tiers of access can be created, with each tier potentially varying in terms of what data are available, how the data can be analyzed, and what data can be publicly released (or even removed from the premises). SIPP currently offers the following tiers: First, SIPP offers a public-use file, in which some data have been removed or altered. Second, SIPP offers access through Federal Statistical Research Data Centers (FSRDCs), which offer potentially complete access to the data without alteration or suppression but require an approval process to get access and to have results publicly released; this currently is offered both at selected physical facilities and through online access, with the same approval process and fees required for both ways. Third, SIPP offers access to synthetic data (which requires a more limited approval process), providing access to only a small subset of the SIPP data plus selected administrative data, all synthesized. Other types of access used by federal agencies include a restricted-use file (provided on CD or by download, with restrictions on how the data may be stored, analyzed, and reported) and secure online data access, in which data users have online access to the data, with restrictions placed on the analysis and reporting of the data.

Copyright National Academy of Sciences. All rights reserved.

Maintaining Usability While Protecting Confidentiality

Full usability is achievable by releasing all of the original data, and full confidentiality is achievable by suppressing all of the data. The difficulty is in finding an appropriate balance between the two. The panel developed several conclusions to guide this process.

Current SIPP users tend to make use of a wide number of modules within the SIPP questionnaire, and they tend to engage in complex analyses, both in the way they manipulate the data and in the statistical procedures they choose. For this reason, neither providing a set of standardized tables nor providing a table generator is likely to meet all their needs, although such tools might increase the number of users, especially given the difficulty of working with SIPP data. Similarly, releasing only selected modules would not meet their needs.

Current SIPP users also frequently rely on the public-use file, so eliminating the public-use file as a way of protecting confidentiality may result in a substantial loss of SIPP users. In particular, current users find the FSRDC process to be highly onerous and would want a mid-tier mode of access to the extent that a public-use file cannot meet all their needs.

SIPP data users need more information about how SIPP data are altered so that they can analyze the data appropriately. Additionally, to make the best choice, they need information about the implications for analysis of choosing each of the available tiers of access.

At the present time, the Census Bureau has not conducted a formal assessment of the trade-off between usability and privacy protection for the SIPP.

RECOMMENDATIONS

Using a Public-Use File and Intermediate Tiers to Facilitate Access

The Census Bureau should continue to produce a public-use file to satisfy the needs of those wishing to perform complex analyses. To the extent that data cannot be included in the public-use file, the Census Bureau should look for an intermediate tier of access that is less onerous than FSRDCs (Recommendation 5-1).

The use of measures of geography, including state identifiers, is a highly important function because policy researchers have used differences across states in state policies as a tool for measuring the impact of those policies (Recommendation 8-1). However, because of the

Copyright National Academy of Sciences. All rights reserved.

capacity of geography to divide the sample and the U.S. population into relatively small groups, it presents likely disclosure risks. For some types of situations, researchers may only need to distinguish between states without necessarily being able to identify them, while other types of research will require precise identification of the states. This divergence of research needs is well suited to offering multiple tiers of access, allowing some types of analysis to be performed on the public-use file, while others will require more secure forms of access.

Strengthening Current Approaches to Measuring the Risk of Disclosure

Current information about disclosure risks is still too incomplete to make a proper judgment about the safety of the current public-use file. Contingent on the findings from future re-identification studies, it seems likely that some variables (including demographic and geography variables) may need to be given statistical disclosure treatment, such as by collapsing to fewer categories, synthesizing, or developing alternative means of access to the variables or their analytical use (Recommendation 4-1); these approaches are discussed in the following section. A regularized process of re-identification studies to evaluate disclosure risks is needed, including several types of analysis not yet specifically addressed by the Census Bureau (Recommendation 3-1). These include considering the use of additional variables for matching, such as home values and education levels, examining the clustering of characteristics across multiple individuals within a household, examining alternative thresholds for identifying putative links, and testing for changes over time. The Census Bureau would benefit from involving experts from the research community in the design of more creative attack strategies tailored to the SIPP data (Recommendation 3-3).

Choosing Among the Tools Available for Disclosure Avoidance

The panel recommends continuing to release a public-use file that is as large as possible while protecting confidentiality (Recommendation 9-3), while also creating new tiers of access as a way of providing access to variables that are dropped altogether in the public-use file or to original variables that were recoded in the public-use data to protect confidentiality (Recommendation 9-4). Data users' needs will vary depending on the particular analysis that is planned and the variables that are involved; for example, some users may be satisfied with categorical version of age (perhaps as long as detailed ages are available for determining

Copyright National Academy of Sciences. All rights reserved.

program eligibility) and others will need more detailed data, while converting state to geographic region would hinder many analyses. The current FSRDC approval process and financial costs are too onerous even with online access now being permitted, and an intermediate solution can be provided for those situations in which the public-use file is not sufficient. Chapter 5 suggests the creation of secure online data access (SODA), which would (1) provide more data than in the public-use file but less than in FSRDCs, (2) have less restrictive procedures than FSRDCs for obtaining access, analyzing the data, and obtaining permission for public release of the findings, and (3) provide a quicker and simpler process for disclosure review of findings for public release. A SODA would function by creating a controlled environment for data access and especially data release, and by requiring signed user agreements to gain access. A key is that to maintain high usability, a SODA should be readily available, with a simple and short application process that can be processed in a timely manner. The Census Bureau will also need an appropriate infrastructure for supporting multiple users of SODA, which may require hiring an outside contractor. (The panel takes no position on whether it is better for the Census Bureau to provide the infrastructure internally or hire an outside contractor. We note that there are contractors who already have such an infrastructure in place, and that the Census Bureau might find their experience useful in determining the best approaches, and that hiring an outside contractor might require less set-up work on the part of the Census Bureau.)

In the short term, Census might also consider augmenting public-use files with synthetic versions of variables that are deemed to be disclosive. These synthetic variables may be created using prediction modelling that is designed to emulate the statistical properties of the original data. Using a set of statistical models, the statistical estimates from the synthetic and original datasets may be compared to assess the usability of augmented public-use files. Current technology can handle a limited set of variables to be synthesized, though considerable research will be needed to assess usability.

A long-term ideal might be a fully synthetic, differentially private public-use data file, but this does not seem realistic in the near term (given current tools for data synthesis and differential privacy) and may prove to be impossible. Research is advancing rapidly, so it is difficult to know what tools may become available in the near future. The Census Bureau should continue to support research on disclosure avoidance and should work actively with both an advisory group and external researchers in developing appropriate tools and strategies

Copyright National Academy of Sciences. All rights reserved.

(Recommendations 3-3 and 9-6). The external researchers might include both privacy protection specialists and SIPP data users.

An important consideration in creating synthetic versions of the variables is the assessment of inferential validity (Recommendation 6-2). This can be assessed by investigating the sampling properties of the analytical results that use synthesized variables and comparing them with the results using the original variables. This can be routinized using several candidate analyses from the literature or a set of analyses developed in consultation with the subject matter experts.

Table generators, or more generally, Remote Analysis Platforms (RAPs), can provide a method of data access with a built-in differential privacy framework. Depending on the interagency agreements that the Census Bureau has with other federal agencies, the table generators might include synthetic versions of linked administrative data. Here the users should be able to submit queries or analysis requests remotely using the SIPP metadata (e.g., codebook) and noise-infused results that are calibrated to be consistent with the results from the actual data could be sent back to the users. In the short term, this framework may be more useful to new users or for preparing research proposals than for the complex analyses often performed by SIPP data users. A fully functional RAP with the ability to perform complex analysis may take longer to develop. The Census Bureau's previous experience with a SIPP table generator was abandoned as being too difficult to maintain; the Census Bureau might consider hiring an outside contractor for such work. This decision might depend on the extent to which the Census Bureau is willing to dedicate its own resources to such a project, and on whether contractors might have systems in place that could be adapted to meet data user's needs with minimal development effort. Table generators can be designed to infuse noise and to limit the production of data involving very few respondents in a cell, so per-user disclosure review may be unnecessary.

Steps That Will Help in Balancing Usability and Confidentiality

Particularly to the extent that new disclosure avoidance procedures are adopted, greatly enhanced communication would be beneficial. Data users need better documentation and tools to properly make use of the data (Recommendation 9-5). The Census Bureau should release information about its actions to protect data, balancing the need to keep some information

Copyright National Academy of Sciences. All rights reserved.

protected or secret from potential intruders with the ability for users to work with the data properly.

In particular, data users need to understand:

- the current risks to confidentiality so they recognize the need for a change in tiers of access;
- how to estimate the impact of data synthesis and noise infusion on the validity and accuracy of statistical estimates (Recommendation 9-1);
- when and how to have their results validated against the Gold Standard File; and
- how to choose which tier of access is most appropriate.

Communications with data users should be in both directions, not only providing tools and documentation for users but learning from data users about their needs and experiences.

A key aspect to protecting confidentiality while enhancing usability is the creation of additional tiers of access. A remote table generator could do this in a limited way, though it would not meet the full needs of current SIPP data users. The provision of secure online data access would provide a more complete tool for allowing customized analysis of the complete (or relatively complete) data file.

One of the hindrances to usability is the current interpretation of Title 13. It may not be necessary to change that interpretation in order to provide secure online data access, since the move would still be toward increased protection of confidentiality as compared with the public-use file. Still, a modernization of the interpretation of Title 13 is desirable, both for secure online data access and for FSRDCs (Recommendation 9-7). Some of the current barriers are the requirements to show a benefit to the Census Bureau, produce a complete research proposal, and go through an extensive background check. None of these barriers is specifically required by Title 13, while more recent legislation assigns the goals of supporting evidence and multiple tiers of access. Census Bureau Policies DS002 and DS006 could be updated to reflect the changing requirement without changing Title 13 itself.

TIMELINE FOR MAKING CHANGES

SIPP data are too valuable to stop data releases while new approaches are being developed. For this reason, the panel recommends a graduated approach in which relatively

Copyright National Academy of Sciences. All rights reserved.

simple changes are made in the short term while research continues on developing longer-range solutions. The use of a graduated approach will also help users to adjust to the changes being implemented in SIPP. The recommendations here may also be difficult for the Census Bureau to implement on its own with its current resources. This may be a time to involve outside organizations for tasks such as maintaining secure online data access and assisting with the disclosure review process for publishing data. The Census Bureau could also benefit by including external researchers when performing re-identification studies and developing disclosure avoidance strategies, bringing in additional skills and perspectives to supplement the work of Census Bureau staff.

Given that it will take time to implement the recommendations in this report, and that more extensive analyses may reveal undesirable levels of disclosure risk, the Census Bureau may wish to consider short-term steps to lessen the risk of disclosure. One approach would be to strengthen traditional statistical disclosure limitation methods like more coarsening of the data, additional top codes, additional geography suppression, and similar methods. The re-identification exercise suggests that age might pose the greatest threats, and our analysis indicates that state identifiers also present risk, so these measures in particular could be considered as high priorities for disclosure avoidance steps. However, these increased protections come at the cost of reduced data utility, and any adoption of increased statistical disclosure limitation methods should seek to minimize the loss of that utility. Another option that has been adopted in a number of European countries is a registration process for those who download the public-use file, requiring users to provide identifying information and state the intended purpose of the data analysis. Alternatively (or in addition), users of public-use SIPP files could be required to sign a user agreement that required them to agree to specific uses of the data and to not disclose any data that could reasonably be regarded as confidential (NCES has adopted a variant of this procedure). While these options would not block a serious and sophisticated intruder from obtaining the data and attempting reidentification, they will make the general user community aware of these issues and sensitive to the problem.

Figures S-1a and S1b display how these various options compare in terms of the trade-off between ease of access and the availability of complete data. Ideally, once the raw data are finalized, a re-identification study is performed to estimate the level of disclosure risks. After the disclosure risks are known, a variety of approaches are possible, depending on the level of risk

Copyright National Academy of Sciences. All rights reserved.

and the planned tier of access. These range from traditional statistical disclosure limitations, such as those currently used by SIPP, to synthetic data (also used by SIPP for a small extract of the file combined with administrative data) and differential privacy (recently adopted for the 2020 decennial census). The four tiers of access are arranged in order of accessibility.

A table generator would be the easiest for users to access in the sense that it requires only internet access, not requiring computer programming skills or software, but at the cost of offering the least information, and with noise added to the data. A public-use file requires greater statistical and programming skills (the SIPP file structure is complex), but it can be downloaded by anyone, it allows data users to customize their analysis, and it makes more data available. Secure online data access places more restrictions on the user, such as signed user agreements and disclosure reviews of what data may be released, but it makes even more data available. FSRDCs tend to be difficult to access, with substantial user fees, background checks, and specific demands on the uses of the data; however, the original data can become available at FSRDCs and the data can potentially be blended with other sources.

While the various data disclosure approaches conceptually could be applied at any of the tiers of access, methods of changing or suppressing the data are most applicable to a table generator or public-use file, while the additional controls placed by secure online data access are intended to allow more complete access to the original data. However, one might still use synthetic data at an FSRDC, not necessarily to protect the SIPP data but rather to allow the merging of other data, possibly to protect the other data or possibly to allow the blending of data that are not directly matchable. Differential privacy, as a tool for controlling what data are output or released, could be used at any of the tiers.

Following the design of the tiers of access, a final step is to use re-identification studies and other empirical tests of privacy attacks to confirm that the disclosure risks have been resolved appropriately without creating new risks.

Copyright National Academy of Sciences. All rights reserved.

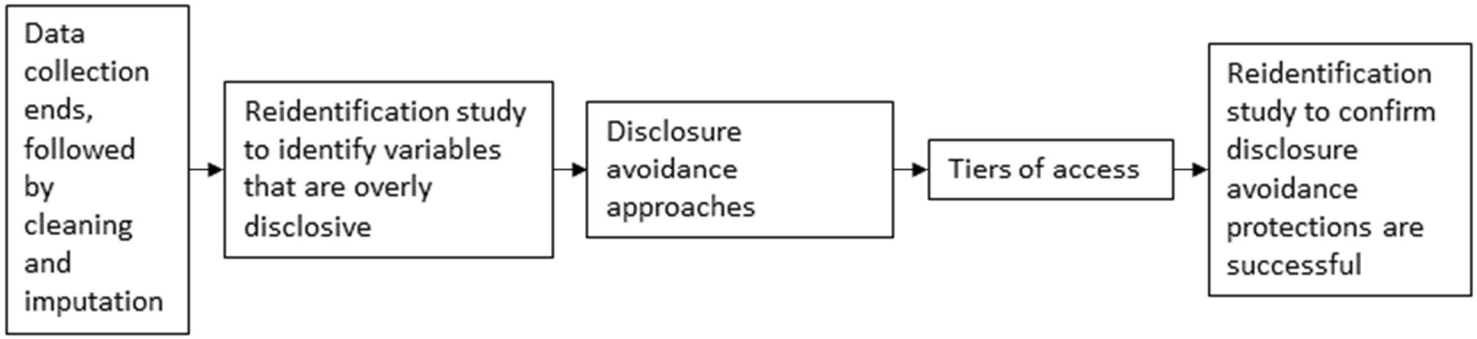


Figure S-1a Stages of disclosure avoidance.

Copyright National Academy of Sciences. All rights reserved.

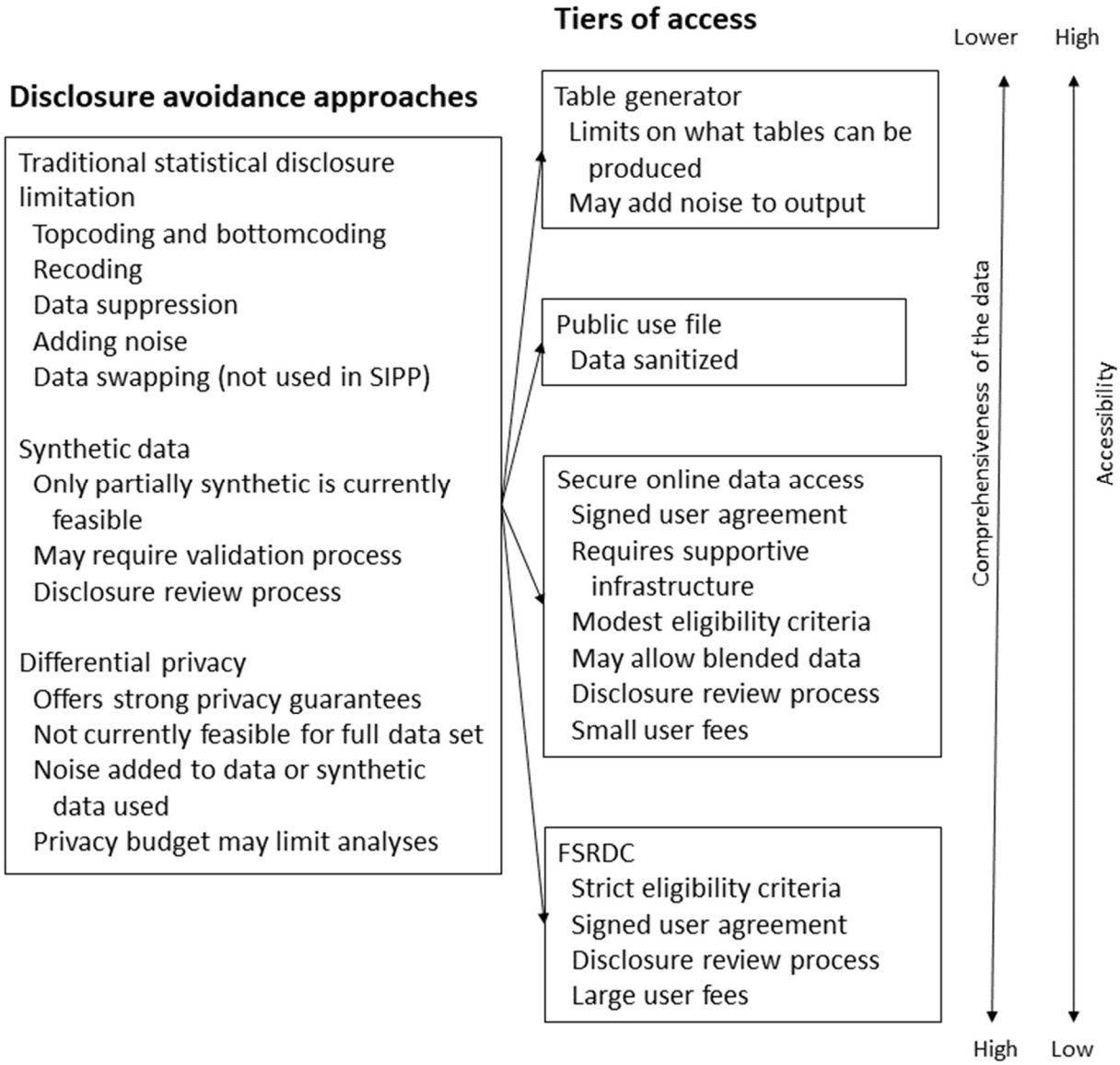


Figure S-1b Disclosure avoidance approaches and tiers of access.

Copyright National Academy of Sciences. All rights reserved.

1
Introduction

Concerns about privacy have changed over time, and the U.S. Census Bureau has changed its approach in response. The earliest (1790) census takers were required to post census lists in the town square for people to check and revise.[2] Until 1929, census records could be bought and shared. Until 1976, the Census Bureau's director had the discretion to grant disclosure exemptions.

The Census Bureau first began making promises of confidentiality to businesses as a way of addressing poor response rates in 1840, while the 1850 Census was the first one in which responses were not posted publicly. Still, promises were not always kept. During World War I, census records were used to support the military draft, despite President Taft's earlier promises of confidentiality. Protections added in 1940 were overturned by the Second War Powers Act in 1942 (though it expired in 1947). As a way of more securely establishing the right to confidentiality, Congress passed Title 13 in 1954 (and amended it in 1962 and 1990), which both places requirements on the Census Bureau and protects the data from others: i.e., "it is against the law to publish any private information that identifies an individual or business such…[as] including names, addresses [including GPS coordinates], Social Security Numbers, and telephone numbers."[3]

Over time, the Census Bureau has also changed its statistical procedures for protecting data, starting with the suppression or compression of potentially disclosive data in 1920 to the progressive adoption of techniques such as data swapping, whole table suppression, rounding, top-coding and bottom-coding. Most recently, for the 2020 decennial census, the bureau adopted

[2] Census Bureau. https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html

[3] Census Bureau. https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html#:~:text=It%20is%20against%20the%20law,Security%20Numbers%2C%20and%20telephone%20numbers

Copyright National Academy of Sciences. All rights reserved.

differential privacy (Dwork et al., 2017; Abowd et al., 2022)—a change that has provoked debate.

American society also has changed, with growth in the ways data are collected and used, including the combination of multiple databases to create more complete records. Helen Nissenbaum (2010) notes some of the important changes:

- "People going about their daily business in urban settings can expect to have their images monitored and recorded an average of 300 times a day by thirty separate CCTV systems;" (p. 33)
- Event-data recorders in modern cars record data such as engine speed, safety belt status, status of brakes during a crash, and acceleration;
- Every interaction in cyberspace is recorded;
- Database technologies have become available to a broad community of individual and institutional users;
- Advances in information science allow the search and retrieval of data dispersed across disparate locales; and
- Advanced statistical techniques allow meaning to be drawn from data that goes beyond the data's raw meaning.

The rapid growth of artificial intelligence has the potential to further increase disclosure risks, increasing the capacity to search across multiple data sources in order to find data that will identify a respondent.

These changes place the Census Bureau's Survey of Income and Program Participation (SIPP) in an increasingly difficult position. The survey collects data that are highly personal and confidential, including data on income, financial assets and liabilities, and household structures. Such data would be of interest to many, including data aggregators, lenders, advertisers, and identity thieves. Based on Title 13, the Census Bureau is legally and ethically bound to protect the confidentiality of SIPP survey respondents. As a practical matter, promising privacy is also a valuable tool for reassuring survey respondents that they may safely provide accurate information, affecting both survey response rates and data quality.

Yet what is needed to protect confidentiality? The simplest and most traditional approach has been to strip clearly identifying information from the data, such as names and addresses.

Copyright National Academy of Sciences. All rights reserved.

Beyond that, the Census Bureau has used tools such as top-coding, bottom-coding, data coarsening, and perturbations to help protect the privacy of the data. Even with these measures, a problem remains, because it is not just single pieces of data that may identify a person, but clusters of multiple pieces of information that may collectively identify a person: "it is well known that a small set of attributes can single out an individual in a population" (PPTTT, 2023, p. 9). As Nissenbaum (2010) writes:

> Data subjects and third-party harvesters alike are keenly aware of qualitative shifts that can occur when bits of data are combined into collages. This is, surely, one of the most alluring transformations yielded by information sciences and technologies. It is anything but the case that an assemblage of bland bits yields a bland assemblage. The isolated bits may not be particularly revealing, but the assemblages may expose people quite profoundly. (p. 123)

Through reidentification studies of the 2010 Census, the Census Bureau found that one in six individuals in the U.S. population could be re-identified using publicly available data (Bowen, p. 37) by using the Census Protected Identification Key (PIK), block, sex, and age:

> Our simulated reconstruction-abetted reidentification attack demonstrated that the tabular summaries from the 2010 Census can be converted into a 100% microdata file with geographic precision to the census block-level. Our simulated attack demonstrated that, depending on the quality of the external data used, between 52 and 179 million respondents to the 2010 Census can be correctly re-identified from the reconstructed microdata (Hawes, 2021).

SIPP differs from the decennial census in many ways, but one core distinction is that it is based on statistical sampling. In a census, it is readily apparent which combinations of characteristics will uniquely identify individuals; in a sample survey such as SIPP, one can use a small set of attributes to uniquely identify individuals among the survey respondents, but it is less clear whether those same attributes are unique within the entire population. The use of sampling does not change whether a particular combination of characteristics is unique, but it changes whether one can *know* that the combination uniquely identifies an individual in the population. However, there are tools for estimating the likely success of an attempt at re-identification: one study, using generative models, estimated that "99.98 percent of Americans would be correctly re-identified in any dataset using 15 demographic attributes" (Rocher, Hendrickx, & de Montjoye, 2019). Further, if external data are available for the entire

Copyright National Academy of Sciences. All rights reserved.

population, then the potential for determining that an individual is uniquely identified is greatly increased.

Two practical examples illustrate the issues that SIPP presents with respect to protecting confidentiality. In the SIPP 2020 data there was only one respondent who had the following characteristics: a black female in Idaho. However, given that 0.9 percent of people in Idaho in 2020 were Black,[4] it seems highly unlikely that such a person is unique in the entire population of Idaho. On the other hand, the SIPP data also show a household in Florida with a black male born in 1946 married to and sharing the household with an Asian female born in 1941 and with a child born in 1968 in the household; given the multiracial nature of the household and the combination of ages, such a combination might very well be unique in Florida. Moreover, many other types of characteristics might be added in from SIPP, such as occupation, educational level, and homeowner/renter status, to make the set of potential matches for the respondent small. Thus, a household that has not already been uniquely identified in the sample has an increased likelihood of being identified by including data on additional characteristics. This is one way in which SIPP may be more disclosive than the decennial census, despite being based on a statistical sample: it contains a substantial amount of highly detailed information, also including changes from one year to another. The potential ability to precisely identify unique households gives good reason to examine disclosure avoidance protections carefully.

Though SIPP faces challenges with regard to protecting confidentiality, these challenges can be met. As the Commission on Evidence-Based Policymaking (2017, p. 1) wrote:

> Traditionally, increasing access to confidential data presumed significantly increasing privacy risk. The Commission rejects that idea. The Commission believes there are steps that can be taken to improve data security and privacy protections beyond what exists today, while increasing the production of evidence. Modern technology and statistical methods, combined with transparency and a strong legal framework, create the opportunity to use data for evidence-building in ways that were not possible in the past.

[4] https://www.census.gov/library/stories/state-by-state/idaho-population-change-between-census-decade.html#:~:text=Race%20and%20ethnicity%20(White%20alone,%25%2C%20up%20from%2054.9%25).

Copyright National Academy of Sciences. All rights reserved.

STATEMENT OF TASK

To determine the most appropriate disclosure avoidance procedures to use for SIPP, the Census Bureau asked the National Academies of Sciences, Engineering, and Medicine to convene an expert panel consisting of experts in statistics, survey methodology, economics, computer and data science, policy evaluation, and sociology.

The following charge (Box 1-1) was given to the expert panel. Numbering has been added to make it easier to refer to different parts of the statement of task and is not meant to imply either a sequential process or prioritization of the different elements.

BOX 1-1
Statement of Task

The National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc panel to consider the challenges facing the protection and release of public-use data from the Survey of Income and Program Participation (SIPP). As part of its fact gathering, the panel will consider:

1. the evolving privacy risks to releasing survey data;
2. developments in methods for protecting privacy and reducing risks of disclosure, including formal privacy methods being implemented at the Census Bureau;
3. the dimensionality and longitudinal nature of the SIPP data;
4. the linking of SIPP data with administrative data;
5. existing SIPP data products and the utility of detailed public-use microdata that enable scientific discovery;
6. selected other SIPP data products, such as a small area estimates program for key SIPP measures; and
7. the need for protecting the confidentiality of the SIPP data, potentially across multiple data releases, while providing timely access for the many research uses of SIPP.

The panel will produce a report with conclusions and recommendations for disclosure protection and data provision from the SIPP program.

HOW THE PANEL DID ITS WORK

The panel's first step was to examine the Statement of Task to determine what information would be needed. We first met with the Census Bureau to discuss their view of the task statement. After internal discussions about the statement, we decided there were four key questions to answer:

Copyright National Academy of Sciences. All rights reserved.

- *What are the needs of researchers and the general public in terms of making use of the SIPP data?* This broad question has several components. We needed to know which data are used by researchers and what types of statistical analysis are performed. For example, one potential solution might be to divide the data into modules if different researchers only use particular sections of the data, and another solution might to be provide an online table builder, but the value of this solution would depend on researchers' need to manipulate the data or perform complex statistical analyses. This question particularly relates to items 3 (restructuring or dividing the data), 4 (a potential benefit of SIPP being the capacity of its data to be combined with other types of data), 5 (the value of the SIPP data), and 7 (supporting the use of SIPP) in the Statement of Task.
- *What are the disclosure risks facing SIPP, and how can they best be assessed?* If there are no current disclosure risks, then the work of the Census Bureau is greatly simplified, though it still might be wise to plan for those days in which disclosure risks increase as more types of data become available from alternative sources. The Census Bureau will benefit from effective tools to determine which data items present the greatest risk of disclosure, and to verify whether its disclosure avoidance practices are effective. This question particularly relates to items 1 (evolving privacy risks) and 3 (challenges presented by the level of detail and measurement of change over time).
- *What tools are available for addressing disclosure risks in SIPP, and how do they compare?* Identifying the problem is only a first step; the next is to develop a solution, or package of solutions, to address the problem. This question particularly relates to item 2 (methods of protection) and how it intersects with items 3 (affecting the complexity of disclosure avoidance), 4 (bringing in additional data, which may increase disclosure risks and involve the need to respect the requirements of other data providers), and 6 (increasing the difficulty in protecting confidentiality).
- *What is the proper balance between protecting the confidentiality of respondents while supporting the best use of the data?* It would be easy to protect confidentiality by denying any outside use of the data, and it would be easy to do the opposite: supporting researchers by providing full access to data. What is wanted, however, is a way of balancing both goals so that both confidentiality and data use are supported.

Copyright National Academy of Sciences. All rights reserved.

This question particularly relates to item 7 (protecting confidentiality while providing timely access), though it is also a foundational question that affects what tools and data can be developed and used.

The panel next examined what information it needed to address the Statement of Task. Collectively, the panel members brought experience and expertise in working with the SIPP data and in disclosure avoidance approaches (including traditional techniques, controlled virtual access, synthetic data, and differential privacy, with different panel members having different specialties. In part, the panel engaged in internal, mutual instruction, so that all panel members would share a common level of understanding; this included briefings and discussions on differential privacy, controlled virtual data access, and balancing privacy and usability.

The panel also required extensive information about SIPP and how it is used. The Census Bureau provided briefings giving us an overview of SIPP and describing key topics that would be relevant. These covered how SIPP works, what its major issues are, what decisions about SIPP and data protection had already been made, and what are SIPP's content, products, and concerns. They also covered SIPP uses of administrative data during production processing, background on SIPP synthetic data and numbers and types of users, SIPP's current and desired level of security, new developments within the Census Bureau on disclosure limitations, SIPP small area estimation, key estimates, and data quality, and a recent re-identification study conducted by the Census Bureau relating to SIPP.

The panel conducted multiple literature searches to identify the uses of SIPP, focusing on four search methods: (1) a bibliography provided by the Census Bureau on its website of research based on SIPP; (2) a search of publications within the last 5 years citing SIPP and focused on the topics addressed and the methods used; (3) a search of articles within the last 3 years on disclosure avoidance, as well as articles on the impact that promising privacy has on survey response rates; and (4) a Scopus citation search of articles listing SIPP in the title or abstract, looking particularly at the 50 publications with the most citations.

The panel downloaded considerable material about SIPP from the Census Bureau website, including the history and design of SIPP, documentation about the SIPP public-use file, and the SIPP 2020 public-use file itself. Selected tabulations were produced from the downloaded data file and then reviewed for accuracy by analysts at the Census Bureau.

Copyright National Academy of Sciences. All rights reserved.

Finally, the panel issued a call for information, asking SIPP data users to complete a short online questionnaire about how they used the data and what problems they experienced (see Appendix E). These data from 65 SIPP data users cannot be considered to be a nationally representative sample, but they provide a greater level of detail about how the SIPP data were used than would otherwise be available. All statistical findings were confirmed by a second National Academies staff member.

Table 1-1 provides a list of all the briefings provided to the panel by both outside and internal experts.

TABLE 1-1 List of Briefings Provided to the Panel

Date	Topic	Presenter(s)
6/6/22	Introduction to Survey of Income and Program Participation (SIPP) and Expectations of the panel	Jason Fields, Census Bureau
6/30/22	The SIPP Synthetic Beta (SSB)	Rachel Shattuck—Census Bureau
6/30/22	Model-based Imputation and Administrative Records in SIPP Processing	Benjamin Gurrentz—Census Bureau
6/30/22	An Introduction to the SIPP Content, Products, and Concerns	Adriana Hernández-Viver, Robert Munk, Yerís H. Mayol-García—Census Bureau
6/30/22	Small Area Estimates for the SIPP	Benjamin Gurrentz, Sam Szelepka—Census Bureau
6/30/22	SIPP Key Estimates and Data Quality	Ashley Westra—Census Bureau
6/30/22	SIPP's Current Level of Security	Holly Fee—Census Bureau
6/30/22	Protecting Respondent Confidentiality in the SIPP	Gary Benedetto, Rolando Rodriguez—Census Bureau
9/7/22	Balancing Data Privacy and Usability in the Federal Statistical System	V. Joseph Hotz—Duke University, Robert A. Moffitt—Johns Hopkins University
9/7/22	A Penny Synthesized is a Penny Earned? An Analysis of Synthetic Earnings Using Survey Responses and Administrative Records	Jordan Stanley, Evan Totty—Census Bureau
9/7/22	Synthetic Data and Census Bureau Directions for Privacy Protection	Jerry Reiter—Duke University
10/3/22	Statistics and Privacy	dana boyd—Microsoft Research and Georgetown University
10/3/22	SIPP User Experiences	Bradford Chaney—National Academies

Copyright National Academy of Sciences. All rights reserved.

11/14/22	A Modern Container-based Approach for Development of and Access to Confidential Data	Lars Vilhuber—Cornell University
11/14/22	Differential Privacy	Salil Vadhan—Harvard University
12/12/22	Restricted Data Access at ICPSR	Amy Pienta—University of Michigan
3/21/23	SIPP 2014 Panel Re-Identification (Re-id) Study Findings and Recommendations	Aref Dajani, Steve Clark, Phyllis Singer—Census Bureau

In some areas the panel relied on its own internal expertise where judgment calls were required or the information, by its nature, could not be clearly documented. For example, this report states that no software for creating synthetic data files is currently ready for handling a file with the size and complexity of SIPP. Such judgments are noted in the text where they occur.

The panel divided into three teams to consider what the report's content should be and draft the report itself. Each panel member served on two teams. The following topics delineated the teams.

A: User perspectives and the balance between privacy and confidentiality,

B: Assessment of privacy and disclosure risk, and

C: Approaches to disclosure avoidance from a SIPP perspective.

Each team reviewed and assessed the report outline to ensure that all important topics were addressed and that the report was well organized. Each team approached its topic by conducting literature reviews to identify the most up-to-date research on it, considering nuances, and conducting assessments of the published research based on analysis of SIPP data. Teams met in separate, closed, remote meetings three times, each meeting followed by a closed panel meeting to review and summarize their progress.

In preparation for the final closed hybrid panel meeting, the teams met a fourth time — making plans to draft their sections of the report and prepare presentations for the panel meeting. At the panel meeting, the draft report was reviewed chapter by chapter, with each team presenting the parts of the report they had prepared. Content was discussed and revisions agreed to. The documents were revised and updated to prepare the first draft of the report.

The panel met again in a closed session to finalize the conclusions and recommendations. After further edits to the text, the report was sent out for review by six independent experts in the following fields: experience with SIPP, disclosure avoidance approaches, demography, and small

Copyright National Academy of Sciences. All rights reserved.

area estimation. The panel met one final time in closed session to discuss the comments received from the outside review.

OUTLINE OF THE REPORT

This report is organized in the following manner. Chapter 2 provides an overall summary of SIPP, describing how the survey is conducted, what data are collected, how the data are used, and what disclosure avoidance protections are currently in place. Chapter 3 examines how disclosure risks occur and how disclosure risk is currently or should be measured. Chapters 4 through 8 collectively discuss the disclosure avoidance approaches that are available. Chapter 4 provides a general overview of disclosure limitation approaches, discussing what approaches are available and how they might be used collectively as a package of different approaches for different situations. Chapter 5 begins a more detailed examination of individual disclosure approaches, looking at secure online data access, while Chapter 6 discusses partially synthetic datasets, Chapter 7 discusses a table generator and remote analysis platform, and Chapter 8 examines the special challenges presented by geographic variables and how they might be addressed.

Differential privacy is both a metric used to guide disclosure avoidance and a framework for developing tools that limit disclosure with respect to this metric. Thus, it is discussed primarily in Chapters 3 (as a measurement tool) and 4 (in the overview of disclosure avoidance approaches), along with references in other chapters as appropriate, and again in Appendix C. Chapter 9 examines ways to create a balance between promoting usability of the SIPP data and preserving confidentiality. It also describes the experiences of SIPP users, based on 65 responses from those users who completed a questionnaire about their experiences. Finally, Chapter 10 provides a summary of the panel's conclusions and recommendations.

Six appendices provide supplementary and generally technical information, particularly including formulas where appropriate. Appendix A provides information on measuring disclosure risk, complementing Chapter 3. Appendix B provides information on making inferences based on synthetic data, as referenced in Chapter 6. Appendix C provides technical information on differential privacy in table generators, as referenced in Chapter 7. Appendix D concerns geography variables, as discussed in Chapter 8. Appendix E provides a description of how data were collected from SIPP users who responded to the call for information, and provides

Copyright National Academy of Sciences. All rights reserved.

more detailed statistical results than are contained in the main text. Appendix F provides biographical sketches on the panel members.

This report does not address how future malicious actions might be anticipated and prevented by identifying suspicious behavior that seems directed toward identifying respondents. It is difficult to identify actions that might be taken with regard to public-use files, since the downloading of multiple files by itself is not a suspicious action; also, the Census Bureau does not require people to register to download a public-use file, and thus does not have records of data users. When operating within a user agreement, such as through an FSRDC, there are mechanisms in place to prevent the release of confidential information, though those mechanisms could potentially be expanded, such as through the use of artificial intelligence.

PREVIOUS CNSTAT STUDIES RELATING TO SIPP

The current study is one of several conducted by NASEM over the last 34 years. Following is a brief description of the earlier studies, along with those conclusions and recommendations from them that are most relevant to the current study.

An Interim Assessment of SIPP (1989)

Five years after the initiation of SIPP, the Census Bureau and the U.S. Office of Management and Budget asked CNSTAT to perform an independent study of SIPP. In an initial interim report, the committee examined the goals of SIPP, how well the survey was meeting the goals, and the quality and utility of SIPP data products. The committee found that "SIPP is making a vital contribution to understanding the characteristics and dynamics of the population at economic risk, and the ways in which federal programs meet—or fail to meet—economic needs" and that SIPP "provides data not elsewhere available that is integral to policy analysis of income maintenance programs" (National Research Council, 1989, p. ix). As part of the study, the committee conducted interviews with selected federal agencies, finding that six of them made major use of SIPP: Food and Nutrition Service, U.S. Department of Agriculture (USDA); Census Bureau, U.S. Department of Commerce; Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services (DHHS); Social Security Administration; DHHS Congressional Budget Office; and Congressional Research Service. Eight other agencies made occasional use of SIPP: Economic Research Service, USDA; National Center for Health Services

Copyright National Academy of Sciences. All rights reserved.

Research, DHHS; Family Support Administration, DHHS; U.S. Department of Education; U.S. Department of Housing and Urban Development (HUD); Bureau of Labor Statistics, U.S. Department of Labor (DOL); U.S. Office of Management and Budget; and U.S. Commission on Civil Rights. Finally, two agencies were found to have a potential use of SIPP: Bureau of Economic Analysis, U.S. Department of Commerce; and the U.S. Department of the Treasury.

Following are the recommendations from the 1989 report that are most relevant to the current study:

- *Recommendation 4*. The committee recommends that the Census Bureau give high priority to matching SIPP with administrative records and that the Office of Statistical Policy in the Office of Management and Budget encourage and promote interagency matching projects. (p. 72)

- *Recommendation 5*. The committee recommends that the Census Bureau commit itself to ensuring the timely release of high-quality public-use microdata files and improving the comprehensiveness of SIPP documentation. The committee endorses the current effort of the Census Bureau to make SIPOP data products available on a more timely basis. (p. 72)

- *Recommendation 6*. The committee recommends that the Census Bureau explore methods to facilitate use of SIPP and put more resources toward efforts to make the data more accessible to users, whether those activities take place inside the Bureau or are conducted by private contractors. (p. 72)

The Future of the Survey of Income and Program Participation (1993)

This 1993 study was part of a reassessment and redesign effort conducted by the Census Bureau after roughly 9 years of SIPP's operation. The panel reviewed the survey's goals, content, and relationship to other data collections; survey and sample design; data collection and processing; publications and other data products; analytical methods for using the complex longitudinal data; methodological research; and management and oversight of the SIPP program. Much of the report concerned the content and methodology of SIPP, which are outside of the charge to this panel, but a few of the report's recommendations are particularly relevant:

Copyright National Academy of Sciences. All rights reserved.

- *Recommendation 6-3*: The Census Bureau should continue to develop improved microdata. (p. 10)
- *Recommendation 6-4*: The Census Bureau should work to improve documentation and related user information services for SIPP. (pp. 10–11)

Reengineering the Survey of Income and Program Participation (2009)

This study was part of a re-engineering of SIPP started in 2006 by the Census Bureau, with the panel focusing on the linking of administrative records and SIPP data. Following are some of the most relevant conclusions and recommendations coming from that study.

- *Conclusion 2-1*: The Survey of Income and Program Participation (SIPP) is a unique source of information for a representative sample of household members on the intrayear dynamics of income, employment, and program eligibility and participation, together with related demographic and socioeconomic characteristics. This information remains as vital today for evaluating and improving government programs addressed to the social and economic needs of the U.S. population as it did when the survey began 25 years ago. (p. 2)
- *Recommendation 3-5*: The Census Bureau should request the Statistical and Science Policy Office in the U.S. Office of Management and Budget to establish an interagency working group on uses of administrative records in the SIPP. (p. 5)
- *Recommendation 3-7*: When considering the addition to SIPP of administrative records values for variables that have never been ascertained in the survey itself, the Census Bureau should ensure that the benefits from the added variables are worth the costs, such as additional steps to protect confidentiality. The bureau should consult closely with users to be sure that the added variables are central to SIPP's purpose to provide information on the short-run dynamics of economic well-being and that their inclusion does not compromise the ability to release public-use microdata files that accurately represent the survey data. (p. 6)
- *Conclusion 3-6*: Multiple strategies for confidentiality protection and data access are necessary for a survey as rich in data as SIPP. Public-use microdata files, which are available on a timely basis and in which confidentiality protection techniques do not unduly distort the relationships in the data, are the preferred

Copyright National Academy of Sciences. All rights reserved.

mode of data release. Some uses may require access to confidential data that at present can be provided only at one of the Census Bureau's Research Data Centers. (p. 6)

- *Recommendation 3-8*: The Census Bureau should develop confidentiality protection techniques and restricted access modes for SIPP that are as user-friendly as possible, consistent with the bureau's duty to minimize disclosure risk. In this regard, the bureau should develop partial synthesis techniques for SIPP public-use microdata files that, based on evaluation results, are found to preserve the research utility of the information. For SIPP data that cannot be publicly released, the Census Bureau should give high priority to developing a secure remote access system that does not require visiting a Research Data Center to use the information. The bureau should also deposit SIPP files of linked survey and administrative records data (with identifiers removed) at all Research Data Centers in order to expand the opportunities for research that contributes to scientific knowledge and informed public policy. (p. 6)
- *Conclusion 4-7*: The release of SIPP data is often not timely. Data from the 2004 SIPP panel were generally released more than 2 years after being collected. Other panel surveys have more timely data release, often within a year of data collection, which enhances their usefulness to external users. (p. 9)

The 2014 Redesign of the Survey of Income and Program Participation: An Assessment (2018)

Following a redesign of SIPP in 2014, this study was designed as an independent evaluation comparing the new design and old design by comparing key estimates, evaluating the content, evaluating the impact on respondent burden, and considering content changes for future improvement of SIPP. Some relevant conclusions and recommendations follow.

- *Recommendation 5-2*: The Census Bureau should continue to investigate effective use of administrative records, such as Internal Revenue Service data mentioned in this report, to enhance the quality of data collected in SIPP. This research should include identification and correction of both false positives and false negatives.

Copyright National Academy of Sciences. All rights reserved.

Efforts should continue to obtain administrative data for important state-run programs such as the Supplemental Nutrition Assistance Program (p. 6).

- *Recommendation 9-1*: The Census Bureau should release SIPP data within 1 year of the end of data collection. (p. 7)
- *Recommendation 9-2*: The Census Bureau should commit to developing a formal release process and timetable for SIPP, including release of data and publications. The publications should contain key estimates and data characteristics, and they should be a formal and publicized part of the process. Timetables should be met. (p. 7)
- Recommendation 9-3: The Census Bureau should continually improve the documentation available to data users. (p. 8)
- *Recommendation 9-5*: The Census Bureau should formally solicit data users' feedback, input, and ideas after those users have had time to work with wave 1 and wave 2 datasets from the redesigned SIPP. (p. 8)

Copyright National Academy of Sciences. All rights reserved.

2
Overview: SIPP and Its Uses

The Survey of Income and Program Participation (SIPP) has been conducted by the Census Bureau since 1983. SIPP was designed to provide data for researchers and policymakers to estimate the costs, impacts, and effectiveness of government social programs designed to help the poor, including how well programs reached their target populations, what the impact of program participation was on participants and their households, what overlap occurred among populations served by different programs, and why and how often people transitioned into and out of programs. SIPP was also designed to answer questions about short-term financial and family dynamics. Since its original design, it has evolved to provide data on a broader range of social issues and has become a general-purpose dataset for information about U.S. society. SIPP data have unique features concerning these topics that are not provided by any other household survey, either federal or nonfederal (see NASEM [2018] as well as the other NASEM reports summarized in Chapter 1).

This chapter provides background information about SIPP: a brief history of the survey, its uniqueness, the content collected, data products produced, historical approaches to protecting data from disclosure, and how the data are used.

FORMATIVE AND EARLY YEARS OF SIPP

Until 1983, when SIPP was introduced, the country's major source of information on income and program participation was the March Income Supplement of the Current Population Survey (CPS). However, the main statutory purpose of the CPS was to collect sufficient data to estimate unemployment rates, and CPS is not designed to collect the kinds of data that affect eligibility for program participation. Nor was it designed to collect general socioeconomic characteristics on the U.S. population. Further, aside from a minor feature of the CPS allowing the analysis of individual and household characteristics as they change over time, the CPS was

Copyright National Academy of Sciences. All rights reserved.

not longitudinal and hence the evolution of household income, poverty status, and program participation could not be tracked.[5]

The Department of Health, Education and Welfare initiated the predecessor of SIPP in 1975 in collaboration with the Census Bureau. The department spent the next five years testing and designing the Income Survey Development Program (ISDP), a longitudinal survey to collect general socioeconomic and social program participation information for a representative sample of the U.S. population. The first nationwide test of this new survey, which sampled approximately 2,000 households, occurred in 1978, and a representative national sample of 8,200 households was surveyed in 1979. In 1981, the Department of Health, Education and Welfare lost its funding for the Income Survey Development Program, and work halted. Work resumed in 1983, when the Census Bureau received funding to continue the renamed Survey of Income and Program Participation, which officially launched in October of that year.

The original design of SIPP focused on a nationally representative sample of households, in the civilian non-institutionalized population with every person age 15 or older in the household being interviewed. Four months after the initial interview, respondents were surveyed again; this process repeated over a 32-month period. Each February, a new group of respondents was interviewed, creating an overlapping series of survey panels.

During the late 1980s, SIPP suffered from erratic and sometimes limited funding. Although each panel was supposed to go through eight interviews, this often was not the case. The 1989 panel, part of which was folded into the 1990 panel, was only interviewed 3 times. Funding problems also limited panel sizes, and the difficulty in field operations caused by having multiple panels in the field at the same time were significant obstacles to data collection and quality. Each panel was supposed to comprise about 20,000 households, but that number was rarely achieved.

The Census Bureau redesigned the SIPP in 1996, with input from NASEM (1993). Instead of conducting 32-month panels, the Census Bureau switched to 48-month panels to improve the quality of longitudinal estimates and to provide better information for policy makers. Field representatives (FRs) interviewed respondents 12 or 13 times over the course of the panel—about three times per year—thereby retaining the rough every-four-month interview

[5] See https://www.census.gov/programs-surveys/sipp/about/sipp-introduction-history.html

Copyright National Academy of Sciences. All rights reserved.

schedule in the prior panels. In 1996 the Census Bureau also introduced computer assisted interviewing (CAI) which allowed for faster data processing and for real-time answer consistency checks.[6]

The difficulty in field operations from four-month interviewing and a variety of other factors led to increasing costs and reductions in interviewing and sample sizes. After much analysis and planning, SIPP was again redesigned in 2014, with input from the National Academies (see National Research Council, 2009). The catalyst for this redesign was SIPP's imminent cancellation. Though SIPP was saved by support from stakeholders, major changes were deemed necessary. The goals of this redesign were to reduce costs and respondent burden and to improve data quality and timeliness. Three major changes were that (1) sampled households would now be contacted only once a year rather than three times, (2) topical modules were eliminated, and (3) the survey instrument was completely redeveloped around an Event History Calendar (EHC) as a tool for collecting data on intra-year dynamics. The National Academies provided an assessment of that redesign and considered content changes for future improvement of SIPP (NASEM, 2018).

The basic design of SIPP has not changed appreciably since the 2014 revision, except that SIPP started using yearly overlapping panels in 2019.

WHAT MAKES SIPP UNIQUE AND VALUABLE

Compared to other countries, the United States is rich in datasets from household surveys that provide important information on the population and society. Despite this richness, the SIPP occupies a unique position in the landscape of social datasets that no other dataset can replicate.

Among datasets that are primarily cross-sectional in nature, the most heavily used alternative household survey is the CPS. As already noted, the CPS was not designed to be a general-purpose dataset collecting information on all economic, demographic, labor market, health, and social program participation domains of importance to understanding U.S. society.

[6] Taken from https://www.census.gov/history/www/programs/demographic/survey_of_income_and_program_participation.html [January 26, 2023].

Copyright National Academy of Sciences. All rights reserved.

The CPS collects only a small subset of the near-3,000 variables collected on individuals and households in each wave of SIPP. Another important household survey is the American Community Survey (ACS), which focuses on local areas with large enough sample sizes to obtain reliable statistics and hence serves a vital role for state and local policy makers. SIPP does not have the capacity to generate statistics at as detailed a level of geography as the ACS. But the ACS is cross-sectional and does not follow households over time. It also has a much smaller set of variables and could not be used to provide the national-level statistics on all the variables that the SIPP can. Finally, the National Health Interview Survey is the country's leading population-representative survey of the health status of U.S. individuals and arguably has better information on health and medical care than SIPP, but it does not collect anywhere close to the number of other domains as SIPP collects.

There are other social surveys that follow individuals or households over time, as SIPP does, and are hence longitudinal in nature. The National Longitudinal Surveys collected by the U.S. Bureau of Labor Statistics, for example, follow birth cohorts of individuals over time, charting their life courses with a particular focus on labor market outcomes. But these surveys only collect data on selected birth cohorts and hence are unsuitable for cross-sectional statistics on the U.S. population as a whole. The Health and Retirement Study, sponsored by the National Institutes of Health, also collects valuable health and economic information over time but only on the older population. The Panel Study of Income Dynamics is the nation's longest general-purpose social information survey, having started in 1968 and continuing to survey individuals who were alive at that time as well as all their children as they start their own families. This survey does not have the same sample size as the SIPP and hence cannot produce statistics for the same detailed subgroups, nor does it collect the detailed social program information or a number of other dimensions that SIPP does. But the uniqueness of the Panel Study of Income Dynamics is that it follows individuals over their entire life course, not just over four years, and it can be used to analyze intergenerational mobility. SIPP is superior in its cross-sectional dimensions and in its short-term dynamics over the typically four years of SIPP panels.

In summary, SIPP is unique because it combines three features: (1) an extremely large number of variables collected on all major domains of importance to U.S. individuals and households; (2) its high level of detail on social program participation across a large number of programs; and (3) its ability, afforded by its longitudinal dimension, to follow individuals and

Copyright National Academy of Sciences. All rights reserved.

households over time to chart the evolution of their status in the large number of domains on which it collects information. Further, by creating an almost-complete record in these dimensions over the past 50 years, SIPP is a unique source for studying historical trends in the country. In the panel's view, preserving these unique characteristics of SIPP deserves high priority.

Two other characteristics also contribute greatly to the value of SIPP, though they might not necessarily be unique. SIPP collects multiple waves of data within a single panel year, thus making a greater number of cases available for analysis in any given year. Moreover, if needed and the questionnaire items have not changed, SIPP data can be combined across multiple years, providing a greater capacity to study small groups. Larger surveys, such as the American Community Survey (ACS), may not have the same need to expand the number of cases for analysis, but SIPP's design makes it very efficient in providing data even for rare groups. SIPP also is well designed for making linkages with administrative data and even provides the SIPP synthetic beta in which the links have already been made (and synthesized). Other surveys also allow linkages, but SIPP's collection of data on program participation makes the potential for linkages especially useful.

SAMPLING DESIGN AND METHODOLOGY FOR SIPP

SIPP is a longitudinal survey, designed to be nationally representative of the civilian non-institutionalized population. In general, since 2014, four years of monthly data have been collected through annual surveys. While the data can be (and often are) analyzed cross-sectionally, focusing on a single point of time, the repeated observations of month-by-month data over 4 years support the analysis of change over time.

Similar to the 1990 through 1993 panels, beginning in 2019 the SIPP panels have been designed to be overlapping in time (with some exceptions). For example, the 2023 data collection includes the first wave of the 2023 panel, the second wave of the 2022 panel, the third wave of the 2021 panel, and the fourth (and final) wave of the 2020 panel. All four of these collectively are considered as the 2023 SIPP; that is, each SIPP is named based on the year of the data collection, not the date of the first wave of a panel. When an analyst wishes to focus on a particular year, all four applicable panels for that year are provided together, greatly increasing the amount of data available for analysis. Alternatively, an analyst can examine change over time

Copyright National Academy of Sciences. All rights reserved.

by combining multiple waves within a single panel. (Different weights are available depending on how the research is to be conducted.)

The primary sampling unit is the household, primarily based on the decennial census. Low-income households are oversampled. Within each household, all adults (age 15 or older) are interviewed, either directly or through proxies if a household member is unavailable. Information about household members under 15 years of age is collected via a proxy interview. In follow-up waves, the original set of adults is interviewed (including those who have moved to new addresses, with a few exceptions such as those living abroad or in institutions), along with any new adults living with the original sample members. New people who join a household are considered to be part of the SIPP panel for as long as they continue to live with an original sample member. People are considered members of the household if they are usual residents (e.g., sleeping there a majority of the time).

Since 1996, all first waves of a panel have contained at least 37,000 households. The 2020 SIPP sampled 57,834 designated Living Quarters, which yielded 50,352 eligible Living Quarters, 21,989 households that were interviewed, and 53,332 person interviews (SIPP 2020 User Guide).

Separate weights are provided or may be constructed for different types of analyses: monthly weights for a cross-sectional sample at a given month, calendar-year weights for samples that provide data in December of the reference year, and longitudinal weights for samples that provide monthly data for more than one consecutive year (with the potential for three different sets of longitudinal weights for reference periods of two, three, or four calendar years). The non-zero final person weights for SIPP 2020 Wave 1 ranged from 387.1 to 39,950.8.[7] The data files also include replicate weights and longitudinal replicate weights which are needed for correctly calculating standard errors, given the complex sampling design.

SIPP has traditionally relied on in-person household interviews (except for those who move out of an area, who are interviewed by telephone), but during the COVID-19 pandemic the mode was switched to telephone interviewing on March 19, 2020.

[7] Weights of zero are assigned when data are not available for all 12 months, or when the person was not part of the sample at Wave 1, with some exceptions.

Copyright National Academy of Sciences. All rights reserved.

To reduce respondent burden, SIPP groups some related people into clumps, limiting the number of people who must be asked certain questions. Identification of the clump is not provided in the public-use data.

SIPP contains different kinds of nonresponse. One form occurs when an eligible household is not interviewed at all, another when at least one person is interviewed but not everyone in the household is, and a third occurs when a respondent refuses to give a response or gives an invalid response to a particular question or set of questions (item nonresponse). For the third of these, while nonresponse rates are low for many variables, they are quite high for some important variables, like household income (49% nonresponse rate) (Tersine, 2022). For the first two types of nonresponse, the Census Bureau provides weights to correct for the missing responses. For the third type, the Census Bureau uses imputation to fill in a value for the missing item. For imputation, the Census Bureau uses a combination of model-based imputation, sequential hot-deck imputation, logical imputation, and cold-deck imputation.

THE CONTENT COLLECTED THROUGH SIPP

Since the 2014 SIPP, the reference year for which respondents are asked to report is generally the preceding year. For example, for the 2020 SIPP, questions were generally asked about 2019. Exceptions to this include questions on disability status and parent-child relationships.

The core of SIPP is the collection of data on income and program participation, including four years of employment data, assets, debts, retirement accounts, and types and amounts of other income including social insurance, social welfare, and other support paid and received such as alimony and child support.[8] SIPP also collects detailed information about health and disability, such as health insurance, functional limitations, health care utilization, and medical expenditures.

Other data collected by SIPP include adult well-being, education, language, education enrollment, demographic data, relationships among household members, citizenship and immigration status, marital history, dependent care, fertility history, biological parents' nativity and mortality, and child well-being. Some of these characteristics, such as demographics and

[8] https://www.census.gov/programs-surveys/sipp/about/sipp-content-information.html

Copyright National Academy of Sciences. All rights reserved.

citizenship, are collected once per wave, while other characteristics, such as residence, family, and household composition, marital status, and commuting and work schedules are collected for each month of the study period.

Topics Covered

Table 2-1 displays the basic types of data collected and provided by SIPP.

Table 2-1 Data Collected in SIPP 2020, by Broad Category

Broad category	Specifics
Demographics	
Basic	Date of birth, sex, race, Hispanic origin
Education enrollment	Months of enrollment, grade, type, full-time or part-time status, credential worked toward, whether a grade was repeated
Educational attainment	Highest level attained, high school diplomas through high school or GED, alternative credentials, professional certification, state or industry license, educational certificate, and organization awarding the certificate
Family and household relationships	Relationship to reference person, parent and child pointers, parent type, monthly relationship of everyone in the household, number in the household, number who are over 65, number who are under 18
Fertility	Number of children ever born, month and year of birth (month omitted from public-use file), multiple partner fertility, grandparent status, age at first birth, whether current union is a childbearing one, number of childbearing unions
Language	Whether speaking a language other than English at home, and if so what language is spoken; how well English is spoken; whether no one 14 or older speaks only English at home, whether no one speaks English very well
Marital status and marital history	Current marital status and identification of the spouse; number of times married; ever widowed or divorced; year of current marriage; year of first marriage; changes in marital status; nonmarital partner in household
Nativity, citizenship, and parent nativity	Born in the U.S.; state of birth; nativity status and country of origin for biological parents; country of birth; citizenship status; how the respondent became a citizen; immigration status on arrival; year of entry
Parent mortality	Biological parent still alive; month and year of death of parent (not in public-use file); month and year that parent was born (not in public-use file)
Residence	Where respondent lived; how long at each residence; whom the respondent lived with; why the respondent moved to the residence; tenure status; type of living quarters; receipt of housing assistance
Veteran status	Ever served on active duty in U.S. Armed Forces; periods of service
Assets, employment and earnings	
Assets	*Assets:* Retirement accounts (IRA and KEOGH, 401(k)/403(b)/503(b), defined benefit), interest-earning assets (government securities, checking accounts, savings accounts, money market, certificates of deposit, municipal and corporate bonds, educational savings accounts), other income-generating assets (stocks, mutual funds, rental property, annuities, trusts), other assets (other real estate, businesses owned as a job, businesses

Copyright National Academy of Sciences. All rights reserved.

	owned as an investment only, life insurance policies, primary residence, cars, recreational vehicles, other financial investments) *Liabilities:* Debts secured by assets (primary residence, rental property, other real estate, vehicle debt, recreational vehicle debt, debt on businesses owned as a job, debt on businesses owned as investment only), debts not secured by assets (credit cards and store bills, student loans and education-related expenses, medical debt, other) *Additional topics:* Rent and mortgage payments, utility payments
Commuting and work schedules	Means of transportation, distance to work, minutes to work, parking and toll expenses, additional commuting expenses, other job-related expenses
Employment and Earnings	When a job was held, different types of earnings, employment and business characteristics, reasons did not hold a job, whether looked for work
Parental leave	Whether worked for pay during pregnancy, whether worked right up to birth, type(s) of leave used after the child was born, how long stopped working prior to birth, type(s) of leave used prior to birth, worked at any time after the birth, how long after the birth they started working
Health and well-being	
Adult well-being	Characteristics of the household; characteristics of surrounding neighborhood; food security
Childcare	Arrangements for children under age 14 (e.g., grandparent care); includes parental, relative, and nonrelative arrangements, federal or employer assistance, and Head Start
Child well-being	Topics such as shared meals, reading, school engagement, outings, activities with nonresident parents
Dependent care expenses	While household member attended work, attended training, or looked for a job
Disability	Hearing difficulty, seeing difficulty, cognitive difficulty, ambulatory difficulty, self-care difficulty, independent living difficulty, development delays; physical, mental, or emotional problems that limit ability to: play with other children, do regular schoolwork, find or keep a job, the kind or amount of work possible
Health care utilization and medical expenditures	Health status; number of days sick in bed and hospitalized; prescription medication use; number of visits to dentists and medical providers; number of visits for uninsured respondents; health insurance premiums; expenses for medical care and supplies; over-the-counter medical expenses; Flexible Spending Account
Health insurance	Type of insurance, who plan holder is, whether the employer pays for premiums, how the plan was purchased, periods of coverage
Program participation and income transfers	
Annuity/life insurance retirement income	Whether lump-sum, each month of receipt, amount received over year
Disability income	Payments from insurance, employer payments, pension, Federal pension, state government pension, local government pension, U.S. military retirement pay, U.S. Government Railroad Retirement, Black Lung benefits, other disability income
Energy assistance	Whether paid separately for utilities, received energy assistance, assistance in form of checks, coupons or vouchers, or sent directly to utility or landlord
General assistance	Month and year, benefit owner, reason receipt began, reason receipt ended, payment amount
Lump sum severance pay/ retirement plan income	Pension/retirement pay, severance pay, deferred payment/final paycheck
Miscellaneous income	Charity, family or friends, roomers or boarders, estates, incidental or casual earnings, miscellaneous cash income, National Guard or Reserve Pay

Copyright National Academy of Sciences. All rights reserved.

Other assistance	Food assistance, transportation assistance, clothing and housing assistance, training assistance
Retirement income	Federal Civil Service or employee pension, local government pension, military retirement pay, National Guard or Reserve Forces retirement, other retirement income, pension from a company, state government pension, U.S. Government Railroad Retirement
School meals	School-provided lunch or breakfast, meals free or reduced-price
Social Security benefits for oneself	Retirement, disability, widowed, spouse, other
Social Security benefits on behalf of a child	Name of children, reason for benefits, monthly amounts
Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)	Month and year, benefit owner, reason receipt began, reason receipt ended, payment amount, type of WIC coverage
Supplemental Nutrition Assistance Program (SNAP)	Month and year, benefit owner, reason receipt began, reason receipt ended, payment amount, type of SNAP coverage
Supplemental Security Income	Month and year, benefit owner, reason receipt began, reason receipt ended, payment amount, type of SSI coverage
Support paid	Number of children outside the household, if paid support, amount of support, time spent with the children
Support received	Presence of and amount of support received for foster childcare, child support, alimony
Survivor income	Income in each of 13 categories (e.g., Black Lung Benefits, Worker's Compensation) for each person 15 years and older who is a widow/widower
Temporary Assistance for Needy Families (TANF)	Name of benefit owner, why receipt began and (if applicable) ended, payment amount, who is covered
Tax returns	Whether filed tax return, whether plans to file return, filing status, whether the respondent was claimed as a dependent, whether received an Earned Income Tax Credit
Unemployment compensation payments	Whether received, type of payment, months of receipt, amount received
Veterans benefits	Whether received, type of benefit, disability rating (if applicable), monthly amounts
Worker's compensation	Months and amounts received
Income and poverty recodes	
Income recodes	Recodes aggregate income in various ways from data supplied in the questionnaire (not new data being collected)
Poverty recodes	Calculated poverty status, based on household and family composition and aggregate income recodes (not new data)

SOURCE: 2020 Survey of Income and Program Participation: User's Guide; 2020 Survey of Income and Program Participation (SIPP): Public Use Instrument Specifications.

The broad categories listed above sometimes hide considerable detail. For example, the adult well-being section in 2020 consisted of 11 questions asking about conditions such as cracks in the ceilings or walls, holes in the floor, problems with pests, plumbing problems, street noise or traffic problems, trash in the streets, staying home for safety reasons, neighborhood safety from crime, difficulty in paying the rent or mortgage, and difficulty paying utility bills.

Over the period 1984–2008, SIPP also contained topical questions that were not repeated in each wave, typically in separate sections but sometimes intermixed with related core questions

Copyright National Academy of Sciences. All rights reserved.

when that improved the question flow.[9] Since the focus of the consensus panel is on disclosure avoidance for the current survey, the prior topical modules are not discussed here in detail.

One of the unusual features of SIPP is that it collects data on each member of the household and their relationships with all other household members. This differs from other household surveys, such as the ACS, which collect data on relationships to the householder only. A total of 20 different relationship types are covered for each household member (see Box 2-1). Thus, though the core of the SIPP data consists of data on income and program participation, other SIPP data can be used to address a wide variety of sociological issues, such as the relationship between health conditions and divorce behavior, transitions in parental presence among children, and (based on a different section of the questionnaire) the prevalence of multiple-partner fertility.

BOX 2-1

Relationship Categories Used in SIPP

1 Opposite-sex husband/wife/spouse
2 Opposite-sex unmarried partner
3 Same-sex husband/wife/spouse
4 Same-sex unmarried partner
5 Biological parent/child
6 Step-parent/child
7 Adoptive parent/child
8 Grandparent/grandchild
9 Biological sibling
10 Half sibling
11 Step-sibling
12 Adopted sibling
13 Other sibling
14 Parent/child-in-law
15 Brother/sister-in-law
16 Aunt, uncle, niece, nephew

[9] https://www.census.gov/programs-surveys/sipp/tech-documentation/topical-modules.html

Copyright National Academy of Sciences. All rights reserved.

17 Other relative
18 Foster parent/child
19 Other nonrelative
99 Self

SOURCE: Table 3-7, 2020 Survey of Income and Program Participation Users' Guide.

The public-use file is not designed to replicate the survey data on a variable-by-variable basis. Some public-use variables have value ranges that differ from those in the original data, some variables are combinations of other variables, and prior to the 2014 SIPP panel some variables were in topical modules rather than in the core wave files.

Selected Studies Show the Range and Depth of SIPP

While the core of the SIPP questionnaire is devoted to financial measures, the broad range and specificity of the data collected has supported a wide range of research. A Google Scholar search (in January 2023) for "Survey of Income and Program Participation" produced over 20,000 results, though those results include citations in addition to independent analyses of SIPP data. Following are illustrative uses of the data that, while far from comprehensive, show the range and depth of research conducted using SIPP.

Economic Data

As might be expected, SIPP has often been used to examine various aspects of the U.S. economy.

- *Occupations and earnings*. The Census Bureau, as part of its Current Population Reports, regularly publishes data on occupation, earnings, and job characteristics, using SIPP as a primary source. The data include what occupational categories workers are employed in, their demographic characteristics, their employment characteristics (full- or part-time employment, holding of multiple jobs simultaneously, sector of employment), work schedule, work-from-home status, median earnings, and health insurance coverage (Gumber & Sullivan, 2022).

Copyright National Academy of Sciences. All rights reserved.

- *Poverty dynamics*. The Census Bureau also uses SIPP to publish data on poverty, including the monthly poverty rate overall and by demographic characteristics, as well as rates of episodic and chronic poverty (Warren & Tettenhorst, 2022).
- *Income sources of older households*. Another Census Bureau report describes the sources of income for older households, including social programs, pensions, and private retirement savings. Reporting on each household income decile, it describes the income ranges, receipt rates of six types of income, the mean income from each type, and the distribution of income types, along with supplementary statistics on certain categories of households (Thompson & King, 2022).

Monitoring of federal programs

The federal government uses SIPP to examine the usage of its programs.

- *Supplemental Nutrition Assistance Program (SNAP*). Over 90 percent of children receiving SNAP benefits also received at least one other form of assistance. The programs include Medicaid; Children's Health Insurance Program; Special Supplemental Nutrition Program for Women, Infants, and Children; Temporary Assistance for Needy Families; Supplemental Security Income; and Social Security (King & Giefer, 2021).
- *Student debt*. Students with student loan debt are more likely than other students to have other types of household debt, with the debt rates described in terms of the rates and amounts of debt, the probability of debt, and the predicted amount of debt by type, student loan status, and student demographic group (Bennett & King, 2022).
- *Joint federal and state programs*. Some federal programs, such as Medicaid, the State Children's Health Insurance Program (SCHIP), and Temporary Assistance for Needy Families (TANF), are administered in part by the states, with different states setting different eligibility requirements. These differences in part can provide a testing ground for evaluating the impact of changes in federal policies, and they are also important when monitoring the impact of federal programs. SIPP is a valuable tool for evaluating such differences (Irvin & Czajka, 2010). For such evaluations and monitoring, state-level identifiers are required.

Copyright National Academy of Sciences. All rights reserved.

Uses by Policy Researchers

Policy researchers use SIPP to examine the need for assistance and the impact of policies and programs.

- *Housing and health quality*. Boch and colleagues (2020) examined the association between poor housing quality and adult health outcomes, adjusting for socioeconomic factors and contextual housing characteristics.
- *Health care*. Carman, Liu, and White (2020) calculated the total costs to households in paying for health care, finding that higher-income households paid the most for health care, but payment as a share of income was greater among lower-income households.
- *Education and earnings*. Carr and Wiemers (2020) found an increase in inequality in long-run earnings for both men and women and a resulting decline in intergenerational mobility.
- *Immigration policies and health*. Examining differences in state policies concerning immigration, Dondero and Altman (2020, p. 1) found that "restrictive immigrant policy climates exacerbate nativity gaps in health provider visits among working-age adults," and that "even immigrant policies not directly related to health have consequences for immigrants' health care utilization."
- *Extreme poverty.* Shaefer and Edin found an increase in extreme poverty between 1996 and 2011 among U.S. households with children, and that adding SNAP benefits reduced the number of such households by 48 percent. Also, accounting for refundable tax credits and housing subsidies reduced the number by 63 percent (Shaefer & Edin, 2013).
- *Occupational licensing*. Gittleman, Klee, and Kleiner found that "those with a license earn higher pay, are more likely to be employed, and have a higher probability of employer-sponsored health insurance offers" (Gittleman et al., 2018, p. 57).
- *The safety net and material hardship*. McKernan, Ratcliffe, and Braga (2021) estimated the impact of participation in TANF, SNAP, or Medicaid/SCHIP on the number of material hardships and the incidence of food insufficiency.
- *Minimum wage*. Comparing states that were bound to increases in the federal minimum wage to states that were not, Clemens and Wither (2019) estimated that the

Copyright National Academy of Sciences. All rights reserved.

minimum wage increases reduced employment rates by at least half a percentage point in the states that were bound.

- *Federal program evaluations.* In addition to academic research, SIPP is a key input for models developed by federal agencies to evaluate programs. The Modeling Income in the Near Term (MINT) model used by SSA is a dynamic micro-simulation model that projects into the future a database consisting of SIPP data matched to Social Security administrative records to evaluate various SSA programs (Smith et al., 2021). In a joint statistical project, the Economic Research Service (ERS) and Census have examined the SIPP linked to the administrative SNAP data (Scherpf et al., 2015), and with the Food and Nutrition Service (FNS), Mathematica developed the MATH SIPP+ simulation model to evaluate SNAP eligibility (Leftin et al., 2014). The Congressional Budget Office (CBO) has also created simulation models using SIPP data (see NASEM [2018], pp. 33–39).

Sociological Research

The breadth of data offered by SIPP also supported sociological research, not always with financial data.

- *Parental presence*. Scherer and Mayol-Garcia (2020) examined the demographic and socioeconomic characteristics of children who experienced at least one transition in parents (including changes in the number of resident parents, changes in which parent was coresident, and changes in a parent's cohabiting partner).
- *Multiple-partner fertility research*. Defining multiple-partner fertility as someone who has had biological children with more than one partner, Monte (2017) found that 10 percent of all adults have had children with more than one partner, as have 21 percent of all parents with two or more children.
- *Leave usage after first birth*. Scherer (2022) found that leave usage among men after the birth of their first child increased over time, and it was more common among men with higher levels of educational attainment.
- *Children in shared households*. Pilkauskas and Cross (2018) found a more than 4 percentage point increase from 1996 to 2016 in the percentage of children who lived in three-generation households.

Copyright National Academy of Sciences. All rights reserved.

DATA PRODUCTS FROM SIPP

Census Bureau Reports

The Census Bureau's primary method for publishing estimates is through its P-70 series.[10] These reports provide access to key SIPP data at a summary level and may be based on the internal data files, without top-coding and other confidentiality protection techniques. They are topical in nature, consist of text, tables, and graphics, and are highly variable in length (recent reports range from 6 pages to roughly 100 pages).

The Census Bureau also publishes table packages (found at Survey of Income and Program Participation Table Packages (census.gov)), which are sets of tables about a particular topic. For example, recent releases include:

- A Child's Day (with tables about children's extracurricular activities and school outcomes),
- State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2019,
- Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2019, and
- Survey of Income and Program Participation (SIPP) Detailed Program Receipt Tables: 2019.

These most recent packages consist of between 2 and 17 tables, and unlike the P-70 reports they consist entirely of tables. They can be used as reference tools for providing key statistics without independent data analysis.

Finally, Census publishes working papers based on SIPP.[11] These tend to be more analytical than the P-70 reports and may include modelling, experimental designs, and other advanced statistical analysis. In length, recent (2022) papers range from a single page (e.g., as for a poster session) to 35 pages. These papers are designed to examine an individual topic in depth, and they are more likely to be cited in the same way that journal articles are than to be used as reference tools for looking up key statistics.

[10]The issue briefs can be found at: https://www.census.gov/programs-surveys/sipp/library/publications.html

[11]The working papers can be found at: https://www.census.gov/programs-surveys/sipp/library/working-papers.html

Copyright National Academy of Sciences. All rights reserved.

Microdata for Data Analysis

For those users who wish to conduct customized data analyses, the Census Bureau provides three options: through public-use files, using SIPP Synthetic Beta, and using the Federal Statistical Research Data Centers (FSRDCs).

Public-use Files

Anyone may download the SIPP public-use data from the SIPP data website.[12] No registration is required to download the data, so it is not possible to directly track users. The number of downloads is potentially countable, but the same person may download the data multiple times or may download the data without using them. Public-use files have some data suppressed and some data altered (i.e., by top-coding or bottom-coding). The public-use metadata report for the 2020 SIPP lists 2,873 variables (not counting status flags,[13] which exist for almost every variable), though data for 414 of these variables were suppressed. The downloadable file contains 5,026 variables and 622,339 observations. Note that SIPP has a separate record for each of the 12 months and for almost every household member, so the number of observations is much greater than the number of households. After downloading the full file, a data user might extract only those records and variables of interest, perhaps also creating summary measures based on combining multiple records.

SIPP Synthetic Beta (SSB)[14]

The SSB combines data from nine SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, and 2008) with linked administrative tax and benefit data from the Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt. Earlier SSB files were partially synthetic, but the most recent version is fully synthetic. Census describes the process in the following way:

> The synthesis process involves estimating the joint distribution of all the variables in the data and taking random draws from this modeled distribution. These draws are then used to replace actual data values. This process is repeated multiple times to create a set of 4 files, which are called implicates.[15]

[12]https://www.census.gov/programs-surveys/sipp/data/datasets.html
[13]Status flags are used to indicate whether and how a particular value was imputed.
[14]SIPP Synthetic Beta (SSB) is described at https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html
[15]https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

Copyright National Academy of Sciences. All rights reserved.

Some advantages of using the synthetic data are that confidential linked administrative data are available that cannot be released publicly, and editing has been performed to make the SIPP data more consistent with the administrative data. The data therefore require less data preparation than would otherwise be the case. On the other hand, the synthetic beta contains only 141 variables, so it has only a small fraction of the data collected through SIPP.

Data users accessed the SSB through an online server, making the data potentially available to all, that is, without requiring travel to a special location. The server is currently unavailable as it transfers from its past host institution (Cornell). The access process requires that users first must submit an application to use the server, providing contact information, a brief description of the proposed project, and a list of the variables to be used. File access is approved or denied solely based on the feasibility of the proposal—that is, whether the needed data are available on the SSB. There are no requirements for a background review of the applicants. Once approved, access to the server is free. Findings based on the synthetic data do not require disclosure review, because the data are considered public-use.

Because the data have been synthesized, statistical results may differ from the original data. Data users are encouraged to have their findings verified by requesting the Census Bureau to rerun their final programs against the Gold Standard File.[16] According to the Census Bureau, it has received 139 validation requests since 2012 from a total of 275 SSB users since 2007.[17] It is not known how many of the 139 requests were from unique users. Thus, the number who made validation requests is at most roughly half the number of users. While no data are available on the number of requests between 2007 and 2012, about two-thirds of the recorded validation requests were received in 2017 or later, suggesting a pattern of increased use over time, in which case the number of requests prior to 2012 (when Census started keeping records) is likely to have been small. Moreover, given that some users may make multiple validation requests, it appears that many SSB users never make validation requests. A separate data collection conducted by this panel found that 11 of 18 SSB users made validation requests, though the respondents to the data collection request might not be representative of all SSB users.

[16]The Gold Standard File is a file created by the Census Bureau containing the original (nonsynthesized) data as a step toward producing the synthetic data. It is also used to verify whether statistics based on the synthetic data are consistent with those using the original data. It is not a master file of all SIPP original data, but rather was created specifically to produce the SIPP synthetic data.

[17]Email communication with the authors from Holly Fee, Aug. 16, 2022.

Copyright National Academy of Sciences. All rights reserved.

The process of validation can be burdensome both for data users, who must prepare a package for the Census Bureau, and for the Census Bureau, which must perform the validation run and then review the results for potential disclosure risks. As part of the disclosure review process, Census also examines the number of requests submitted by a user, looking at whether the collective set of runs may present a disclosure risk that would not be present in a single run. The total Census review process typically ranges from 2 to 15 hours; a recent instance that involved a lot of back-and-forth with a user involved about 60 hours of analyst time.[18] The Census Bureau has declared a desire to move to a system in which the validation is fully automated.[19] The validation process also has been subject to disruption. In August 2022, the Census Bureau announced that the synthetic beta server operating at Cornell University would be shutting down on September 30, 2022, and that new validation requests would not be accepted after September 2, 2022. The Census Bureau has declared an intention to find another server location, with the location and timing undetermined at the time of this writing.

Federal Statistical Research Data Centers (FSRDCs)

A third way to access the SIPP data is at FSRDCs. These research data centers are based on partnerships between universities, nonprofit research institutions, and government agencies. As of 2023, SIPP data were available for the 1984–2021 panels.[20] The Census Bureau lists 33 open FSRDCs, located throughout the country.[21]

The advantages of accessing the data through FSDRCs are that they are more detailed than the public-use files, with no or minimal top-coding, uncollapsed variables, and finer geographic detail, and that they allow access to some linked data, such as SNAP data. The data remain stripped of personally identifiable information, instead having a Protected Identification Key.

Access to FSRDCs requires multiple steps. The researcher must attain Researcher Special Sworn Status, which requires a background check, a lifetime oath of confidentiality, and completion of yearly training to maintain the status. Until the COVID-19 pandemic, to access the data users also were required to be physically present at an FSRDC site. Each of these sites has a

[18]Email communication from Holly Fee, Sept. 12, 2022.
[19]Briefing by Rachel Shattuck, June 30, 2023.
[20]https://www2.census.gov/programs-surveys/sipp/2015/2015%20PAA%20-Using%20SIPP%20in%20the%20FSRDCs.pdf
[21] https://www.census.gov/about/adrm/fsrdc/locations.html

Copyright National Academy of Sciences. All rights reserved.

secure connection to the Census data facility in Bowie, MD, where special security conditions are in place; for example, no Internet access is allowed and printing is strictly controlled. However, a pilot extension allowing approved unpaid Special Sworn Status supporters to work from their personal residences with virtual access to Census Bureau data has now been indefinitely extended,[22] although linked data from other agencies, such as the Social Security Administration and the Internal Revenue Service, would be subject to the requirements set by those agencies.

Additionally, there are limits on what analyses may be performed; tabulations are discouraged, with model-based analyses preferred instead. The analyst must submit a proposal that defines the research agenda, reviews CES Working Papers, states the benefits to the Census Bureau, justifies the data requirement, and provides a project timeline. Census reviews the proposal based on scientific merit, the need for nonpublic data, the benefit to Census Bureau programs, feasibility, whether a sensitive population is involved, and the absence of a disclosure risk to an individual or business. The Census Bureau states that approval requires a maximum of 12 weeks unless approval by additional data providers is needed, after which the data requester must obtain a moderate security clearance.[23]

The Census Bureau does not charge a fee for use of its data. However, the research organizations that provide access may charge various types of fees.[24] For example, the Berkeley FSRDC charges hosting fees of between $15,000 and $20,000 per year for new users.[25] If an FSRDC is not nearby, researchers may also face transportation costs, though the recent allowing of virtual access for projects using stand-alone SIPP data (i.e., with no administrative record linkages) removes this potential cost.

The process for applying for access was revised and simplified in December 2022 with the creation of a standard application process through an online portal.

As with the FSRDC access to SIPP linked to administrative data, the federal agency projects (e.g., Social Security Administration, Food and Nutrition Service, and Economic Research Service) use similar agreements or joint statistical projects and operate under similar restricted access and disclosure procedures. The Social Security Administration has had special

[22] Jan. 12, 2023 email from Michael Snow, DSEP coordinator to David Johnson.
[23] https://www.census.gov/topics/research/guidance/restricted-use-microdata/standard-application-process.html
[24] https://www.census.gov/about/adrm/fsrdc/about/fsrdc-network-fees.html
[25] https://dlab.berkeley.edu/data/fsrdc-data-and-guidelines

Copyright National Academy of Sciences. All rights reserved.

agreements to collect supplemental data for use in their MINT model. It is expected that these will continue under the current agreements.

DISCLOSURE AVOIDANCE IN SIPP FROM THE 1990S TO THE PRESENT

SIPP disclosure avoidance procedures are described in Box 2-2. Separate procedures are considered for data published in tables and for data released in microdata files.

BOX 2-2

SIPP Disclosure Avoidance Procedures

Tabular data

Due to the small sample size with large weights, SIPP tabular data are published for very large areas for data quality reasons. No additional disclosure avoidance techniques were deemed necessary to protect the data.

Microdata

Removal of direct identifiers. The Census Bureau removes direct identifiers from the file such as name, address, and phone number.

Geographic threshold*.* All geographic areas identified must have a population of 250,000 or more. When calculating this population, all geography-related variables in the file are cross-tabulated to obtain the final population count of an area that can be identified as a piece of geography. This population threshold is larger than the 100,000 threshold used for the Current Population Survey and the American Housing Survey. The Microdata Review Panel, predecessor of the Disclosure Review Board,* determined in the early 1990s that SIPP should have a higher population threshold than some other surveys because the Public Use Microdata File contained a very large amount of detailed income items not available in most surveys, and the panel was concerned about an inflated disclosure risk.

After that ruling, the Census Bureau could not publish a SIPP Public Use Microdata File for all states, because Primary Sampling Units were public information, and when crossing those areas with all geographic variables on the Public Use File, some states did not make the 250,000 threshold and had to be combined. A few years later, the Census Bureau decided not

Copyright National Academy of Sciences. All rights reserved.

to reveal Primary Sampling Units to the public for any survey, so all states can now be identified.

Top-coding and bottom-coding. The Census Bureau uses top-coding and bottom-coding to eliminate outliers in a file for continuous variables such as wages and salary. A top-code (cut-off) is in place for 0.5 percent of all values or 3 percent of all nonzero values, whichever is the larger of the two. Bottom-codes are the same except on the other side of the distribution. A bottom-code might be applied to gross income. For variables that are part of a sum, the individual summands are top-coded prior to their summation. Originally, all top-coded values were replaced with the top-code cut-off itself. Beginning in 1996, the top-coded values were replaced with the mean of the top-coded values. However, a small number of variables still use the top-code cutoff itself (rather than the mean). At least three values from three different respondents must be top-coded or the top-code is lowered to meet this requirement. Bottom-codes are the same except on the other side of the distribution. A bottom-code might be applied to gross income or years received for a given benefit. For variables that are part of a sum, the individual summands are bottom-coded prior to the summation.

Rounding/recoding. Each category of a categorical variable must contain at least 10,000 weighted people or households (depending on the universe for that particular variable) nationwide. If a category does not meet this threshold, it must be combined with other categories until it does. Some variables' categories, such as language spoken at home and immigration status, are combined even more. Dollar amounts must follow one of two rounding/recoding schemes. Round to two significant digits, or use this recoding scheme:

- Zero rounds to zero.
- 1 to 7 rounds to 4.
- 8 to 999 rounds to the nearest multiple of 10.
- 1,000 to 49,999 rounds to the nearest multiple of 100.
- 50,000 and greater rounds to the nearest multiple of 1,000.

Any totals or other derivations are calculated using the rounded numbers.

Noise infusion. Noise infusion is used to hide very unusual characteristics of a person or household at a given point in time. For example, consider a woman with sextuplets or a 10-year-old in college or a household with 13 people in it. Such unusual circumstances are often well known and sometimes in the news. Census Bureau editing procedures capture and alter many,

Copyright National Academy of Sciences. All rights reserved.

but not all, of these types of unusual circumstances. Unlike many Census Bureau surveys and censuses, SIPP collects month of birth, but noise is added to this variable. In addition, SIPP is longitudinal and changes in personal or household characteristics can often be found in public records. For example, a birth, death, marriage, or divorce would be reflected in the SIPP while a given household is in the sample. The Census Bureau does not publicly release the details of how noise is added to protect these types of data that pose a disclosure risk.

SOURCE: L. McKenna, *A History of the Survey of Income and Program Participation and Disclosure Avoidance* (Washington DC: Research and Methodology Directorate, U.S. Census Bureau, 2019) https://www.census.gov/library/working-papers/2019/adrm/CED-WP-2019-006.html

*NOTE: Census Bureau data products must be approved by the DRB before dissemination.

Administrative Data Added to SIPP

Administrative data are merged with SIPP data for three purposes: to support imputation of missing data, to verify and sometimes correct survey responses, and to provide supplemental data not contained in the SIPP data (e.g., the entire earnings history, not limited to the four years covered by SIPP).

No administrative data are directly included in the public-use files; when imputation is performed, the imputed values are based on statistical models rather than on directly inserting the administrative data. The SIPP synthetic beta (described above) differs by including administrative data in synthetic form, including sometimes by replacing survey values with the values in the administrative data. SIPP data housed at the FSRDCs offer the most opportunities for merging with administrative data or external data sources. Additional information on these three tiers of access is presented later in this chapter.

Following are some examples of how data linkages have been used analytically:

- The Census Bureau uses administrative records to impute missing data for selected variables in SIPP.[26] For example, SSA has data on work history, W2 earnings, self-employed earnings, deferred earnings, job count, benefits received,

[26] B. Gurrentz, Model-based imputation & administrative records in SIPP processing, presentation to panel on June 30, 2022.

Copyright National Academy of Sciences. All rights reserved.

payment history, and receipt of Supplemental Security Income (SSI); the IRS has data on earnings, number of W2 jobs, deferred earnings, and firm characteristics; and Numident provides birth dates.

- The Census Bureau uses administrative data from SSI to evaluate and correct for respondents' confusion between SSI and Social Security income, making a combined modeled and logical edit.[27]
- Capps and colleagues (2018) combined data from SIPP, which identifies unauthorized immigrants, with data from the ACS, which does not, to examine how many unauthorized immigrants reside in the country, where they reside, and the conditions in which they live.
- Carman and colleagues (2020) created a synthetic population, combining data from SIPP, the Medical Expenditure Panel Survey, Kaiser Family Foundation/Health Research Educational Trust Employer Health Benefits Survey, ACS, and National Health Expenditure Accounts. They then estimated payments at the individual level for financing health care services and the dollar value of the services received.
- Eggleston and Reeder (2018) combined data from SIPP and IRS 1040 tax data to perform a methodological study measuring the accuracy of respondents' self-reporting depending on whether or not they consulted financial records, also looking at the impact in terms of total duration of the SIPP interview.
- Hryshko and colleagues (2017) combined SIPP and Social Security earnings to examine inequality in the earnings of married couples and the instability of their combined earnings.
- Maples (2017) examined combining SIPP and ACS data to improve state estimates of disability.
- Kim and colleagues (2018) match SIPP data with longitudinal earnings records from administrative tax information to assess the predictive power of variables such as occupation, education, and short-term earnings on cumulative earnings over 20 years.

[27] B. Gurrentz, Model-based imputation & administrative records in SIPP processing (presentation to panel on June 30, 2022).

Copyright National Academy of Sciences. All rights reserved.

HOW SIPP DATA ARE USED

An earlier National Academies panel commissioned a literature search to determine the topics examined using SIPP data with the intention of assessing the 2014 redesign of SIPP. This panel has a somewhat different focus, being interested in how the selection of topics might affect disclosure avoidance and how the methods used in the statistical analyses affect how the data need to be accessed.

Topics Covered in Research Using SIPP

Table 2-2 provides the results from the earlier NASEM study associated with the SIPP redesign. Though the bulk of the references concern economic applications, the results show a wide variety of research topics.

TABLE 2-2 Survey of Income and Program Participation Bibliographic References, 2000–2014, by Topic

Topic	Number of Citations
Employment	84
Disability	64
Health, Health Care, and Insurance	48
Household Structure	46
Retirement	39
Immigrants	31
Earnings	27
Welfare Reform	27
Income Dynamics	25
Wealth	24
SCHIP/CHIP	23
Childcare	21
SSI	19
Poverty	18
Survey Methodology	17
Program Eligibility	16
Education	15
Material Hardship	15
Program Participation	12
SNAP/Food Stamps	14
Medicaid	11
TANF/AFDC	11
Transitions	11
Unemployment	10

Copyright National Academy of Sciences. All rights reserved.

Child Welfare	5
WIC	4

SOURCE: The 2014 Redesign of the Survey of Income and Program Participation: An Assessment, National Academies of Sciences, Engineering, and Medicine 2018, Table 3-1, p. 29.

It is evident that SIPP data are highly valued for analyzing the U.S. population. The wide variety of topics covered by SIPP and the wide range of studies that have been conducted demonstrate its immense contributions. The three unique features of SIPP described here—the large number of individual and household characteristics, the detail on social program participation, and its longitudinal nature, allowing studies of the evolution of individuals and households—are not available in this combination in any other dataset and are responsible for these contributions.

It is now recognized that these contributions come at an increasing risk of disclosure of private information to outside parties. In the rest of the report, we describe the nature of this problem and what steps might be taken to reduce disclosure risk, if necessary, while preserving the main benefits of the SIPP and preserving the utility and the usefulness of the data for understanding U.S. society. We first discuss how to assess whether the current public-use files released by the Census Bureau run the risk of unacceptably high levels of disclosure, and draw conclusions and recommendations on this issue. We then discuss methods for limiting disclosure risk, if and when it is found or believed to be at an unacceptably high level. We also discuss methods for providing access to users in alternative ways if they cannot be provided in public-use form. After discussing these topics, we end with a discussion of how to balance the utility and usability of the SIPP with methods for reducing disclosure risk.

CONCLUSIONS

Conclusion 2-1: The Survey of Income and Program Participation has made major contributions to the understanding of U.S. society and to the evaluation of the effects of government policies across an extremely wide variety of domains.

Conclusion 2-2: There is great value in a public-use data file of the SIPP, which appears to be the most used mode of SIPP data access, and in including at least the minimum

Copyright National Academy of Sciences. All rights reserved.

number of variables needed across a broad set of research topics. The lack of a public-use file would result in tremendous loss in research and a reduced number of users, and ultimately it would lower the support for SIPP.

Conclusion 2-3: Accessing restricted data through physical or virtual presence at an FSRDC is extremely time- and process-intensive and costly, making access prohibitive to researchers who are not working in locations with FSRDCs and those who do not have the financial resources to pay fees when required, resulting in inequitable access to the data.

Copyright National Academy of Sciences. All rights reserved.

3
Measuring of Disclosure Risk and Ways of Assessing It

This chapter first provides an overview of the way disclosure risk is measured, followed by a look at the specific disclosure-risk challenges that SIPP faces.

TYPES OF DISCLOSURE RISK

Disclosure risk can be defined in multiple ways and forms, but the most common types include *reconstruction, re-identification,* and *membership inference*. Disclosure risk is a major concern in data privacy and confidentiality when it comes to sharing SIPP data for research or public-use purposes. If not managed properly, sensitive information, such as income, health insurance status, and fertility history, could be revealed, which could have significant consequences for the individual involved.

Reconstruction. First proposed by Dinur and Nissim (2003), this type of disclosure risk occurs when an intruder takes all of the published statistics and solves a large system of equations or an optimization problem to reconstruct the underlying microdata as well as possible. Intuitively, if the number of published statistics is larger than the number of unknowns in the underlying microdata and it is sufficiently accurate, reconstruction is very likely to be effective (because it involves solving a system with more equations than unknowns). Once the underlying microdata are reconstructed, reidentification attacks can be applied.

Re-identification. When microdata (datasets where each record corresponds to an individual data subject) are released, the most immediate threat is re-identification, which can occur when an attacker can identify a specific individual with one of the released data records. This is typically done by matching the released record on a set of characteristics that make the individual unique and are known to the attacker through auxiliary datasets or by direct knowledge about the specific individual. For example, combinations of indirect identifiers, such as state of residence, age, race/ethnicity, sex, education attainment, and employment status, could

Copyright National Academy of Sciences. All rights reserved.

be known to the intruder and be unique in both the sample and the population. Occupation is also collected at a highly granular level in SIPP; corresponding general external databases lack the same level of granularity, but the highly granular SIPP data still may be useful in identifying individuals. This is particularly a risk when data are highly granular and some combinations are particularly rare; for example, Asian males are relatively uncommon among flight attendants (Ruggles et al., 2023), so adding just a few more characteristics may uniquely identify such a person. By re-identifying an individual, an attacker learns other previously unknown characteristics about the individual. There is a large research literature showing that re-identification risk is pervasive in datasets containing many attributes per individual, and virtually any attribute can be used as a potential identifier (Narayanan & Shmatikov, 2010). There are also computational and statistical methods for profiling individuals in data collected and released over time, which increases the risk for longitudinal studies such as SIPP (Tournier & de Montjoye, 2022). See below for more on how re-identification risk can be estimated.

Membership inference. Yet another kind of attack on aggregate statistics is a membership inference attack, whereby an attacker uses published statistics and data about a particular individual to determine with high statistical confidence whether an individual was in the dataset used to compute the statistics. The possibility of membership inference attacks demonstrated by Homer et al. (2008) led the National Institutes for Health to restrict access to genomic summaries in the Database of Genotypes and Phenotypes (dbGAP) (Couzin, 2008). The reason that membership inference is a concern is that sometimes being in a dataset, or a particular subset of a dataset, reveals sensitive information, such as when the dataset consists of individuals diagnosed with a particular disease (a "case group"). More recently, membership inference attacks have been developed against machine learning models (e.g., Carlini et al., 2022; Shokri et al, 2017). With black-box access to a machine-learning model, an attacker may be able to determine, with high statistical confidence, whether an individual was in the dataset used to train the model.

Methods for performing all these kinds of attacks are becoming increasingly sophisticated with time.

A different kind of disclosure risk appears when a respondent reveals that they completed the survey. This allows a more direct search for a person in the sample with particular characteristics, potentially removes the uncertainty often associated with statistical sampling as to whether a person with a unique combination of characteristics is actually that particular

Copyright National Academy of Sciences. All rights reserved.

respondent, and negates the reduction in privacy-loss that sampling offers differential privacy. Such self-disclosure can have the effect of negating the disclosure avoidance protections established by the Census Bureau. Sometimes surveys have encouraged people to acknowledge their participation in a survey as a way of encouraging others to participate; in this case it may be better for the Census Bureau to advise people of the risks of self-disclosure.

MEASURES OF DISCLOSURE RISK

Disclosure risk can be measured using a probabilistic disclosure risk measure (Duncan et al., 2011; Duncan & Lambert, 1986, 1989; Lambert, 1993; McClure & Reiter, 2012) that refers to the probability that an individual's sensitive information is revealed or inferred, or that an individual is re-identified based on attributes and patterns in a dataset through the usage of statistical, machine learning, or computational techniques. The disclosure risk can be quantified or measured using either an absolute or a relative scale. Absolute disclosure risk measures how likely it is that the intruder learns the sensitive information about a respondent in a dataset from some released information, given the intruder's prior knowledge (see examples given in Chapter 2). Relative disclosure risk measures the intruder's relative gain in ability to identify the respondent, beyond the intruder's prior knowledge. Though both measures are useful in assessing risk, the relative disclosure risk can be controlled by the data producer, who only has the decision to release or not release a particular dataset. On the other hand, the absolute disclosure risk provides additional information on what the level of risk is after release. For example, it will show the data releaser whether the risk after release is 1 in 100 versus 1 in one million.

When microdata or individual-level data are released and the intruder's goal is to re-identify a target individual in the release data, probabilistic re-identification risk can be measured as the probability of correctly identifying a "sample unique" (i.e., an individual in the sample data who has a unique set of attribute values) given a set of variables considered as quasi-identifiers. Probabilistic disclosure risk measures, which are often estimated through statistical modelling, are interpretable, and we can estimate the risk for each individual and every sensitive attribute in a dataset. Based on probabilistic disclosure risk assessment results, individuals at high risk of disclosure can be identified and corresponding disclosure limitation methods can be developed to reduce the risk.

Copyright National Academy of Sciences. All rights reserved.

Assessments of both types of disclosure risks, absolute and relative, require making assumptions about the intruder's prior knowledge and specification of the intruder's model for learning an individual's sensitive information (the re-identification attack method discussed below necessarily makes just such assumptions). This implies that the probabilistic disclosure risk measurements would vary by the assumptions and the intruder model specifications. In addition, it cannot be guaranteed that probabilistic disclosure risk can be controlled at a prespecified level a priori, given a dataset when designing or implementing a disclosure limitation procedure; rather, the risk is assessed in a post-hoc manner before and after a disclosure limitation procedure is applied. In summary, probabilistic disclosure risk depends on the prior knowledge of data intruders, the type and characteristics of the dataset from which information will be released, and the data releasing mechanism itself.

Differential Privacy and Disclosure Risk

Differential privacy (DP) is a broad framework for privacy protection, as formulated in 2006 (Dwork et al., 2006, 2017). This topic has since been the subject of intensive research and has been used by both large technology companies (Abadi et al., 2016; Differential Privacy Team, 2017; Ding et al., 2017; Erlingsson, et al., 2014; Messing et al., 2020; Wilson et al., 2020; Yousefpour et al., 2021) and by the U.S. Census Bureau for the 2020 Decennial Census and other data releases (Abowd et al., 2022; Foote et al., 2019; Machanavajjhala et al., 2008). Within this framework, researchers can conceptualize a relative measure of disclosure risk, whereby instead of comparing an intruder's knowledge before and after a data release, they compare the intruder's knowledge after the data release to a counterfactual scenario in which the release is performed with any one individual's data removed. Intuitively, every individual's privacy is protected because what is released is indistinguishable from what would have been released if that individual had opted out from the survey or study. For a disclosure avoidance method or mechanism to satisfy differential privacy, this "indistinguishability property" must hold for every potential dataset and for every individual represented in the dataset. This requirement is achieved by carefully injecting small amounts of random noise into data and/or statistical computations to obscure the effect of each individual data point while still allowing a statistical signal about the population to come through. This leads to a privacy-utility trade-off (e.g., see Slavković & Seeman, 2023, for a general discussion), whereby more noise provides stronger privacy

Copyright National Academy of Sciences. All rights reserved.

protections but less accurate statistics, and less noise provides more accurate statistics but weaker privacy protections. The level of privacy protection is measured in differential privacy (and its variants) by one or more privacy-loss parameters, often denoted by the Greek letter epsilon (ϵ), which bounds the relative increase in risk compared to the counterfactual opt-out scenario.

A number of variants of differential privacy are commonly used, including in the 2020 Decennial Census (Abowd et al., 2022). One common variant, known as "bounded differential privacy," requires indistinguishability when any one individual's data are *changed,* as opposed to being *removed* in the original notion described above (which is sometimes called "unbounded differential privacy" for contrast). Bounded differential privacy is suitable when the number of observations is publicly known information, and it can allow for some modest improvements in utility under this assumption. Unbounded differential privacy has the benefit of not requiring this assumption, and in particular applies also when analyzing subsets of a dataset, whose size may be unknown even if the overall size is known. Gong and Meng (2020) show that using bounded differential privacy makes it easier to formally compare probabilistic disclosure risk measures with differential privacy. Another axis along which differential privacy variants vary is how indistinguishability or privacy loss is quantitatively measured. The basic version, where privacy loss is typically denoted by a single parameter ϵ, is called "pure" differential privacy, with some variants known as "approximate," "concentrated," "Renyi," and "Gaussian" differential privacy. Interpretations of these variants can be found in Kifer et al. (2022) and many other variants have been described in Desfontains and Pejo (2020).

One important feature that all the variants of differential privacy share is transparency. They do not allow the privacy protections to rely on the secrecy of the noise infusion mechanism used to protect the data. Thus, the exact mechanism, including the parameters of all noise distributions used, can be made available to the public, as was done for the 2020 Decennial Census.[28] (This transparency is in contrast to some traditional approaches to disclosure limitation, such as swapping, where the swap rate is often kept as a secret parameter. Such secrecy is an intentional part of protecting confidentiality, but it also leaves the data user with a limited ability to adjust for the impact on data analysis of changes made to the data.) Transparency allows for public input into the design of the algorithm, such as what form of

[28]Census Bureau (2021). DAS_2020_Redistricting_Production_Code, https://github.com/uscensusbureau/DAS_2020_Redistricting_Production_Code

Copyright National Academy of Sciences. All rights reserved.

statistical utility or types of analysis it is optimized for, and allows data users to estimate the impact of the noise infusion on the accuracy and uncertainty of their analysis.

Comparing the Measures of Disclosure Risk

One can design general randomized mechanisms that satisfy differential privacy guarantees, independent of local data or data intruders' prior knowledge or models, and then apply those mechanisms to release statistics calculated from datasets so that the privacy loss from them is automatically controlled at a prespecified level. By contrast, statistical disclosure limitation procedures can provide a certain level of protection, but it is only quantifiable after they are applied. Probabilistic disclosure risk can be assessed before and after a disclosure limitation technique is applied, including a differential privacy mechanism, to evaluate the effectiveness of the procedure in a post-hoc manner.

It is important to recognize that different disclosure avoidance mechanisms, including noise infusion to attain the differential privacy metric (see Chapter 4), may differ in their limitation of the alternative disclosure risks. For example, Abowd and Vilhuber (2008) examine the relation between differential privacy and probabilistic disclosure risk in the setting of releasing multinomial data. Gong and Meng (2020) use the ε-DP mechanism, which imposes the same direct bound on both the differential privacy disclosure risk measure and the relative probabilistic disclosure risk defined in McClure and Reiter (2012) but only bounds the absolute probabilistic disclosure risk measure in proportion to the disclosure risk in the absence of the released data, i.e., the prior risk. These differences are assessed by McClure and Reiter (2012) in the context of releasing fully synthetic microlevel data (see Chapter 4 for a definition of synthetic data) that satisfies ε-differential privacy at different levels of ε and under different assumptions about what information is available to potential intruders.

Based on their simulations, McClure and Reiter (2012) note several findings about the relationship between the use of differential privacy mechanisms to bound absolute disclosure risk and their use to measure relative probabilistic disclosure risk. First, these two measures of risk do not necessarily provide the same answer to the level of risk. Their simulations show that relative risk can be low while absolute risk remains high. The simulations also demonstrate that absolute risk may be quite low, in which case the data steward can allow a higher level of relative risk as long as absolute risk stays below a chosen level. Second, the authors argue that data stewards

Copyright National Academy of Sciences. All rights reserved.

may wish to evaluate disclosure risk over a wide range of possible levels of intruder information to determine the worst-case scenarios for absolute disclosure risk. Third, given their findings, they propose a strategy for using both measures of disclosure risk, combining the assumption of strong prior knowledge by intruders (which underlies the differential privacy approach) with the interpretability and data-specific relevance of a fuller range of statistical risk measures.

Finally, the relevance of different measures of disclosure risk noted above may differ depending on the objectives of data curators, such as the Census Bureau. This matter is discussed in Hotz et al. (2022). For example, Hotz et al. note that schemes like that proposed in Abowd and Schmutte (2019)—in which a data curator seeks to design and calibrate formal privacy disclosure limitation mechanisms that trade privacy protection for the benefits of alternative uses of the data—require assessing the absolute disclosure risks. By contrast, the differential privacy mechanisms that focus on a different privacy-utility trade-off focus on the differential privacy measure of disclosure risk.

It is challenging to develop satisfactory definitions and measures of disclosure risk that both achieve the desired privacy protection and also allow for useful statistical releases. Here we enumerate some of those challenges.

- *Distinguishing privacy from statistical utility.* It can be difficult to draw the line between the inferences we want to prevent data users from reaching (for the sake of privacy protection) and those we want to enable (so that the data release provides societal value). In particular, generalizable inferences about an entire population (which is often the goal of statistical releases) also allow potential adversaries to draw new inferences about individuals. For example, if a data release teaches us that there is a high correlation between smoking and lung cancer in a population, it also allows us to infer that the smokers we know are more likely to have lung cancer. Thus, it is difficult to limit absolute or probabilistic disclosure risk without also limiting generalizable inference. Differential privacy is expressly designed to allow generalizable inference (and any inferences about individuals that are derived from it); it protects against learning features that are specific to an individual, ones that cannot be inferred from knowledge about the population.
- *Modelling an attacker's auxiliary information and attack strategy.* It is difficult to anticipate what auxiliary information, such as external datasets and background

Copyright National Academy of Sciences. All rights reserved.

information about a specific individual, a future attacker may have and therefore what attack strategy that attacker may carry out. Traditional assessments of disclosure risk (such as those described below) are done by estimating, either experimentally or analytically, the effectiveness of specific attack strategies (e.g., reidentification via linkage) using specific background information (e.g., specific commercial datasets). Thus, even if such assessments do not find evidence of risk, they do not rule out the possibility that an intruder possessing different background knowledge, such as through a coworker or ex-spouse of one of the data subjects, or by carrying out an unanticipated attack strategy (e.g., a sophisticated reconstruction attack, as discussed below) can violate the privacy of the data subjects. On the other hand, the protections of differential privacy hold regardless of the attacker's auxiliary information and computational strategy.

- *Composition.* A major challenge in measuring disclosure risk is how it accumulates when many statistical releases are performed. If we know that each release has a small risk, can we conclude that over many releases about the same or overlapping sets of individuals, the risk remains small? This is particularly relevant when releasing longitudinal datasets. Unlike most other measures of disclosure risk, differential privacy satisfies a variety of *composition theorems,* which quantify how much privacy loss can accumulate over multiple releases.

Assessing Disclosure Risk Through Analytic or Empirical Approaches

Disclosure risk can be either computed analytically or estimated empirically. Analytically, measuring disclosure risk means using a statistical analysis or proof to bound the disclosure risk measures described above. For absolute and relative disclosure risk, this requires carefully modelling the data distribution, the intruder's prior knowledge, and the inferences one seeks to prevent. Assessing differential privacy requires much less modelling, as it is primarily a matter of determining the unit of privacy protection, such as individuals, households, or firms, and then determining how those units are represented in the data; the privacy loss can then generally be bounded using a mathematical proof.

Empirically measuring disclosure risk means carrying out experiments using existing attack methods to see how vulnerable a given dataset and disclosure avoidance mechanism is to the attacks. The results obtained, while informative, are thus lower bounds on risk. That is, if the

Copyright National Academy of Sciences. All rights reserved.

experimental attacks are successful, then we know that the disclosure avoidance mechanism is insufficient. But if they are unsuccessful, it still remains possible that an attacker using different auxiliary data sources or attack methodology will be able to violate confidentiality. For this reason, it is important that the experiments be as comprehensive as possible, drawing on a wide range of auxiliary data sources (e.g., federal administrative data, state databases, and commercial databases) and using the most sophisticated attack methods known.

When measuring disclosure risk empirically or using statistical modelling, it is important to explicitly consider different intruder models. The auxiliary data sources, attack methodology, and level of tolerable disclosure risk may differ widely depending, for example, on whether the attacker is a foreign intelligence agency trying to manipulate the U.S. electorate, a commercial data aggregator trying to acquire data to monetize, a domestic abuser trying to learn about their victim's behaviors, or a researcher evaluating the effectiveness of a government program under the supervision of an Institutional Review Board (IRB). For each mode of access an agency provides, the kinds of adversaries who could gain access may differ. Assumptions about the intruder are key considerations in every measurement of disclosure risk.

Quantifying Risks of Re-identification

This section provides an overview of attack scenarios and disclosure risk measures for measuring the risk of re-identification. Disclosure risk assessment methodologies have been developed for using external data, for example through record linkage, and also through the use of probabilistic modelling without the availability of external data.

Quantifying Re-Identification Risk Through Matching to External Data Sources

A matching exercise that uses common variables in an attempt to link the records in one file with the same cases in an external file can be used to quantify re-identification risk. For example, such an exercise might try to link the planned public-use file, along with sufficient information to determine if a match was correct, with public administrative data or commercially available databases. If exact matching does not work well, probabilistic record linkage (Fellegi & Sunter, 1969; Winkler, 2006) can replace exact matching to quantify the likelihood of successfully linking a case. Exact matching might not work well if common variables are defined slightly differently, if data points are collected at different points in time in the two files, or due

Copyright National Academy of Sciences. All rights reserved.

to measurement or sampling error, among other things. The records and variables with high re-identification risk can be targeted for perturbation in the risk mitigation process. The percentage of correct matches or the percentage of records whose best match is the true record can be reported as the risk measure.

As an example of this setting, early work by Yancey et al. (2002) and Domingo-Ferrer et al. (2001) used probabilistic and distance-based record linkage to scope out the risk of re-identification in public-use files. The linkages were carried out between the public-use file and the original confidential microdata, and a quantitative measure, such as the distance between the public-use file record and the correct record, were used to quantify disclosure risk. This approach is generally an overestimation of disclosure risk since it assumes that intruders have access to the original confidential data. In addition, this approach does not take into account the protection afforded by the sampling, which is a benefit if the survey inclusion status is unknown. There are limitations: the approach is dependent on the choice of quasi-identifiers, and it only assesses the specific strategies of probabilistic and distance-based record linkage. However, it is a good way to measure risk in the case of a 'nosy household member' scenario, that is, a situation where someone may know the true and reported values of indirect identifiers in the sample.

More recent work on measuring the risk of re-identification has been based on intruder penetration testing and simulating attack scenarios based on freely available information open to the public, such as Facebook, Google maps, and consumer/lifestyle databases that can be purchased. This approach requires a deep understanding of the data environment and what information is available in the public domain that can lead to an attack in the public-use file through linkages. See Tudor et al. (2014) for an example of an application.

A Heuristic Method for Quantifying Re-Identification Risk

Another approach for measuring the risk of re-identification is based on the concept of the "special unique" developed in Elliot et al. (2002); see the software code titled SUDA (Special Uniques Detection Algorithm). A special unique is based on the assumption that a record in the released confidential microdata, which is a sample unique on coarser, less detailed information, is riskier than one which is unique on a finer, more detailed set of information. Extensive empirical work has shown that special uniques were more likely to be population uniques and were correlated with the reciprocal of the population equivalence classes on a set of key

Copyright National Academy of Sciences. All rights reserved.

variables *K*. The SUDA software classifies special uniques according to the size and number of subsets to obtain the minimal sample uniques. Examining all possible subsets of key variables *K* for special uniques requires high computational time. In addition, this approach also does not take into account the protection afforded by the sampling, but it would be relevant to determining the risk in the case of a 'nosy household member' scenario as mentioned above.

Quantifying Re-Identification Risk Using Probabilistic Models

Another quantitative disclosure risk measure for the risk of re-identification, which was proposed in Skinner and Elliot (2002), is called the Data Intrusion Simulation measure. In contrast to the typical scenario of matching a record from the confidential data to the population, this approach uses a 'fishing' scenario, that is, one where the intruder has information about a data subject in the population, then explores whether the data subject is in the confidential data, and if so, provides the probability that there is a correct identification given that it is a unique on a set of variables that are quasi-identifiers (QIDs). Suppose the quasi-identifiers are used to create a cross-classified table, and the sample uniques in the cross-classified tables are identified. Suppose that population sizes in the corresponding sample unique cells are also known or estimated. The Data Intrusion Simulation is defined as the ratio of the number of sample unique cells to the total population of the corresponding cells. For more technical details, see Appendix A.

The 'fishing' scenario is generally not sufficiently relevant for statistical agencies that need to assess the consequences of releasing a whole data public-use file (Cohen, 2022). Therefore, much literature has since been devoted to estimating more relevant disclosure risk measures based on the attack scenario of an intruder having access to the confidential microdata and wanting to make an identification based on an external data source containing information about the population. Two main disclosure risk measures[29] are

(1) Expected number of sample uniques that are population uniques; and

(2) Expected number of correct matches for sample uniques assuming a random choice of potential matches within the cell. For example, if a sample unique matches to three individuals in the population, the match probability for that sample

[29] These are defined mathematically in Appendix A.

Copyright National Academy of Sciences. All rights reserved.

unique would be 1/3. Aggregating all match probabilities over the sample uniques leads us to a measure of disclosure (Shlomo & Skinner, 2022).

We note that the probabilistic modelling approach accounts for the protection afforded by the sampling. Limitations of this approach are that the choice of quasi-identifiers cannot be too large and often a number of iterations are carried out based on different sets of key variables to assess the risk of re-identification. For data released in the form of households and individuals, two different models are applied, one at the household level and one at the individual level, and the results of the risk assessment are merged.

Alternative Focus on Disclosure Limitation: Prevent Reconstruction

In the discussion so far, the focus of disclosure risk has been on estimating the re-identification risk from the released data. Somewhat surprisingly, even releases of what appear to be simply aggregate statistics (e.g., contingency tables) are also subject to attack. Experiments with reconstruction attacks motivated the U.S. Census Bureau to adopt differential privacy for the 2020 Decennial Census (Abowd, 2021), and reconstruction attacks have also been demonstrated to be effective against commercial privacy software (Cohen & Nissim, 2020).

As attack scenarios have become more advanced, there is value in more perturbative methods being applied to the outputs. In recent years, more focus has been directed at the risk of reconstruction rather than the risk of re-identification following the extensive work carried out at the U.S. Census Bureau. Even releasing data in aggregate form has an associated disclosure risk of reconstruction. For example, tabular data are susceptible to attacks by "table differencing" (i.e., when one takes the differences between two similar tables to generate estimates for a small subpopulation) and "table linking" (i.e., when one strings together information from several tables in order to form a microdata record on a singleton).

The database reconstruction theorem (Dinur & Nissim, 2003) has made data producers cautious about the number of data estimates they publish per number of data values in the confidential microdata. For example, if a very large number of tables is produced from a small dataset, then it is likely that most, if not all, of the dataset can be reconstructed by the published tables. Such reconstruction attacks are a great risk for outputs based on tabular data. A reconstruction attack is one that emulates an intruder using the full set of tabular products to reconstruct the underlying confidential microdata. The threat comes from an understanding of

Copyright National Academy of Sciences. All rights reserved.

the constraints within a series or set of aggregate releases generated from confidential microdata. Using non-perturbative disclosure limitation methods, such as cell suppression, reconstruction algorithms can easily be built as a system of equations to be solved. Garfinkel (2019) provides details of a reconstruction attack as it relates to census products. Even when one limits the releases of aggregate data, for example by publishing partial tables or rounded statistics such as underlying rates in the tabular data, others can easily reconstruct the relevant tables by relying on the same ideas, using a system of equations to be solved (for a few simple examples, see Slavković & Fienberg [2004] and Fienberg & Slavković [2005]). As mentioned, using perturbative methods and the preset disclosure limitation rules in a remote table builder can mitigate the risk of reconstruction (see Chapter 7 for additional discussion of table builders).

Furthermore, attack scenarios have become very sophisticated, to the point that disclosure risk checks are important even for perturbed output data or synthetic data. For example, membership inference attacks can successfully detect the data used to create the synthetic or perturbed output by observing the output, even without the synthetic data model's parameters. Another attack scenario, called the memory-based attack, focuses the attack on getting information in or data out of the system's memory, which is less detectable than direct attacks on file directories (Daily, 2022; Francis, 2022; Jarmin, 2021; McKenna, 2019).

EXAMINING DISCLOSURE RISKS FACED BY SIPP

As described in Chapter 2, SIPP has characteristics that both help and hinder the task of disclosure avoidance. The use of statistical sampling is largely helpful by making it more difficult to identify whether individuals are unique within the population, though including data defining the statistical strata acts to group similar respondents in ways that present the potential for risks. SIPP's complex sample design, including oversampling of primary sampling units with higher concentrations of low-income households, would lessen the protection provided by sampling to some degree. The use of data imputation for item nonresponse, sometimes for a large number of cases (again, see Chapter 2), in effect introduces synthetic data even into the public-use file. Several factors complicate the task of disclosure avoidance, including the large number of variables containing potentially identifying information, the high granularity of the data, the presence of data not simply on individuals but on households, and the capacity to use longitudinal data to measure change over time.

Copyright National Academy of Sciences. All rights reserved.

While the disclosure risks associated with some of these features have been assessed in other data releases, both by the Census Bureau and by other data providers, the existing literature does not provide much guidance on the disclosure risks associated with the *combination* of these features, such as disclosure risks of longitudinal data on households, nor on the combination of the above features of SIPP. There have been two previous incidents in which disclosure avoidance became an important issue for SIPP: one occurred in 1996, when data on adolescents in a special self-administered section could not be publicly released because they could easily be linked to the household data, allowing parents to identify their children's responses. In this case the lack of a public release disappointed many who planned to use the data. The other incident involved a husband hearing his wife's previous responses as part of what had been standard procedures to shorten the length of the interviews, and because of this newly learning that his wife had been previously married. As a result of these two incidents, Census modified its confidentiality protections by explaining that confidentiality between members of the same household cannot be protected (McKenna, 2019).

The Level of Detail in SIPP Presents a Challenge

The ability to identify unique respondents for the purpose of matching depends on the number of variables available, the granularity or coarseness of the data, and the degree to which particular values are relatively rare (e.g., identifying someone as Native American would do much more to identify someone uniquely than identifying someone as White). All data, when used in combination, are potentially disclosive, but data that are highly granular simplify the task of intruders seeking to identify respondents.

The SIPP data are highly granular, as shown below in the statistics from the SIPP 2020 file among the primary respondents in month 12.[30] The dollar amounts are rounded to varying amounts. For example, the variable TJSCHKVAL, or total share of a joint shared checking account, only appears to be rounded for values of $1,000 or more, then rounded to the nearest $10, but is rounded to the nearest $100 starting with $10,000, and has a top-coded value of $85,300; with 136 observations resulting in 473 unique values, 253 of those values are associated with only a single person. The occupation data take 522 values, and the state identifiers take 52

[30]This section contains custom tabulations provided by the Census Bureau to support the panel's analysis, all based on the public-use file.

Copyright National Academy of Sciences. All rights reserved.

values. Three standard variables alone (occupation by race by sex) are sufficient to produce 613 singletons in the 2020 panel, and adding state of residence as a fourth variable is sufficient to produce 5,627 singletons.

As a sample survey, however, the weights for each observation in SIPP range from 387 people to 39,951. Thus, it is not clear how many of the singletons within SIPP represent actual singletons in the population. Note that one could potentially use a different source with more observations and thus generally smaller weights, such as the American Community Survey, to estimate the number in a subgroup more precisely, but these still would be weighted statistics. The true test of uniqueness is through using the most comprehensive databases that might be available to data intruders.

The granularity of the data is increased when one considers the multidimensional nature of the data. The above frequencies are based on seeking to find a match for a single individual. However, the SIPP data also include information about other people in that person's household, as well as changes over time. If one has data about these other features, then even a relatively small number of variables applied over a full household can often identify a household uniquely. Table 3-1 illustrates how the ability to identify a respondent increases as these extra aspects of data are considered. For the 2020 SIPP file, looking for simplicity's sake only at Wave 1 households, there were a total of 8,004 households. When looking at the primary respondent only, there were 440 different combinations of sex/race/ethnicity/state, and 99 of those combinations described only a single household in the file. Considering that SIPP is a sample-based survey, these combinations of households are unlikely to represent unique households from a truly national perspective. (As noted above, the smallest weight is 387.) Still, as the number of household members being matched is increased, the number and percentage that are unique households increases greatly. When the characteristics of the first five adults in the household are considered, then 15.8 percent of households have unique characteristics, based only on sex, race, ethnicity, and state. If two more variables are added (occupation and age), again allowing for up to five adults, then 93.9 percent of the households (or 7,513 of 8,004 households) have unique combinations of characteristics. In fact, almost the same percentage of households (93.6 percent) can be uniquely identified based on only two adults and the same characteristics. Again, without matching these characteristics against a national database (such as the decennial census or some commercial databases), one cannot precisely calculate the

Copyright National Academy of Sciences. All rights reserved.

disclosure risk to any household. However, with 7,513 households uniquely identified, it seems likely that some households can be uniquely identified in a national database, using only a small number of all of the variables available in SIPP for matching. Adding in measures of change over time within the year 2020 increases the disclosure risk even more. This analysis demonstrates how the dimensionality of the data contained in SIPP is a key consideration when evaluating disclosure risk.

Table 3-1 Percentage of SIPP Households That Are Unique in Wave 1, Based on the Number of Types of Information Included

Number of characteristics	Primary member only	Two adults	Three adults	Four adults	Five adults
Sex/race/ethnicity/state	1.2	7.4	11.6	14.8	15.8
Add occupation	40.0	63.6	66.7	67.3	67.4
Add age	77.8	93.6	93.8	93.8	93.9
Add number of children	86.5	93.9	94.0	94.0	94.0
Add change over time in number of children	94.0	94.0	94.0	94.0	94.0

NOTE: Data are for Wave 1 of SIPP 2020. Change over time is measured within 2020.
SOURCE: U.S. Census Bureau, custom tabulations.

One way to further examine the likely rarity of the above unique combinations of characteristics is to examine their frequency across both waves included in the 2020 SIPP file.[31] If a set of characteristics appears only once across multiple waves, then its frequency is likely to

[31]A SIPP data collection year may include up to four panels, each with a different wave. However, there was no 2017 SIPP, and the 2019 SIPP was limited to a single wave. Thus, the 2020 SIPP had wave 1 from the 2020 panel and wave 3 from the 2018 panel.

Copyright National Academy of Sciences. All rights reserved.

be less than the minimum weight of 387. Frequencies across all waves are shown in Table 3-2.[32] This results in dropping some sets of combinations that are no longer unique across all waves, while adding some that appeared only in a different wave. The percentage of cells with unique combinations of characteristics is reduced relative to Table 3-1, but the same basic pattern emerged: the likelihood of unique combinations increases as the number of characteristics increases and as the number of adults is increased. Once one has added age and includes the two primary adults, then about 87.5 percent of the cells are unique, with little gained by adding more characteristics or adults in the same household.

Table 3-2 Percentage of SIPP Households That Are Unique Across Waves, Based on the Number of Types of Information Included

Number of characteristics	Primary member only	Two adults	Three adults	Four adults	Five adults
Sex/race/ethnicity/state	0.4	4.5	8.5	11.0	12.0
Add occupation	28.8	56.9	60.9	61.5	61.7
Add age	64.3	87.5	88.0	88.1	88.1
Add number of children	74.6	88.7	89.1	89.1	89.1
Add change over time in number of children	89.2	89.2	89.2	89.2	89.2

NOTE: Data are for Waves 1 and 3 (combined) of SIPP 2020. Change over time is measured within 2020.

SOURCE: U.S. Census Bureau, custom tabulations

Given that many databases may not contain occupation codes, particularly at the level collected in SIPP, Table 3-3 provides a similar breakdown by omitting occupation and instead

[32]SIPP 2020 contained data only for waves 1 and 3, reflecting the 2018 and 2020 panels. SIPP was not conducted in 2017, so there is no wave 4, and the 2019 SIPP panel was discontinued after a single year.

Copyright National Academy of Sciences. All rights reserved.

including the highest level of education attained.[33] Adding the level of education results in making 59.2 percent of the households unique, even when considering only the primary respondent, and over 90 percent of the households are unique if two or more adults are considered. Again, given that the smallest weight was 387, unique combinations of characteristics within SIPP may not be unique within the population. If someone knew that a particular person had participated in SIPP, then it is highly likely that person could be identified based on only a small number of variables. To determine which of these combinations were unique in the entire population, one would need to examine an alternative database.[34]

Table 3-3 Percentage of SIPP Households That Are Unique, Based on the Number of Types of Information Included, and Replacing Occupation with Highest Level Of Education

Number of characteristics	Primary member only	Two adults	Three adults	Four adults	Five adults
Sex/race/ethnicity/state	0.4	4.5	8.5	11.0	12.0
Add birth year	21.4	72.3	74.9	75.2	75.3
Add highest level of education	59.2	91.4	92.0	92.0	92.1
Add number of children	77.8	92.6	93.0	93.1	93.1
Add change over time in number of children	93.1	93.1	93.1	93.1	93.1

[33]SIPP collects considerably more detail on the education level than would appear in most databases (e.g., with categories such as "less than 1st grade"). To better simulate what might be available on external databases for matching purposes, the categories were collapsed to: less than high school diploma, high school graduate, some college, associate's degree, bachelor's degree, master's degree, and professional or doctorate degree.

[34]An alternative approach, which may not be feasible, is to computationally consider the space of all possible tables that satisfy some of the percentages and other statistics that are reported, akin to linear or integer programming.

Copyright National Academy of Sciences. All rights reserved.

NOTE: Data are for Waves 1 and 3 (combined) of SIPP 2020. Change over time is measured within 2020.
SOURCE: U.S. Census Bureau, custom tabulations.

Choosing Which External Databases to Compare with SIPP

One factor not discussed in the overview of SIPP in Chapter 2 is the degree to which variables provided in SIPP are also available in other databases for potential matching. Data from displays some of the relevant variables available on three major commercial databases, though this list is incomplete (e.g., Acxiom says that it measures 10,000 attributes). Other potential databases that might be used for matching include federal and state databases. Collectively, these databases present considerable data available for a data intruder to use for matching and identifying SIPP respondents, as well as to identify the likely uniqueness of various combinations of attributes.

Table 3-4 Key Areas in Which Three Commercial Databases Have Data That Correspond to SIPP Data

Broad category	Data Axle	Equifax	Acxiom
Demographics			
Basic			
G1 *Date of birth*	X		X
Sex	X		X
Ethnicity	X		X
Race			X
Geographic data		X	X
Educational attainment	X		X
Family and household relationships			
Number of adults in household	X		
Number of children in household	X		
Age range	X		
Gender	X		
Marital status	X		
Households likely to buy items for children		X	
Family wireless plan subscribers		X	
Marriage/divorce			X
Birth of children			X

Copyright National Academy of Sciences. All rights reserved.

Broad category	Data Axle	Equifax	Acxiom
Demographics			
Basic			
G1 *Date of birth*	X		X
Sex	X		X
Ethnicity	X		X
Race			X
Geographic data		X	X
Educational attainment	X		X
Family and household relationships			
Number of adults in household	X		
Number of children in household	X		
Age range	X		
Gender	X		
Marital status	X		
Households likely to buy items for children		X	
Household size			X
Number & ages of children			X
Adult living relatives			X
Nativity, citizenship, and parent nativity			
Country of origin	X		
Residence			
New homeowners			X
New Movers			X
Moves			X
Assets, employment and earnings			
Assets			
Home value	X		
Home sales price	X		
Years at residence	X		
Home age	X		
Mortgage date	X		
Single family home	X		
Multi-family home	X		
Boat owner	X		
Aircraft owner	X		
Mortgage amount	X		
Loan-to-value percentage	X		
Invested assets		X	
Estimated household assets		X	
Credit behaviors and outstanding balances		X	
Student borrowers		X	
Net worth		X	

Copyright National Academy of Sciences. All rights reserved.

Broad category	Data Axle	Equifax	Acxiom
Demographics			
Basic			
G1 *Date of birth*	X		X
Sex	X		X
Ethnicity	X		X
Race			X
Geographic data		X	X
Educational attainment	X		X
Family and household relationships			
Number of adults in household	X		
Number of children in household	X		
Age range	X		
Gender	X		
Marital status	X		
Households likely to buy items for children		X	
Homeowner status			X
Home purchase			X
Auto purchase			X
Employment and Earnings			
Household income	X	X	
Expendable income	X		
Occupation	X		X
Health care utilization and medical expenditures			
Health-related measures			X
Insurance behavior			X

NOTE: These firms do not publish comprehensive lists of their variables, and they may have data comparable to SIPP that are not listed here. Thus, this table may understate the degree to which the databases have information that might be matched against SIPP. The firms also have a substantial amount of data that do not directly correspond to SIPP variables and are not shown here.
SOURCE: Data from www.dataaxle.com; www.equifax.com; www.acxiom.com; and marketing materials provided by Heather Conklin at Data Axle.

The Census Bureau's Measurement of Risk Through a Re-identification Study

To help gauge the level and nature of disclosure risks associated with the SIPP Public Use Microdata Files, the Census Bureau completed an initial re-identification risk study for the 2014 SIPP panel, referred to as the "Re-id Study," that involved a statistical matching process on

Copyright National Academy of Sciences. All rights reserved.

tax filers from SIPP panel data with four years of IRS data.[35] In doing so, high-risk areas in the data, in terms of records or groups of records in conjunction with variables or combinations of variables, can potentially be identified and explored further for a more concentrated risk assessment, leading to consideration for data treatment solutions. In this manner, re-identification studies are an iterative process.

The Census Bureau's Re-id Study is an example of an intruder attack at one point in time with one realization of the sample/population pair of files. It assumes that the intruder does not know the sample inclusion status of the population, but knows information relating to a certain set of indirect identifiers. One external "population" file is used, the linked SSA/IRS file, which is not publicly available and therefore is considered a conservative choice. Prior to conducting the study, the data used in the statistical matching were harmonized to have consistent definitions and categories across databases. The extent of the recoding for categorical variables, and the extent of the skewness in continuous variables, have a bearing on the risk results.

The Re-id Study assessed the disclosure risk from combinations of 36 indirect identifying variables that were used together in a record linkage process. The study also took into account the longitudinal data structure by including four years of data for the same household records. Movements of households were addressed, to a limited extent, over the length of the panel. In addition, the study took into account that the SIPP data are from a sample, as well as oversampling or differential sampling rates, by linking about 28,000 records from the SIPP 2014 panel to over 90 million records from a linked SSA/IRS data file that contained IRS data from 2013 to 2016 and SSA Numident for 2013. Linking in this way made possible many potential correct matches (or ties) on the 36 variables, where the true match may or may not be part of that group of potential correct matches.

The statistical matching in the Re-id Study used two distance measures, called the Taxicab and the Euclidean metric, that were computed for each pair of observations across the SIPP panel data and the SSA/IRS file, within strata based on state by metro/non-metro status. For each of the four waves, the variables used were mover status, age, race, sex, presence of

[35]Aref Dajani, Steve Clark and Phyllis Singer, "SIPP 2014 Panel Re-Identification (Re-id) Study Findings and Recommendations," Presentation to CNSTAT Panel on A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (SIPP), March 21, 2023. This briefing provided a summary of the reidentification study; the full report had not yet been released at the time of this writing. The write-up in this section is based on that briefing.

Copyright National Academy of Sciences. All rights reserved.

spouse in the household, number of children, and six sources of income. Note that only age is not susceptible to changes over time, and indeed it is identifiable. The results from the Re-id Study were classified into suspected matches using undisclosed thresholds. Re-identification rates using ratios of suspected to total count were computed. The "true" matches were identified using a third internal file (referred to as the Internal Use File). Knowing the true matches allowed two more rates to be computed: (1) the confirmed rate, which is the number confirmed over the total record count in the stratum; and (2) the conditional rate, which is the number confirmed over the number suspected. The Census Bureau uses the latter as a measure of precision.

Overall, the findings of the Re-id Study were that the risks of re-identification in the 2014 panel of the SIPP Public Use Microdata Files were "better than expected." For state by metro/non-metro area strata, the Study found that 72.2% of strata had zero confirmed re-identifications using the Taxicab metric and 13.9% had zero confirmed risk using the Euclidean metric. Mover households in the SIPP were re-identified at similar rates to those households that did not move over the course of the panel, although more states had zero confirmed re-identifications for nonmover households than for those that moved. Housing units in metro areas were re-identified at lower rates than nonmetro units. The study also found that no clear patterns emerged in rates of confirmed re-identifications across the various "linking variables" used to link with the external databases such as age, race, sex, presence of spouse, number of children and categorical versions of alternative sources of income, across the panel's four waves. Finally, the Re-id Study found that age was the most disclosive variable among those used, both individually and with other linking variables. Based on this preliminary study, the Census Bureau concluded that no changes were needed to the public-use data file. (However, some limitations of this preliminary study are discussed in the following section.)

RECOMMENDATIONS TO CENSUS FOR ON-GOING DISCLOSURE RISK ASSESSMENT STRATEGY FOR THE SIPP

In this section, we offer a set of recommendations to the Census Bureau suggesting what elements should be included in an ongoing program for assessing the disclosure risks associated with both current and possible ways of releasing SIPP data, including public release files (public-use files) of microdata. We begin with an assessment of the Census Bureau's recent re-identification study of the 2014 SIPP panel, a summary of which was presented to our panel for

Copyright National Academy of Sciences. All rights reserved.

this report. We then offer our recommendations for the future program the Census Bureau should establish to assess the disclosure risk of SIPP data products.

Limitations of the Census Bureau's Re-Identification Study

Our panel was impressed with the work that went into the Re-id Study. Furthermore, the authors of the Re-id Study were careful to note many caveats, limitations, and assumptions, some of which were noted above. Nonetheless, our panel had several reactions to the study that bear on the future efforts of the Census Bureau to assess the disclosure risks of SIPP data products, especially its public-use files.

First, we note the lack of an acknowledgement in the study that many variables in the SIPP are imputed. In particular, a large portion of the income data are imputed. One might hypothesize that data that are imputed have less predictive power, and thus may produce more incorrect matches than would an analysis based only on actual survey responses. Still, if the imputed value comes reasonably close, it may support finding a match (certainly much more so than if one were working with a missing value). A data intruder might seek to focus only on non-imputed data to get better matches. Whether that strategy would work is not clear: a study by Krenzke et al. (2014) showed that the use of non-imputed data can lead to overestimation of disclosure risk, while including imputed data corrects for that overestimation. Typically, for measures such as income one does not look for an exact match but rather at the closeness of the values in the examined databases, and an imputed value might still come sufficiently close (plus, respondents' self-reports are also subject to error). The use of imputation also provides deniability to the respondent (e.g., "I didn't say that; that is an imputed value"). The public release of imputation flags has multiple implications: (1) it may affect whether a respondent has deniability, and (2) it changes what strategies are available to a data intruder when seeking to identify a respondent. Ultimately, whether or not imputation has been performed, a re-identification study provides a test of the extent of disclosure risks, but it could also be used to explore whether a data intruder might benefit by focusing on actual reported values. As a side note, synthetic data are a type of imputed data and can contain disclosure risks just as imputed data can (Hotz et al., 2022). Again, one depends on tools such as re-identification studies to confirm the safety of the data.

Copyright National Academy of Sciences. All rights reserved.

Second, data intruders may show tremendous creativity and persistence when seeking to identify respondents, while the Census Bureau's analysis did not explore the full range of options that are available to data intruders. The panel received only a summary of the Census Bureau's findings, and only data that had been approved by the Census Bureau's Disclosure Review Board could be released, but following are some additional types of analysis that might have been performed.

- *Alternative thresholds*. It would be beneficial to examine the use of alternative thresholds for identifying putative links, since data intruders are not limited to the thresholds that the Census Bureau used. One might anticipate that the closest observations will be the ones most likely to be correct matches, so that lowering the threshold might reduce the number of false positives and increase the ratio of correct matches to false positives. In particular, the Census Bureau's use of the five closest observations will tend to produce a false positive rate of 80 percent or higher.[36] Indeed, the best practice for privacy attacks accepted in the research literature is to set thresholds to reach a very small false positive rate (Carlini et al., 2022; Dwork, et al., 2017; Narayanan & Shmatikov, 2008); in such a regime, any re-identifications (even if only a few) are highly confident and thus more worrisome. For example, Narayanan and Shmatikov (2008) only declare a putative re-identification if the second-closest match is at least 1.5 standard deviations further than the closest match. With regard to publishing tables, a common rule is that at least three observations must appear in a table cell in order for that cell's data to be published,[37] again suggesting that a threshold of five is higher than in standard practice.
- *Change over time*. The analysis of change over time was limited to whether respondents had moved, and that measure was a summary measure across four years

[36]When there are exactly five observations deemed as being the closest, then the false positive rate is either 80 percent or 100 percent, depending on whether one of the five is a correct match. Lower false positive rates can appear when there are ties. For example, suppose that three observations are tied as being the fourth, fifth, and sixth closest. The Census Bureau's procedure would be to exclude those three, leaving only those that are first, second, or third, resulting in a false positive rate of either 67 percent or 100 percent. However, in such cases, a lower threshold would also succeed in including a correct match. Thus, in general the larger the threshold is, the higher the false positive rate will be, even though increasing the threshold will also increase the chances of including the correct match.

[37]Federal Committee on Statistical Methodology. https://nces.ed.gov/fcsm/dpt/content/3-2-2-1

Copyright National Academy of Sciences. All rights reserved.

rather than a year-to-year measure. More precise measures might have led to more precise links. The Census Bureau did include measures that changed over time (such as the number of children and income levels), but did not look at changes in those measures over time.

- *Multiple members of households*. The analysis did not look at how combining information across multiple members of a household might be disclosive. For example, as discussed in the section on data granularity above, having data on the ages of two adult members in a household is much more disclosive than having data on only a single member. For that matter, even the number of adults living in the household is potentially a disclosive variable, without considering the characteristics of each adult.[38]
- *Comparisons apart from IRS data*. The analysis was limited to comparisons with IRS 1040 data as merged with SSA NUMIDENT data, based on the logic that this is an extremely tough test using highly confidential data that had previously shown disclosure issues on the American Community Survey. The panel agrees this provides a tough test, but (1) it is not a pathway that a data intruder is likely to take since a data intruder normally will not have access to such data, and (2) the ability to create correct linkages depends largely on the number and types of variables used to create links, and there are a number of potentially disclosive variables available in external data that might be used for matching but are not contained in the IRS and NUMIDENT data (these include home values, education levels, assets and liabilities, and state-provided income support). Also, the Census Bureau chose to focus on nationally available data, but there may be much greater risks for respondents living in states that release identifying information on program participation.
- *Limitations of IRS databases*. The ability to find correct matches also depends on how each database is structured, and how complete each database is. In this respect the IRS data have important limitations.

[38]The degree of risk or ability to match a person will depend on the database being used. A person who does not pay taxes will not be on the IRS database but may be in other databases (including of those receiving Social Security benefits and in databases maintained by credit agencies). IRS data are not designed to consider households with multiple filers, though perhaps they could be aggregated by address to create more comprehensive measures of households.

Copyright National Academy of Sciences. All rights reserved.

- The IRS data are incomplete in the sense that the re-identification study looked at tax filers only,[39] while roughly 40 percent of tax units do not file tax returns, and are thus omitted from the database.[40] One might argue that such households are not at risk of disclosure, which would be true with respect to comparisons with IRS tax data, but would not necessarily be true with respect to comparisons with other databases, even regarding data on incomes. The Census Bureau could potentially augment its intruder file by including all people with Social Security numbers, by adding those who pay payroll taxes, and/or adding those who receive Social Security benefits (or other benefits).
- The IRS data and SIPP data define households differently. A SIPP household might include adult children living with the parents, other relatives such as parents living with the primary respondents, or even nonrelatives. Thus, a SIPP household might be represented in the IRS data through multiple tax returns, and household-level data may not be comparable unless the IRS data are somehow aggregated. On the other hand, if a SIPP household is broken down into individuals, then the IRS data for married couples filing jointly will not be comparable.
- In the re-identification study, the Census Bureau chose to simplify the analysis by picking a single member of each household to represent the household, that is, the household member earning the highest total monthly income within a household in the first year of the panel. This approach excludes some people from the confidentiality analysis. For example, roughly half of adults ages 18-24 lived with their parents in 2019 (Fry et al., 2020) and 18 percent of young adults ages 25-34 did the same (Creamer et al., 2020);[41] if those young adults are employed and

[39]Email from Holly Fee, Census Bureau, May 16, 2023.

[40]https://www.taxpolicycenter.org/model-estimates/tax-units-with-zero-or-negative-federal-individual-income-tax-oct-2022/t22-0131 A tax unit is defined as "an individual, or a married couple, that files a tax return or would file a tax return if their income were high enough, along with all dependents of that individual or married couple." https://www.taxpolicycenter.org/resources/brief-description-tax-model This is different from a SIPP household, which may consist of multiple tax units.

[41]https://news.prudential.com/58-young-adults-are-still-living-at-home-impacting-their-parents-path-to-retirement.htm#:~:text=More%20than%20half%20(58%25),are%20living%20with%20their%20parents

Copyright National Academy of Sciences. All rights reserved.

filing tax returns, then they potentially could be matched against the IRS data (or other sources), but their risks of disclosure are not being examined. Looking at the SIPP 2020 data (waves 1 and 3), 3,334 out of 21,504 households (16%) would have been excluded because no one in the household filed federal income taxes, and approximately 10,000 of 24,000 tax filers (42%) would have been excluded through the choice to focus on the household member earning the highest income.[42] The disclosure risk for such respondents thus has not been measured; one might speculate that because they are a smaller subgroup, they are therefore more likely to be unique and face higher disclosure risks. A data intruder might make different assumptions or work with a different database that is better structured for working with the SIPP data.

We recognize that the Census Bureau faced multiple constraints, including what datasets were readily available, the time and resources available for the analysis, and the limitations placed by the Bureau's Disclosure Review Board on what could be released to our panel or to the public. Most of the reporting focused on the precision results, which are not very helpful in identifying those who are at risk of disclosure. Rather, one wishes to know which subgroups faced disclosure risks so that they can be addressed. We suggest that using classification trees in lieu of logistic regression will allow a deeper dive into the results by determining "pockets" of high-risk areas. In doing so, risks among combinations of variables will be more apparent than risks are when looking at results one variable at a time.

More generally, the Panel concluded that more can be learned from the Re-id Study than was contained in the presentation provided. Such information and detail would be needed for the Panel and others to draw strong conclusions about the findings of this study and its adequacy. Furthermore, as discussed below, future studies of the re-identification risks of SIPP public-use files would benefit from additional analyses that were not contained in the Re-id Study. For

[42]This latter set of statistics is an approximation based on the following: The SIPP data show 32,647 who filed federal income taxes. Of those, 9,855 filed a joint return, and of those 8,555 were in a household that filed two or more returns. To avoid double-counting, those with two or more returns were halved, leaving approximately 32,637 – 4,377 = 28,370 unique tax returns from 18,170 SIPP households. Thus, about 10,000 filers would have been excluded as belonging to households in which someone else was being counted.

Copyright National Academy of Sciences. All rights reserved.

example, using other external files and other key identifying variables, the statistical models mentioned in this chapter can be used to gauge the risk and may identify what subgroups (e.g., SNAP participants) pose the highest risk in more time-efficient first passes of the data. Then focus can be given to those subgroups by getting a better data source for them to conduct another matching study. It is important to maintain a continued focus on variables that are identifying and susceptible to change over time. Other areas of improvement to consider would be the use of preference weights in the linkage process to give priority to some variables over others. Another is to try different recoding scenarios to see the impact on risk. Yet another is to use different thresholds to see the impact on risk.

Nonetheless, we return to our earlier observation that this study was a very good start, for which we wish to commend the Census Bureau staff who completed it. Our recommendations below are intended to build on this start and to help the Census Bureau develop a disclosure risk and re-identification program for SIPP products.

Longitudinal Data Increases the Risk of Disclosure

Specific to the SIPP, where there are longitudinal data in the public-use file, it is important to include a time dimension in the key variables. In this manner, Li et al. (2021) applied a log-linear modelling approach (Skinner & Shlomo, 2008) to the Survey of Doctoral Recipients public-use files as an example to demonstrate a way to measure the re-identification risk while incorporating the longitudinal nature of the data. To assess the re-identification risk due to longitudinal data, the cross-sectional risk assessment model was re-processed by adding binary change indicators that were derived from variables that were more susceptible to changing values over the years. The variables for which the change indicators were derived from across survey years were academic position, number of children, federal government support, and year of highest degree. Marital status is another variable that would have an important time element to take into account. In another example, El Emam et al. (2011) examine and assess longitudinal risk due to demographics (age, gender, and postal codes) among Montreal residents for data gathered over 11 years by estimating the proportion of individuals who are unique in the population.

Copyright National Academy of Sciences. All rights reserved.

Making Disclosure Analysis an Ongoing Activity

We note that although the probabilistic modelling approach can be quite complex, it is a one-time study so long as the design of the survey remains unchanged. However, if a new sample design is introduced, or new variables or new data products are introduced, then new risk assessments are appropriate. Each data product initially would be assessed for re-identification risk. To start, a re-identification risk assessment can be conducted using the measures discussed above on the current SIPP public-use file to identify subgroups of greatest risk, while incorporating the longitudinal structure and linkable sources of data.

At the same time, it becomes appropriate to update the re-identification studies for the SIPP as external data sources, especially commercially available ones, change in both their scope of coverage and their content. For example, because these databases that contain names and addresses of individuals have recently been used to construct longitudinal data on households on annual changes in income and household composition and structure, assessing the risk of re-identification of these dimensions of SIPP longitudinal data requires monitoring external database development and their capacities.[43]

Finally, the Census Bureau has limited researchers' access to the existing SIPP Synthetic Beta program, which synthesizes SIPP data linked with Social Security Administration earnings and benefit data (see Chapter 2), restricting access to approved research projects. Also, any output produced with the linked Gold Standard data has been subjected to disclosure review. But the Census Bureau might consider including these data in its disclosure risk assessment program. Assessing the SIPP Synthetic Beta data would provide a formal assessment of the disclosure risks of such data. This in turn could provide useful input into expanded use of synthetic methods for either public release files or files that can be accessed in modes other than through Federal Statistical Research Data Centers (FSRDCs), as discussed in Chapter 6. Additionally, such assessments may provide guidance for other modes of access to SIPP linked with these and other administrative record linkage, such as through a remote flexible table generator, as discussed in Chapter 7.

[43]A. Ziff, J. V. Hotz , & E. Wiemers, The Association between the volatility of income and life expectancy in the U.S. (under review); made available by one of the authors.

Copyright National Academy of Sciences. All rights reserved.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 3-1: Risk assessment may be quantified on either an absolute or relative scale, and differential privacy can be viewed as a type of relative risk measure.

Conclusion 3-2: There are several measures of disclosure risk that have been developed and assessed in the literature and they differ with respect to their ease of measurement and their appropriateness for addressing the trade-offs between protecting data privacy and ensuring data usability.

Conclusion 3-3: While analytical approaches to risk assessment can be useful, the SIPP's complex and longitudinal data structure make it important to conduct empirical assessments using resources that would realistically be available to the intruders, various intruder models, and potential data release strategies. Intruder models range from intruders who know a lot about a specific model to those using databases to fish for someone to attack.

Conclusion 3-4: Widely available sources potentially used by intruders include federal administrative data, commercial data, and state program data. It is important to consider all these potential sources, also allowing for changes over time as new data become available and for the potential use of longitudinal data.

Conclusion 3-5: SIPP data are highly granular and may be subject to intruder attacks by matching. Potential identifying variables include age, state, country of birth, occupation, household income, home value and mortgage amount, program participation, demographic characteristics, household composition, and change over time in income, program participation, and household composition. The exact list may vary depending on what is collected through SIPP and what is available through other sources, both of which may change over time.

Copyright National Academy of Sciences. All rights reserved.

Conclusion 3-6: Current knowledge about SIPP disclosure risks is limited. The Census Bureau's recent re-identification study for SIPP is helpful in this regard but still incomplete, not fully accounting for the range of content contained in SIPP and the inclusion of longitudinal data.

Conclusion 3-7: Analyses of disclosure risk conducted by outside researchers may provide valuable perspectives and can be important tools for guiding the improvement of disclosure limitation for the SIPP.

Recommendation 3-1: The Census Bureau needs to establish a systematic and on-going program of assessing disclosure risk for SIPP data products, especially microdata public-use files. This program should include both empirical and analytic approaches to assessing the risks of re-identification attacks, comparisons with a variety of external databases, probabilistic models, and the risks present in both cross-sectional and longitudinal data. All forms of data releases should be evaluated.

This program should include (a) conducting re-identification attacks and matching them with a range of external databases (e.g., federal government databases within the Census Bureau, state and local social program databases, commercial databases such as InfoUSA/Data Axle USA and Experian); (b) the use of probabilistic models to assess probabilities of re-identification for households/individuals with unique combinations of data; and (c) assessment of the disclosure risks associated with within-wave cross-section data and the longitudinal data across waves. If the results of these exercises reveal a set of sensitive variables that are at high risk of disclosure, it will be important to investigate the alternative disclosure limitation methods discussed in the rest of this report, in each case weighing the risk reduction against reductions in data usability. Each of those alternative statistical disclosure limitation methods—including differential privacy mechanisms, partial or full synthetic methods, traditional SDL methods (these are all discussed in the next chapter) and other approaches—should be made the subject of disclosure risk assessments to gauge the reduction in risk that each obtains. These assessments should be considered for both existing and proposed SIPP data products.

Copyright National Academy of Sciences. All rights reserved.

We recommend that the Census Bureau establish a "baseline" assessment of these risks to guide its choices among Disclosure Avoidance Systems methodologies and alternative products, including microdata releases, and data available through Table Generator and Remote Analysis platforms (discussed in Chapter 8), and possibly among data available through secure online data access (discussed in Chapter 6). Furthermore, we note that this Disclosure Risk Assessment system for the SIPP may require periodic re-assessments of disclosure risk to keep up with the development of new or more accessible external databases and (or) new or more challenging intruder attack models.

We think it is important that the Census Bureau find ways to release to users and the public the findings of its assessments of disclosure risk. Thus, we make the following recommendation:

> **Recommendation 3-2: The Census Bureau should find ways to communicate to researchers and the general public the findings of its assessments of re-identification attacks and disclosure risk so that data users will understand the need for disclosure avoidance techniques.**

Finally, while the Census Bureau has the primary obligation for conducting the disclosure risk assessments of the SIPP public release data, we think it is important that the Census Bureau find ways to involve others in improving the Bureau's understanding of these risks. As discussed in this chapter, the academic research community has developed a variety of sophisticated methodologies for re-identification and privacy attacks. Performing only internal risk assessments using standardized strategies can lead to a vast underestimation of risk that does not represent what a creative and resourceful intruder might be able to do. Thus, the Census Bureau would greatly benefit from finding ways to involve external experts in its risk assessment program. This leads us to make the following recommendation:

> **Recommendation 3-3: The Census Bureau should find ways to partner with and involve external researchers and other experts in its risk assessment research program. Such involvement will expand the Bureau's capacity and allow it to tap into state-of-the-art developments in the area of disclosure risk assessment.**

Copyright National Academy of Sciences. All rights reserved.

When appropriate protections are in place, the involvement of external researchers in re-identification studies can be important for making progress in the statistical disclosure limitation of SIPP data products and more broadly. Privacy protection specialists can help in identifying potential attack strategies that intruders might use and in state-of-the-art methods for protecting the data, and SIPP data users can guide the Census Bureau on what data are needed and how the data might be accessed. Additionally, the involvement of data users would help them to feel that they are partners rather than people whose work is being undercut by disclosure avoidance procedures that they do not value. Non-experts such as college students might also be helpful in risk assessment, perhaps bringing a different kind of lateral thinking.

Copyright National Academy of Sciences. All rights reserved.

4
Overview of Disclosure Limitation Approaches

This chapter provides a broad overview of the disclosure limitation approaches that have been suggested in the literature on privacy protection. We also note which ones are currently used for the SIPP. These approaches are not mutually exclusive, but rather may be used in combination. We divide them into two broad classes. One class involves modifying data to protect confidentiality and disseminate widely. This typically involves statistical modelling to preserve important population-level signals and relationships when modifying sample data to yield valid inferences. The second class imposes restrictions on how the data are accessed and how they are approved for release. Typically, there is a tradeoff between these two classes of approaches: data may be modified so that they are considered safe from disclosure and are widely disseminated, or controlled access is provided to data in ways that are safe from disclosure. There may also be intermediate levels that combine the approaches from the two classes, in which some access controls are relaxed while some data modifications are also imposed.

A major consideration, whether modifying the data or offering multiple modes of access, is the impact of such decisions on data usability. Usability includes not only the ability to obtain valid inferences across a wide range of applications, but also includes issues such as equity in access. Maintaining usability is a theme that appears throughout this report because there is a trade-off between data privacy protection and data usability, and neither can be considered in isolation from the other. We discuss usability in depth in Chapter 9 after the various disclosure avoidance approaches have been discussed.

PROTECTING CONFIDENTIALITY THROUGH OFFERING MULTIPLE MODES OF ACCESS

Because the appropriate disclosure avoidance system depends on the mode of access, we first list the possible modes of access. The value of creating multiple tiers of access is motivated

Copyright National Academy of Sciences. All rights reserved.

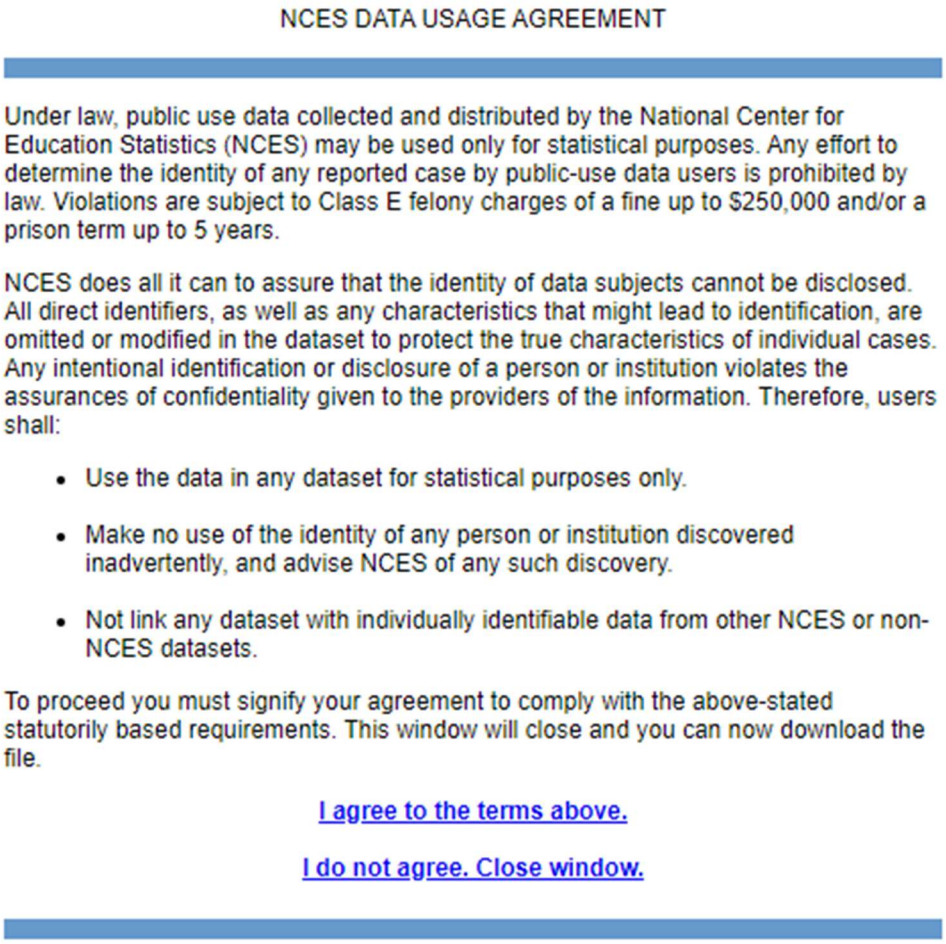
by three factors: the risk elements mentioned in Chapter 3 (e.g., longitudinal structure and integrated administrative data), the balancing of risk reduction with data usability, and satisfying existing users while also broadening the user base. We use the term *access* somewhat liberally in this report to refer to the entire process of working with and releasing the data: getting approval for access to the data, the means used to work with the data, verifying and validating the data, and the process of publicly releasing the data. These are all both important and interrelated, and different tiers can be expected to impose different conditions.

The most common mode of access is through the provision of a public-use file, with records on each individual or household released together with a set of survey-reported variables on each record. A public-use file is typically provided without any check on the identity of the user, although some form of registration for their use can be required. Users typically download these files to their own computing environments for use. While this mode has the widest accessibility, it also carries the greatest threat of disclosure if too many (or highly disclosive) variables are released on each record. However, there are ways of seeking to further protect confidentiality even with public-use files. One way is to require data users to agree to a data usage agreement. For example, the National Center for Education Statistics posts the notice reproduced in Box 4-1 when users seek to access certain of its public-use files.

BOX 4-1

Sample Data Usage Agreement

Copyright National Academy of Sciences. All rights reserved.



NCES DATA USAGE AGREEMENT

Under law, public use data collected and distributed by the National Center for Education Statistics (NCES) may be used only for statistical purposes. Any effort to determine the identity of any reported case by public-use data users is prohibited by law. Violations are subject to Class E felony charges of a fine up to $250,000 and/or a prison term up to 5 years.

NCES does all it can to assure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted or modified in the dataset to protect the true characteristics of individual cases. Any intentional identification or disclosure of a person or institution violates the assurances of confidentiality given to the providers of the information. Therefore, users shall:

- Use the data in any dataset for statistical purposes only.
- Make no use of the identity of any person or institution discovered inadvertently, and advise NCES of any such discovery.
- Not link any dataset with individually identifiable data from other NCES or non-NCES datasets.

To proceed you must signify your agreement to comply with the above-stated statutorily based requirements. This window will close and you can now download the file.

I agree to the terms above.

I do not agree. Close window.

SOURCE: https://nces.ed.gov/surveys/frss/downloads.asp#FRSS30

An agreement of the type shown in Box 4-1 has no formal approval process that a user or the Census Bureau must go through, and the file is available for use immediately by clicking to show agreement to the terms. Given the sensitive nature of much of the SIPP data, the panel does not view this approach as providing adequate protection in itself, and at the least would like to see some type of registration process so that the name and institution of the data user are identified. Also, a registration process would not be fully adequate in the long term, but might be part of a temporizing approach while the Census Bureau develops more elaborate precautions. As a variation on this approach, the Census Bureau could also ask people to register online before downloading the file; this would further emphasize the importance of maintaining confidentiality.

Copyright National Academy of Sciences. All rights reserved.

A second mode is through the provision of synthetic data for variables that are not included in the public-use file. Synthetic variables are constructed using statistical or machine learning models that emulate the distributional and relational properties of the original variables. These synthetic variables can be augmented to the data that have been made accessible under the first mode. Sometimes synthetic variables are created only for a subset of respondents instead of all respondents. An extended version is to create a fully synthetic dataset of all variables for all respondents in the SIPP survey and then release them as public-use files. That is, the original data are completely restricted and never given to the user. We discuss this last mode in Chapter 6 and note there that users usually have the option to query the data custodian for statistics on the original data to ensure that the statistics they obtain on the synthetic data are close to those on the original data. In this mode, the data custodian controls both the construction of the synthetic data and the nature of the statistics released in response to such a user query.

A third mode is secure online data access. The Census Bureau offers one type of such access through the Federal Statistical Research Data Centers (FSRDCs) (see Chapter 2). FSRDCs offer a highly restrictive form of such access, while Chapter 5 presents a much more general (and accessible) type of data access. In this latter model, users must be licensed, but they have online access to selected SIPP variables that are deemed too sensitive to be released or that are highly coarsened in the public-use file. Access requirements or proposal approval processes are set to limit disclosure, but these can offer wider access than is currently offered through FSRDCs, whose use requires considerable processing time, a difficult approval process, and added expenses. Access through FSRDCs may still be needed for SIPP data that are comingled with administrative data, which have other legal restrictions on data access.

A fourth mode is what is known as query-based mode, made available to the general public and without any restriction or screening of users. A specific type of query-based access provides users the ability to request a table of statistics, using a table builder constructed by the data custodian. This mode is sometimes used by the Census Bureau to allow the general public to obtain national summary statistics on demographic and economic variables. As with public-use files, use of this mode is typically unrestricted to the general public, but it allows the data custodian (i.e., the government agency) to restrict the statistics released to provide sufficient privacy protection, and it may include the addition of noise to the estimates as a way of protecting privacy. We discuss table-builders in general and for the SIPP in Chapter 7.

Copyright National Academy of Sciences. All rights reserved.

Currently, the SIPP is primarily provided through the first mode, a public-use file. Users in the general public, without restriction or registration, can download the data to their own computing environments from the SIPP website. The disclosure limitation techniques described below and in Chapter 2 are applied to the dataset before it is posted on the website. However, the Census Bureau also provides data through FSRDCs, which provide greater access to the original data but with different disclosure restrictions (see Chapter 2). The Census Bureau also for several years provided a synthetic data file in which administrative earnings were matched to a subset of survey variables (as we discussed in Chapter 2 and will discuss further in Chapter 6). As of September 30, 2022, access to the synthetic data has been temporarily shut down until a new server location is established, leaving FSRDCs as the only currently available alternative to the public-use files.

METHODS OF DISCLOSURE LIMITATION INVOLVING CHANGES TO THE FILE

The classical, or traditional, approaches in statistical disclosure limitation (SDL) to render sensitive data confidential include these five: (1) *top-coding*, where all values above (or below) a given threshold are set to a fixed value; (2) *recoding*, where variables are coarsened into a smaller number of categories; (3) *variable suppression*, where a variable is deleted from the dataset (for example, when smaller geographical areas are not included); (4) *data swapping*, in which selected values are swapped between similar respondents; and (5) *noise infusion* (e.g., by adding a random increment to the original data) added to selected sensitive variables (Hundepool et al., 2012). These methods historically have been most commonly applied to data released as public-use files. As discussed in Chapter 2, the SIPP currently uses these types of methods for disclosure avoidance (except for data swapping). Some benefits of these approaches are that they are relatively easy to implement (e.g., without requiring new technology or software) and that the Census Bureau and other statistical agencies have a long track record of experience in using them. They have long been considered as sufficient to protect privacy (with the proviso that data are never completely safe from disclosure).

A problem with these traditional SDL approaches is that it is not easy to evaluate how protective they are, using the concepts and approaches described in Chapter 3, and it is sometimes unclear whether the disclosure risk is below a tolerable risk threshold after an SDL technique is applied. Additionally, the impact of these approaches on data utility may not always

Copyright National Academy of Sciences. All rights reserved.

be clear or fully assessed. In principle, it is possible to conduct multiple re-identification exercises to assess the likelihood of disclosure for a given traditional SDL treatment, but this is costly and is not completely reliable unless a formal risk metric is computed. Consequently, the nature and severity of the specific traditional SDL methods used tend to be ad hoc and best guesses by the data custodian on how large the risk is. In addition, most often these traditional methods, when applied to longitudinal datasets, are imposed wave by wave and ignore risk from the use of multiple waves of data and therefore of the longitudinal nature of the data. (In fact, the SIPP data collect longitudinal data within each wave since much of the data have monthly values.)

A method of disclosure avoidance which has the capability of being assessed with formal privacy measures is the general method of noise injection or perturbation, particularly for the sensitive variables. In general, perturbation approaches to protect microdata involve applying a controlled stochastic treatment procedure to replace a subset of the original variable values by perturbed values, with the aim of introducing just enough noise or uncertainty into the microdata to reduce the disclosure risk to an acceptable level. Perturbation also can be applied to aggregate statistics and their generation procedures such as table generators; see Chapter 7. Perturbation methods control change from the original values depending on whether the original variable is categorical or continuous. For continuous variables, perturbation methods generally take the form of generating noise from a given distribution having a mean of zero and a predetermined level of variance, and adding (or multiplying) the noise to the original values of the variable. For categorical variables, perturbation methods take the form of changing values of the variable to other values according to a probability transition matrix and the outcome of a random draw. If perturbation methods are applied to the indirect identifiers, they should be applied prior to calibrating the survey weights.

One approach for noise injection or perturbation is to generate *synthetic data*. The basic idea of synthetic data approaches is to randomly generate data from a statistical or generative model so that the aggregate measures and relationships in the original data are well preserved. Rubin (1993) first proposed the idea of creating fully synthetic data based on the multiple imputation framework. For example, Shlomo and De Waal (2008) use a multivariate or sequential approach to maintain multivariate associations during data synthesis. Yu et al. (2022) generate synthetic data using semi-parametric and non-parametric models. Besides full synthesis,

Copyright National Academy of Sciences. All rights reserved.

there is also partial synthesis (Little, 1993), which selects a subset, either quasi-identifiers or sensitive variables or both, to be synthesized. Although both fully synthetic data and partially synthetic data are generated under the multiple imputation framework, the variance estimation approaches differ (Reiter, 2003; Reiter, 2005). It should be noted that the disclosure risk is higher for partially synthetic data, since true values remain in the dataset, though disclosure risk is not completely removed even for fully synthetic data. Reiter and Mitra (2009) described methods to assess the remaining disclosure risk in the partially synthetic datasets. It is equally important for partially synthetic data to satisfy both the standards on data utility and those on disclosure risk. The synthpop R package (Nowok et al., 2016) on the CRAN network and IVEware (Raghunathan et al., 2016) offer several ways to employ synthetic data methods. We discuss synthetic data more in Chapter 7, where we note that one of the key questions about synthetic data is whether they can faithfully capture all the relationships in the original data that users are interested in.

Other perturbation methods include swapping (Dalenius & Reiss, 1982; Fienberg & McIntyre, 2005). Swapping has been used in some public-use microdata released by government agencies as a means to reduce disclosure risk by introducing "uncertainty" into the microdata. It is designed to switch values of certain fields in a record with data values of the corresponding fields of another record, while controlling the swap pair within strata ("boundary" variables) related to the survey outcome. This creates uncertainty by creating false positive matches that are indistinguishable from true positive matches by the data intruder. Some argue that while the intruder can match the data with the known key characteristics or with an external file, the intruder can never know for sure if the matched data are true when the perturbation rate and variables impacted are kept secret, as they typically are. Often geographies are swapped because it is believed that, conditional on geography, many of the survey target variables are independent.

In general, a combination of classical disclosure limitation approaches is carried out. Often, deterministic perturbation methods are applied to the identifiers and random perturbation methods are applied to the sensitive variables. With the movement toward transparency, which would call for the parameters of the data swapping (e.g., perturbation rates and swapping variables) to be released to the public, swapping might be evaluated further for data protection and data utility.

Copyright National Academy of Sciences. All rights reserved.

Output may need to be reviewed for accuracy or disclosure risk avoidance. When synthetic data are used, it is possible that statistical results from the synthetic data will not be the same as those from the original data. In the case of SIPP data, synthetic data users are given the option to have Census personnel replicate the analysis on what is called the Gold Standard File to determine if the results have changed. Second, agencies may require that output be reviewed by specialists or by a Disclosure Review Board to determine what data can be publicly released.

Differential privacy is an alternative framework that proposes a specific metric for evaluating disclosure risk (see Chapter 3) but is also commonly associated with a specific method that involves noise injection, making a statistical release "indistinguishable" from one in which any one individual's data are removed from the data.[44] This approach carefully injects random noise into statistical computations so as to ensure that the effects of each individual's data are obscured. This leads to a privacy-utility tradeoff, whereby more noise provides stronger privacy protections but less accurate statistics, and less noise provides more accurate statistics but weaker privacy protections. A large research literature has developed around mathematically proving that appropriately perturbed statistical methods satisfy the requirement of differential privacy with a specified level of privacy loss.

The initial focus of the differential privacy literature was on descriptive statistics (Blum et al., 2005; Dwork et al., 2006), but over time the literature turned to more ambitious tasks such as synthetic data generation (e.g., Blum et al., 2013; Bowen & Liu, 2020; Snoke & Slavković, 2018; Tao et al., 2022; Vietri et al., 2022), machine learning (e.g., Abadi et al., 2016; Chaudhuri et al., 2011; Kasiviswanathan et al., 2011), and statistical inference (e.g., Duchi et al., 2018; Kamath & Ullman, 2020; Vu & Slavković, 2009).

Here we briefly mention some of the areas of the differential privacy research and practice that are most salient for SIPP, some of which are covered in more detail in later chapters.

Managing the privacy-loss budget. Mathematical theorems about reconstruction attacks and membership inference attacks (see Chapter 3) tell us that there is an inherent trade-off between privacy, usability, and the amount of statistical information released: if too many statistics about the same set of individuals are published with too much accuracy, then the data

[44]Absolute and relative disclosure risk are alternative metrics. See Chapter 3. Noise injection can also be used to apply to those metrics.

Copyright National Academy of Sciences. All rights reserved.

will necessarily become vulnerable to reconstruction or membership inference attacks (see Dwork et al., 2017). The composition theorems of differential privacy allow us to measure and control the accumulation of risk as more statistics are published. A common practice is to preset an overall "budget" for the maximum privacy loss one is willing to incur, and then manage the releases so that the budget is never exceeded. Alternatively, each extra release over the initial budget has to be evaluated for whether the benefit of the additional release warrants increasing the budget and hence the disclosure risk. This is particularly salient for SIPP, as many data releases are made over time on the same panel. Managing the privacy-loss budget exposes difficult choices about which statistics are the most important to include with high accuracy and which can be omitted or released at a coarser level.

Differential privacy for longitudinal data. The simplest way to manage the privacy-loss budget for longitudinal data is to perform each release independently and use the composition theorems for differential privacy to bound the accumulated privacy loss. However, it is possible for the privacy-utility trade-off to be improved by carefully correlating the releases (i.e., using past releases to help construct new ones), especially for statistics that are not expected to change dramatically between releases. So far, the differential privacy literature has only developed methods for tracking simple statistics (e.g., counts) over time (e.g., Chan et al., 2011; Dwork et al., 2010; Jain et al., 2022), and more research along these lines would be beneficial for SIPP.

Differential privacy for complex survey designs. A common intuition is that using a random sample when collecting or releasing data aids in privacy protection. For differential privacy, there are theorems that demonstrate that indeed the privacy loss is reduced when a differentially private algorithm is applied to a simple random sample (Balle et al., 2018; Beimel et al., 2014; Kasiviswanathan, 2011). However, for complex sampling designs, the situation is more subtle, and in some cases, the differential privacy protections may even degrade (Bun et al., 2022) Thus, more research is needed to understand the impact of the SIPP sampling design on differential privacy, or how the sampling design can be modified to interact better with privacy.

Differential privacy for relational and joined data. The most standard model and implementation of differential privacy applies to performing releases based on a tabular dataset, where each record corresponds to one individual's data (so the opt-out counterfactual amounts to removing one record from the dataset). However, for more complex data (such as from SIPP) that also encodes relationships between individuals or between individuals and other units that

Copyright National Academy of Sciences. All rights reserved.

need privacy protection (like households), it is better to use software that is designed to express and provide privacy protections on relational databases. There is a growing research literature and set of software tools for this setting (see the survey in Near & He, 2021). Some such tools can also support joins between different tables and control the impact of the joins on the privacy loss; this functionality is important for releases that involve joining SIPP data with administrative data.

Differential privacy synthetic data. There are striking theoretical results showing that, in principle, differential privacy can support very rich synthetic data generation. In particular, it is possible to produce differentially private synthetic data that supports accurate answers to exponentially many more queries than would be possible if each query were answered independently with differential privacy (Blum et al., 2013; Hardt & Rothblum, 2010). There has been a substantial body of work aimed at making these differentially private methods more practical, but research challenges remain in tackling high-dimensional complex-structured data like SIPP (see Chapter 7).

Differential privacy query systems. In addition to being used to perform statistical releases (tables, synthetic data, model parameters, etc.), differential privacy can also be used in an interactive query system, through which analysts can issue custom queries (e.g., regressions or models they want to fit) and receive differentially private results. Such systems have an advantage in that the agency holding the data does not need to anticipate what analyses data users will want to perform. However, query systems do raise additional challenges. The most notable is managing the privacy-loss budget. To ensure a given level of privacy protection, the system would track the accumulated privacy loss as each query is issued and stop answering queries once the budget is expended. For this reason, differential privacy query systems may only be practical in more controlled environments, where the number of queries per analyst can be limited and there can be assurance (e.g., through user agreements) that multiple analysts do not collude to exceed their individual limits. In the context of SIPP, a differential privacy query system could be deployed using secure online data access (see Chapter 5) or in a verification and validation server to accompany a synthetic data release (see Chapter 6), or it could be transformed into a noninteractive flexible table builder (see Chapter 7).

Copyright National Academy of Sciences. All rights reserved.

DESIGNING DISCLOSURE AVOIDANCE TO MAXIMIZE USABILITY

Several alternative methods for increasing privacy protection and reducing disclosure risk are available. However, it is important to consider usability when adopting methods for the SIPP data. Usability can be identified through three different dimensions—how accurate results are in a basic statistical sense (accuracy), whether the data can be used to answer significant questions (feasibility), and how accessible the data are to users (accessibility). It is important to assess usability by measuring various approaches for reducing disclosure risk across these three dimensions.

While most people have an intuitive idea of what usability is—the ability to use a dataset and conduct the kind of analyses on it that they want to—the concept is actually fairly complex and requires careful consideration. For example, for the purposes of considering what disclosure avoidance methods to use for SIPP, it is important to quantify usability to the maximum degree possible. This is because all disclosure avoidance methods, including those discussed in Chapters 4–8, will reduce the usability of the data in at least one dimension and possibly more. Because of this, it is important, when considering any Disclosure Avoidance System, to assess how much, and in what ways, it reduces usability. If one has knowledge of the exact nature of this privacy-utility trade-off, a decision can be made on whether the benefits of the reduction in disclosure risk associated with any particular Disclosure Avoidance System are too costly given the expected reduction in usability.

The panel considered three dimensions of usability when evaluating the impact of disclosure avoidance methods on usability:

- *Accuracy*: Validity and reliability of parameter estimates
- *Feasibility*: Ability to conduct analyses to answer significant questions, and
- *Accessibility*: Equal access among users.

The first dimension of usability, "accuracy," refers to the accuracy of parameter estimates generated from the data to answer specific scientific questions. It is perhaps the narrowest and most traditional definition of usability, and it builds on basic concepts of statistics. This dimension is affected by the degree and way noise is added to the data, the way the data are coarsened, and (or) the way synthetic data are produced. Accuracy has implications for the believability of estimates generated from the data.

Copyright National Academy of Sciences. All rights reserved.

The second dimension of usability, "feasibility," has multiple aspects. One is that feasibility refers to the ability of analysts to use the data to investigate the range of scientific questions that the data were designed to answer, which is a much broader concept than accuracy per se. Feasibility is affected by the particular variables made available in data releases, the granularity of such measures (such as whether state identifiers are provided), and the facility with which users can manipulate and analyze the data. But feasibility also includes the ability of users to obtain accurate and valid estimates from the protected files with easily available software. This has consequences for the range and types of substantive questions that may be answered with the data. A final dimension of feasibility, related to the second, is the requirement that users be informed of the nature of any disclosure avoidance systems that have been implemented so that they can take measures to conduct valid estimations and reach valid inferences from the data.

Finally, the third dimension of usability, "accessibility," refers to how difficult it is for potential data users to access the data and generate results. Accessibility may be reduced by policies that make it more difficult or time consuming for users to access the data, such as background checks, prolonged proposal requirements, "seat fees" to conduct research in an FSRDC, and requirements to conduct research in person in an FSRDC. Accessibility is also affected by whether data users need specialized training or software to generate results from the data, which could reduce the number and type of users who can analyze SIPP data. Still another aspect of accessibility involves the disclosure review process. A major advantage of a public-use file is that, once a file has been approved for release, no additional clearance process is required either of the Census Bureau or data users. To the extent that some data users may need to shift their work to other tiers of access, data users will benefit from having a clear and transparent process for data review, knowing what limitations will be placed on the data release, how the disclosure review package should be prepared, and what time constraints will be involved. Overall, reductions in accessibility will likely increase inequality in data access. As accessibility decreases, potential SIPP users with the least resources (in terms of time, money, institutional support, and expertise) will be the most likely to stop using SIPP or be deterred from using it altogether. Details about the usability of various methods suggested in this report are discussed in Chapter 9.

Copyright National Academy of Sciences. All rights reserved.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 4-1: Multiple tools for disclosure avoidance are available, requiring differing levels of sophistication among data users and imposing differing levels of control.

Conclusion 4-2: To the extent that changes to the database are not sufficient to protect confidentiality while also providing reliable statistics, providing multiple tiers of access can be a solution.

Conclusion 4-3: For datasets with large numbers of variables, like SIPP, just adding noise to each sensitive variable prior to public release may bias the analytical results with unknown magnitude and direction. Analysis correcting for additional noise may require information about the noise distribution and special statistical expertise.

Conclusion 4-4: Users will benefit by having information on disclosure limitation methods applied to the data so they will better understand how to interpret results and make valid inferences with the data.

Recommendation 4-1: The Census Bureau should modernize and strengthen disclosure limitation methods applied to SIPP to satisfy standards on both data utility and disclosure risk while moving toward methodological transparency. In particular, the Census Bureau should continually assess how the science and technology for privacy are advancing and consider implementing new tools when advantageous.

Recommendation 4-2: Based on the results from the disclosure risk assessment exercises recommended in Chapter 3, the Census Bureau should evaluate what disclosure avoidance methods are needed to achieve acceptable levels of disclosure risk for each mode of access.

Depending on what data may be available to an intruder, respondents are never completely safe from exposure, and the only completely safe option is to not release the data. However, the panel views SIPP data as being extremely valuable and views the release of the data as providing important public benefits. Statistical agencies historically have sought a compromise—providing strong protections of confidentiality while recognizing that there remains some risk of disclosure. There is no consensus on how to define an acceptable level of risk. Even when the level of risk has been defined through formal privacy, that level has

Copyright National Academy of Sciences. All rights reserved.

sometimes changed.[45] The acceptable level of risk may vary depending on the potential uses of the data, the sensitivity of the data, and the potential harm resulting from disclosure.

Recommendation 4-3: The Census Bureau should evaluate using additional modes of access to provide products with more disclosure protection than public-use files offer but less onerous access requirements than FSRDCs impose.

[45]E.g., see https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html

Copyright National Academy of Sciences. All rights reserved.

5
Disclosure Limitation Approaches: Secure Online Data Access (SODA)

This chapter discusses the potential for creating a new tier of access to the SIPP data that would be intermediate between a public-use file and access through Federal Statistical Research Data Centers (FSRDCs). Such a tier would offer access to more data than the public-use file and less than at FSRDCs, and it would have different mechanisms for determining eligibility for data access and for maintaining data security. A commonly used term for such a tier is a virtual data enclave (VDE). However, the Census Bureau already makes SIPP data available through two mechanisms that might be considered VDEs. To avoid confusion with these other tiers, we instead use the term secure online data access (SODA) to refer to what we propose.

An important context to this suggested change is that systems like SODAs are already used widely to provide access to confidential data and have a record of providing appropriate security and protection for human subject confidentiality. They have been embraced by the scientific community. Such approaches are in use by other federal agencies in a way that can serve as a model for the Census Bureau processes for protecting SIPP confidential data. For example, the Bureau of Labor Statistics (BLS) asks people to sign a BLS Agent Agreement to gain access to its confidential data accessed using a VDE. The Letter of Agreement establishes the data user as a temporary BLS agent, and no additional work (such as special sworn status) is needed. (The Census Bureau is subject to Title 13, which requires such temporary staff to swear to observe the limitations imposed by Title 13, but does not specify the nature of how the agent is sworn in.) Additionally, a SODA-like approach is used by federal agencies including the National Center for Education Statistics, the Bureau of Justice Statistics, the United States Department of Agriculture, and the Millennium Challenge Corporation. The Food and Drug Administration and National Institute on Drug Abuse operate a VDE to provide access to the restricted use files from the Population Assessment of Tobacco and Health. NORC at the University of Chicago, Westat, and the Inter-university Consortium for Political and Social

Copyright National Academy of Sciences. All rights reserved.

Research are several examples of organizations contracted to operate VDE infrastructure for various data and partners and to provide access to confidential data to authorized users.

SODAs also have an important role in fulfilling the requirements of the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act), which formalizes a goal of increasing access to data, and the Information Quality Act, which makes tiered access a tool for supporting that move.

Currently, SIPP data are primarily offered through two vehicles: a public-use file and FSRDCs. Data access through FSRDCs originally required the researcher's physical presence at an FSRDC, but starting with the COVID-19 epidemic and later as a more permanent change, access to Title 13 data at FSRDCs (including SIPP) is now also available online, though with the same approval process as for physical attendance.[46] Thus, data access has now been expanded to three tiers. An additional tier, but much more limited in terms of the amount of SIPP data that are available, is the SIPP synthetic beta;[47] in this tier, the SIPP synthetic data can be accessed online, with a somewhat different (and easier) process for gaining approval for access than for FSRDC access.

The reason for considering a new tier of access is that, given the risks of high data granularity discussed in Chapter 3, it seems likely that the public-use file in its current format will soon be found to be too disclosive, either based on current data availability or the general trend toward more types of data becoming available. Though a reduced public-use file could still be useful to many researchers, the FSRDC process is too onerous, costly, and time-consuming to be practical for those researchers whose data needs would not be met. SODA can provide a solution, but its success will depend both on data users' motivations (e.g., whether the public-use file will meet their needs, and what other databases are available as alternatives) and on the particular conditions and tools that the Census Bureau provides for accessing the data through SODA. If SODA access requirements are largely the same as FSRDC access requirements while the public-use file becomes less useful, then SIPP may face a precipitous loss of users. It will be important to monitor user access and the level of burden imposed upon users, and to design a system that supports access while providing data security. Multiple factors are important,

[46]Such access may not be available when other data are comingled with the SIPP data, depending on the requirements set by the agencies providing the data.

[47]At the time of this writing, access to the SIPP synthetic beta has been suspended until a new host site is established.

Copyright National Academy of Sciences. All rights reserved.

including the initial process of approval for access, the mechanisms available for access, the presence of tools to simplify analysis, and the disclosure review process before data can be released.

We note that a long-standing concern of the Census Bureau has been to maintain the confidentiality requirements specified in Title 13. In this particular case, where the move would be to make some data more secure (moving from a public-use file to a more secure tier) rather than making FSRDC data less secure, it is not clear whether any changes to Title 13 procedures are required. However, part of Chapter 9 also discusses how the policy interpretation of Title 13 is actually much stronger than what is required by Title 13 itself. Thus, there is also a possibility that Title 13 could be reinterpreted in a way that would recognize the federal government's move toward increasing tiers of access and promoting the provision of data for evidence building.

PRIVACY PROTECTIONS IN THE SODA

A SODA would be very valuable for providing access to Title 13 data for several reasons. First, a SODA places a less onerous burden on the data requestor than the FSRDC, as it involves less credentialing work in becoming authorized for data access than using an FSRDC requires. Researchers seeking to use data in an FSRDC currently must obtain a Special Sworn Status, which requires applicants to meet U.S. citizenship and residency requirements and to pass a background check administered by the Department of Homeland Security (DHS). Applicants must also submit fingerprints, complete online training, submit residential, travel, employment, and education histories, provide references, and complete an interview.

SODA user credentialing can incorporate a system and protocols to vet users and monitor the terms of their controlled access. Users generally submit applications that demonstrate their need for restricted data and the scope of their analysis plans. The approval process may also consider applicants' statistical training, educational background, institutional affiliations, and titles, and it may require a supervisor as a co-applicant (a faculty advisor in the case of student applicants). Requiring Institutional Review Board approval or exemption and a signed Data Use Agreement by an institutional representative adds additional layers of security and accountability, though they can become administrative barriers to nonacademic researchers. The Association for the Accreditation of Human Research Protection Programs, Inc., offers an accreditation process for organizations, including independent Institutional Review Boards,

Copyright National Academy of Sciences. All rights reserved.

which can offer ethical oversight for those wishing to access SIPP data in SODAs who are not at academic institutions.

ENHANCING EQUITY IN THE SODA

It is also possible to augment the application process with additional user training, which would be valuable to creating researcher behaviors that keep the SIPP data from being disclosed. There are other difficulties inherent in FSRDCs, such as (1) a limited number of seats, (2) limitations in computing power, thus slowing responses, and (3) high cost to the applicants. A SODA stands to offer a much broader set of researchers with quicker access to SIPP confidential data while maintaining Census Bureau authorization of users and disclosure review controls applied to any output. Expanding access through SODA will alleviate the additional administrative and cost burden of applying to multiple physical research data centers for multi-site research teams, and will ease the transition if researchers change institutions during the project. A SODA is not limited by physical computing space (for example, using secured cloud computing) and can more easily expand its computing capacity in response to shifting researcher demand, though SODA users will still be subject to disclosure risk review on their analysis output. Thus, the SODA approach would be more equitable across different user and research communities as compared with those cost and experience barriers that are associated with FSRDCs. The process in gaining access would necessarily be more involved than simply downloading a dataset, but institutions such as the Inter-university Consortium of Political and Social Research (ICPSR) have worked to make data both secure and accessible.

There is another equity issue other than access but related to disclosure avoidance. Minority groups of all types tend to be at greater risk of disclosure simply because of their smaller numbers, and thus their greater probability of having unique combinations of characteristics. If one were to apply disclosure avoidance tools differentially to those groups in order to protect their confidentiality, then one might also be weakening statistical analyses concerning these groups. SODA helps to address this issue as well by providing a secure way of accessing the data without the data being perturbed.

Copyright National Academy of Sciences. All rights reserved.

SODA CAPABILITIES

A SODA environment simulates a non-Internet-connected desktop computer and can be configured to meet security requirements for most confidential data. The SODA would ensure secure access to SIPP data files that could be reinterpreted as controlled unclassified information. Thus, the SODA would then need to adhere to CIPSEA and NIST 800-53 (or NIST 800-171) confidentiality and security requirements for controlled unclassified information in a federal system. SIPP data accessed through a SODA would still be restricted from general dissemination, protecting the confidentiality of the individuals and organizations represented in the data. From a technological perspective, a SODA could be similar to a virtual data enclave provided through an FSRDC; the real difference would be in the application process and the disclosure review process.

A SODA provides authorized users with a mechanism with technical controls to securely access confidential data from their local desktops. It is a virtual environment launched from the researcher's desktop—operating on a remote server meeting FISMA security controls or in the FedRAMP Cloud-based environment. For the user, launching the SODA is akin to remotely logging into another physical computer or secure network, something that many researchers are accustomed to doing. The virtual environment is isolated from the users' physical desktop computers, restricting users from emailing, printing, copying, or otherwise moving files outside the secure environment. Therefore, the user never has access to a physical copy of the data; the data remain within the SODA, giving the organization managing the environment the ability to monitor data access and terminate access within minutes. A SODA can also be configured to provide access at specified locations, such as through an approved IP address. In a SODA, desktop security and data protection are centralized and applications can be quickly added, deleted, upgraded, and patched. The SODA is both useful and cost-effective for providing a wide array of desktop editing and statistical software.

A SODA can be configured to restrict researcher access to the datasets they are approved to use and can support combination with other publicly approved datasets and other restricted Federal Statistical Agency datasets. The virtual environment can be deleted when the researcher logs off of the SODA, and a new virtual environment is created each time the researcher logs back on to the SODA. A SODA can be configured to offer a secure workspace for each approved

Copyright National Academy of Sciences. All rights reserved.

project, one that is isolated from others' access. Thus, authorized users retain access to any files previously saved in the SODA, but not to the workspaces. The SODA also facilitates collaborative research with multiple individuals within and across institutions being able to access a shared folder for their work.

A SODA also allows an agency to manage legal and social controls for accessing the data. SODA access is supplied to authorized users only after Census Bureau user approval is in place (such as through a Census Bureau Agent Agreement). A SODA can be configured to provide agreement tracking with notifications for expiring agreements to manage compliance. Data access can be suspended quickly. Authorized users request output and files that they want to receive outside the SODA to be reviewed against a set of disclosure review guidelines (either by a Census Bureau Disclosure Review Board or supported with automated checking). Census Bureau staff or a designee conducts a disclosure review of the output. It is possible to add disclosure protections to the output, including a differential privacy process. The cleared output will have no confidential information. Finally, SODA access can automatically terminate at the end of the agreement period (typically one year) unless the SODA user applies and is approved for an extension.

REQUIREMENTS FOR ORGANIZATIONS MANAGING A CENSUS BUREAU SODA

The Census Bureau's infrastructure solution ideally would be connected to government efforts to streamline access to its data. The Evidence Act, enacted in January 2019, which has the goals of advancing data and evidence-building functions in the federal government, has already made it easier to access government data in ResearchDataGov.org, a single application portal for discovering and requesting access to restricted microdata from federal statistical agencies. Another outcome of the Evidence Act is accelerated planning for a National Secure Data Service hosting a secure infrastructure where researchers and other users (such as program evaluators) would be able to: (1) submit proposed projects for approval; (2) link and access data for research and analyses; and (3) have project results privacy-protected, then prepared for public dissemination (Potok & Hart, 2022).

We note that creating and maintaining SODA will require some cost and effort on the part of the Census Bureau. This includes creating a mechanism for approving people for access (even if it is simpler than at FSRDCs), creating and maintaining the hardware and software

Copyright National Academy of Sciences. All rights reserved.

infrastructure, and reviewing statistical findings for potential release. The panel has not attempted to estimate those costs, though the timeline presented in this report makes allowance for the development time needed. Possibly the National Secure Data Service (NSDS) will be helpful in this regard. To the extent that SODA allows access to the actual data rather than to synthetic data, SODA has the advantage of not requiring a separate verification and validation process, thus avoiding one source of burden on the Census Bureau; still, the results will need to go through disclosure review. Ideally, some parts of the disclosure review process might be automated, though doing so for complex analysis would require a long-term development process. The Census Bureau might also consult with other agencies (including through NSDS) and external organizations to find ways of lessening the burden.

The Census Bureau might also individually implement security and confidentiality requirements for its own SODA infrastructure using a contractor as a way of reducing the burden on the Census Bureau and potentially making use of systems that are already in place. The Census Bureau would require any organization under contract to select, train, and maintain rigorous security processes for employees who have access to the SIPP confidential data (documenting employee training in confidentiality and security procedures) and have the ability to destroy data at the end of any contract period. An organization contracted to manage the SODA infrastructure would also be required to maintain a secure worksite and system to prevent unauthorized access. A contractor would implement security controls and maintain documentation and security controls for any system holding SIPP confidential data, obtaining a FISMA moderate Authority to Operate.[48] The Census Bureau may use security inspections to ensure that the system and worksite where the data are stored are in compliance and secured. Sometimes the Census Bureau's use of private contractors has gone poorly, especially when the contractors were working outside of their areas of expertise or lacked sufficient infrastructure for the large-scale operations that are sometimes needed (Carless, 2020; Ruggles & Magnuson, 2020). If the Census Bureau makes use of a private contractor, it is important to verify the contractor's experience and capabilities for maintaining a SODA. If a contractor is given

[48]FISMA scores are used to grade the security of a company's or agency's internal systems, with three levels of security: low, moderate, and high. The standards are set by the National Institute of Standards and Technology as part of its statutory responsibilities under the Federal Information Security Modernization Act of 2014 (FISMA). See https://doi.org/10.6028/NIST.SP.800-53r5

Copyright National Academy of Sciences. All rights reserved.

responsibility for disclosure review, then it may be that certain types of decisions could be delegated completely to the contractor, while others require consultation with the Census Bureau; even in the latter case, the contractor may be able to organize the request in a way that lessens work on the Census Bureau.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 5-1: Secure online data access is a viable method for protecting confidentiality for access to a more comprehensive set of SIPP data than would be available in the public-use file.

Conclusion 5-2: There are elements of SIPP synthetic and virtual data access (provided through FSRDCs) approaches that could be used to develop new modes of access to protect confidentiality and limit disclosure.

Recommendation 5-1: If disclosure risk assessment studies find that the current public-use file does not provide adequate disclosure avoidance, the Census Bureau should consider secure online data access as a mode likely to support both access and security.

Copyright National Academy of Sciences. All rights reserved.

6
Disclosure Limitation Approaches: Synthetic Data

OVERVIEW OF DATA SYNTHESIS

Data synthesis is a technique for statistical disclosure limitation, as discussed in Chapter 4. It generates individual-level data (microdata), synthesized based on information in the original dataset. The resulting surrogate data have the same structure as the original data and are, generally, generated using predictive modelling. If the data synthesis model preserves the statistical properties and population-level signals of the original data, then inferential approaches for analyzing synthetic data are appropriate, and the approach enables data users to make valid inferences for a variety of analyses.

Synthetic data generation as a disclosure limitation approach, which was originally proposed by Rubin (1993), is referred to as the *full synthesis* approach. The idea of full synthesis uses a standard missing-data paradigm, by which survey outcomes for individuals not sampled in a survey are treated as missing and are multiply imputed using predictive modelling. This generates synthetic populations. Samples from these synthetic populations are then released as public-use files. Given the large population size in the synthetic populations, the original survey values are excluded when drawing these samples. Valid inferences can be constructed when these public-use files are analyzed properly using the combining rules discussed in Raghunathan et al. (2003). It has been argued, however, that in certain cases a single implicate could be sufficient (see Klein & Sinha, 2015; Klein et al., 2019; Raab et al., 2018).

It may not be necessary to fully synthesize a survey dataset to achieve the desired level of disclosure risk. Little (1993) proposed the *partial synthesis* approach, where only a subset of variables on sampled individuals are synthesized. Assume that the variables in a dataset can be

Copyright National Academy of Sciences. All rights reserved.

categorized as either quasi-identifiers or sensitive variables.[49] The pseudo-identifiers are characteristics of individuals (such as sex, race, and Zip code) that are available to data intruders in publicly available sources. Sensitive variables are the responses collected in a study or a survey and may include sensitive information about individuals. Partial synthesis methods may focus on only synthesizing sensitive variables so that intruders would only know the synthetic rather than the actual responses, or synthesizing only quasi-identifiers to prevent record linkage through them, or both. Synthesizing only quasi-identifiers is better in terms of retaining information, especially when response variables outnumber quasi-identifiers significantly, since only a smaller set of variables would be sanitized to protect the information of a much larger set. Reiter (2003) developed combining rules for partially synthetic datasets.

Little et. al. (2004) proposed *selective synthesizing (partial synthesis)*, where we synthesize for only a selected subset of the sample individuals, such as those at high probabilistic disclosure risk or the cases deemed most likely to be identified by data intruders. For example, synthesizing quasi-identifiers for selected individuals may involve probabilistic swapping of quasi-identifiers between paired cases. Synthesizing selected cases in the sample incurs less information loss and reduced sensitivity of inferences to synthesis model misspecification than partially synthesizing all cases, at the expense of weakening the guarantee of privacy protection. As with the full synthesis approach, multiple synthetic sets of identical structure can be released using partial synthesis to propagate the uncertainty arising from the synthesis process.

Figures 6-1 to 6-3 illustrate the three versions of synthetic datasets (taken from Raghunathan, 2016).

- Figure 6-1 illustrates fully synthetic data where all the variables are synthesized on all respondents. This is generally a very difficult task for a large and longitudinal dataset like SIPP.
- Figure 6-2 illustrates partially synthetic datasets where a selected set of variables that are deemed sensitive or key to potential identification are synthesized on all respondents in the survey. The list of variables to be synthesized may be

[49]The distinction between quasi-identifiers (QIDs) and sensitive attributes have increasingly blurred. Sensitive attributes can be used as QIDs and some QIDs may also contain sensitive values. In other words, QID and sensitive attributes may vary case by case and differ in their context (see Narayanan and F. W. Felten (2014); Sei et al. (2017); and Wong et al. (2019).

Copyright National Academy of Sciences. All rights reserved.

determined, for example, based on a re-identification study, as described in Chapter 3. Suppose S is a set of variables to be synthesized and P is a set of variables that can be included in the public-use file. The partial synthesis can be created based on the prediction modelling, $f(S|P)$.

- Figure 6-3 illustrates a situation where the variables in S are synthesized for a subset of respondents. The approach is same as in the partially synthetic data, namely, based on a prediction modelling, $f(S|P)$, but only based on the subset. The subset is determined based on the level of individual risk of disclosure. Note that the synthesis of the full sample data (that is, everyone in the sample is regarded as a sensitive subject) without using any population information is a special case of the scenario depicted in Figure 6-3. This is also referred to as full synthesis in the literature.

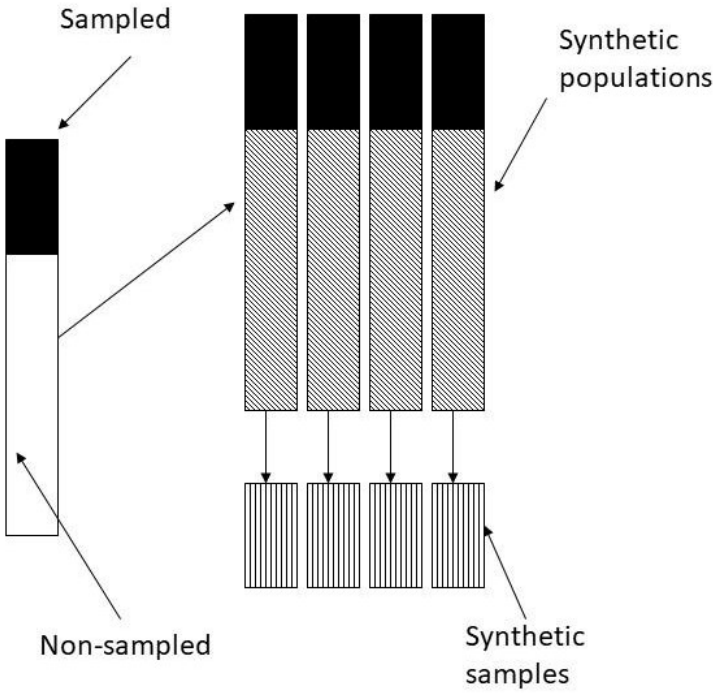


Figure 6-1 Fully synthetic data.

SOURCE: Raghunathan, 2016.

Copyright National Academy of Sciences. All rights reserved.

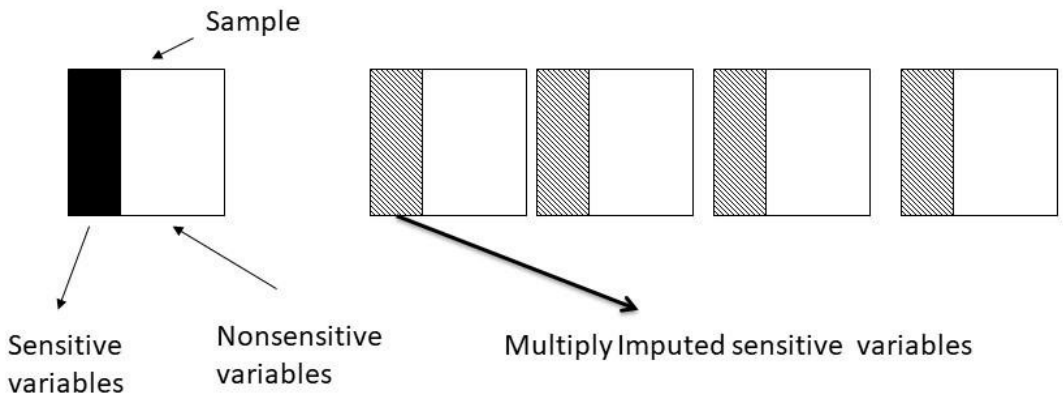


Figure 6-2 Selected variables are synthetic.

SOURCE: Raghunathan, 2016.

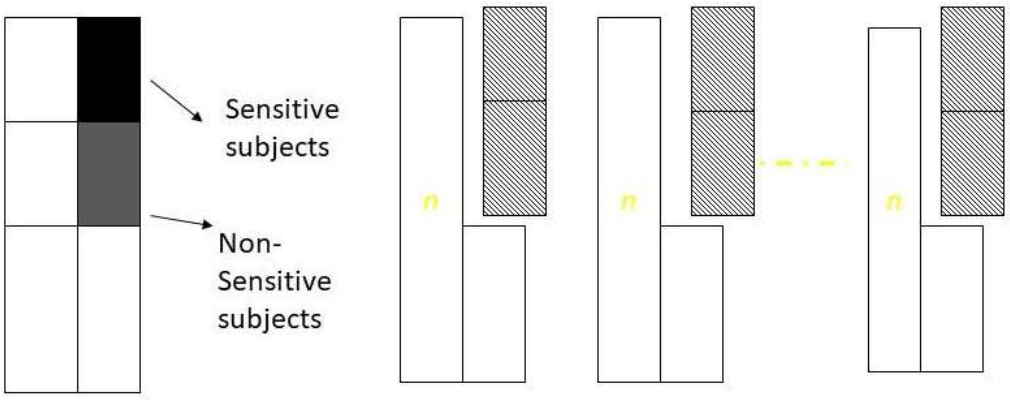


Figure 6-3 Variables are synthesized for selected respondents.

SOURCE: Raghunathan, 2016.

In summary, "synthetic datasets, generated to emulate certain key information found in the actual data and provide the ability to draw valid statistical inferences, are an attractive framework to afford widespread access to data for analysis while mitigating privacy and confidentiality concerns" (Raghunathan, 2021).

In addition to the information provided in this chapter, Appendix B provides technical information about inferences based on multiple synthetic data.

Copyright National Academy of Sciences. All rights reserved.

DATA SYNTHESIS APPROACHES

Many approaches and techniques have been developed in the past several decades for both partial and full synthesis. We loosely group the data synthesis methods into two categories: *distribution-free* or *non-parametric* approaches and *distribution-based or parametric* approaches. In the former, a data synthesizer is constructed without making assumptions that the sample data are drawn from a particular probability distribution, and in the latter it is constructed based on an assumed parametric distribution or a statistical model for the sample data.[50]

A straightforward distribution-free approach is to first estimate the probability distribution of the sample data using a non-parametric method, such as histograms and kernel density estimation. Then samples are generated from the estimated distribution to create synthetic data. Both histograms and kernel density estimation are well studied and have established theoretical results. Despite their simplicity, they are known to be subject to the "curse of dimensionality."[51] Reiter (2005b) examines the feasibility of using CART (Classification and Regression Trees) to generate partially synthetic data.[52] Burgette and Reiter (2010) propose using regression trees as the conditional models in the sequential regression approach to impute missing values, which yield more reliable imputed data than standard parametric sequential regression techniques. The technique was later used to generate synthetic data in Bonnéry et al. (2019), Drechsler and Reiter (2011), El Emam et al. (2020), Quintana (2020), and Raab et al. (2018), among others.

Sequential trees as a data synthesis have the advantage in that datasets can be heterogeneous in the variable types, even with a significant amount of missing values, without making any assumption about the distribution of each attribute.

An and Little (2007) propose a non-parametric stratified hot-deck procedure through creating strata and drawing removed quasi-identifiers with replacements from each stratum, and they tested the approach in longitudinal studies. Pistner et al. (2018) propose using quantile

[50]There are other categorizations along similar lines, such as parametric vs. non-parametric approaches. Note that some data synthesis approaches are distribution-free but are model-based and contain parameters, such as deep generative models.

[51]As the number of dimensions increases, the available data become increasingly sparse, making it difficult to obtain meaningful estimates.

[52]CART works by recursively partitioning the sample data into smaller subsets given the observed attributes and creating a tree-like model that can be used for prediction or classification of an outcome.

Copyright National Academy of Sciences. All rights reserved.

regression to address the heavy-tailed heteroscedastic nature of the Longitudinal Business Databases (LBD) to create a version of synthetic LBD, and compare their proposal in both utility and privacy with the use of CART. Yu et al (2022) use non-parametric and semi-parametric regression techniques to generate longitudinal synthetic data.

Modern data generation based on deep learning techniques also belongs to the distribution-free category, as it does not make any assumptions about the underlying probability distribution for the sample data. These deep generation techniques work with various types of data, such as numerical data, categorical data, and unstructured data (such as graphs, networks, and text).[53] The techniques have been shown empirically, in numerous experiments, to be capable of generating synthetic data that resemble the original. On the other hand, they do involve training large deep neural networks with tens of thousands of parameters, and there is still a lack of understanding on *why* they work.

In contrast to distribution-free synthetic data approaches, distribution-dependent data synthesis starts with the specification of an appropriate statistical model or a probability distribution for the observed data, referred to as the *synthesis model*, based on which synthetic data will be generated. This method then estimates the parameters in the model or distribution given the data, and finally generates data from the model given the estimated parameters. To account for the uncertainty around the estimated parameters in the synthesis model, a Bayesian approach can be useful and easy to implement as well. In such an approach, first one draws model parameters from their posterior distribution, and then one builds synthetic data from the distribution assumed for the observed data given the drawn posterior samples in the previous step. When a dataset has a large number of variables, including longitudinal data, synthetic data can be generated using sequential regression (Raghunathan et. al., 2001). The order in which the sequential regressions are performed can be randomized. The sequential modelling framework has been successfully implemented in practice for missing data imputation and data synthesis, for example for IVEware (Barrientos et al., 2018; Kinney et al., 2011; and Raghunathan et al., 2016;) and is currently the method behind SIPP Synthetic Beta.

[53]For example, variational autoencoders (VAE) (Kingma & Welling, 2013), generative adversarial networks (GAN) (Goodfellow et al., 2014), normalizing flows (NF) (Papamakarios et al., 2021; Rezende & Mohamed, 2015), and diffusion models (Ho et al., 2020; Song & Ermon, 2019).

Copyright National Academy of Sciences. All rights reserved.

Little et al. (2005) provide a couple of approaches for model-based partial synthesis of quasi-identifiers. The first approach is to synthesize the quasi-identifiers in a mixed set of individuals who are at high disclosure risk and who are at low risk, where the latter group is similar to the former in terms of the sensitive attributes based on a similarity measure, such as the Mahalanobis distance. Synthesis models need to be carefully defined, since the selection of the low-risk group is not random. An alternative is a model-based probabilistic swapping of quasi-identifiers between individuals who are at high disclosure risk with individuals who are at low risk, both groups having similar or matched values on sensitive attributes.

Distribution-dependent approaches may not be as robust as distribution-free approaches, in the sense that synthetic samples would be biased if distributional assumptions imposed on the observed data are inappropriate. Regardless of whether a synthesis model makes distributional assumption about the sample data or not, to enable analysts to perform unbiased analyses on synthetic data, the synthesis model needs to be "expansive" and must include, as submodels, all models that may be run by analysts on the synthetic data. In the literature on missing data, it is well known that if the imputer model is "uncongenial" (Meng, 1994), for example, the model to create multiple imputations is a submodel of the one used by an analyst, which is the true model, then inferences based on the imputed data in the analyst model would be invalid. On the other hand, if the imputer's model is more general and includes the analyst model as a special case, then the inferences based on imputed data, though consistent, may not be as precise if the information from the imputer's model is not shared with the analyst or incorporated in the analyst's model.

Disclosure Risk

Work has demonstrated that synthetic data, even fully synthetic data, may still be subject to privacy attacks and are not immune to privacy leaks, although the risks may be different case by case. Privacy attacks on synthetic data can be membership attacks that determine whether a particular individual (target) is in the original sample used to generate the synthetic data. Other privacy attacks on synthetic data include re-identification attacks, attribute inference attacks, where an attacker attempts to learn information on sensitive attributes of individuals, and reconstruction attacks, where an attacker tries to reconstruct the original data from the synthetic

Copyright National Academy of Sciences. All rights reserved.

data. With advancements in machine learning and the availability of vast amounts of public data, many of the attack tasks can be performed using only the released synthetic data and without knowing the details of the synthesis model; such attacks are known as black box attacks.

Reiter and Mitra (2009) show that the partially synthetic data with quasi-identifiers being replaced by synthetic values are still subject to re-identification attack. Zhang et. al (2022) formulate a membership attack by learning useful representations of observed data without specific assumptions about the generative model. They perform such attacks against synthetic health datasets and find that the records in synthetic data that have a one-to-one correspondence with real records in the population are vulnerable to membership inference at a very high rate. By contrast, synthetic records that are random samples from a distribution and generative model without direct mapping to real records are only marginally susceptible to attack and, in most cases, could be deemed sufficiently protected from membership inference. El Emam et al. (2020) argue that if a model synthesizer is overfitted to the original data, a synthetic record, even in fully synthetic data, can be mapped to a real person in the population, who is thereby subject to disclosure risk. Deep generative models have also been demonstrated to pose serious privacy risks, because they inadvertently memorize some of their training data (Carlini et al., 2021, 2023; van den Burg & Williams, 2021).

To make synthetic data safer and more robust against privacy attacks, it is important to carefully design data synthesis algorithms and carefully incorporate privacy concepts, such as differential privacy, into the data-generation process.[54] Blum et al. (2008) show that in theory there exists a differentially private algorithm that takes a dataset and outputs a synthetic dataset that preserves an exponential number of statistical properties of the original data. However, there is a computational bottleneck to implement such methods in full generality (Ullman & Vadhan, 2020). If one is willing to make some assumptions regarding the distribution of the data or to give up some accuracy on some statistics, differentially private synthetic data of good utility can still be generated, even for data with more complex structures (e.g., see Nixon et. al, 2022; Snoke & Slavković, 2018).

Indeed, many techniques and methods have been developed for releasing differentially private individual-level data in the past decade. In the multiple synthesis framework without

[54]It should be noted that incorporation of differential privacy in data release does not imply that privacy risk is reduced to zero even if privacy loss approaches 0. See an example in Liu and Zhao (2022).

Copyright National Academy of Sciences. All rights reserved.

formal privacy guarantees, the more sets of synthetic data are released, the higher the disclosure risk becomes by aggregating information across the multiple sets. However, when undertaking data synthesis with differential privacy guarantees, the privacy loss of each release can be prespecified and the overall privacy loss over multiple releases can be bounded and controlled using composition theorems for differential privacy. Like the distribution-dependent synthesis approaches without differential privacy, the methods with differential privacy guarantees can also be roughly grouped into distribution-free approaches and distribution-dependent approaches.

When the original data are numerical, a straightforward approach for differentially private data synthesis (DIPS) is to sample from privacy-preserving histograms constructed from the original data, such as by using a perturbed histogram or smoothed histogram approach (Wasserman & Zhou, 2010).[55] To generate privacy-preserving histograms, one can perturb the bin counts in the original histograms using a differential privacy mechanism, and then synthesize numerical data by drawing a bin according to the relative perturbed frequencies of the histogram bins and then sampling data from the uniform distributions bounded by the sampled bin endpoints in the previous step. Histogram-based approaches have several limitations. First, to form histograms on the numerical data, discretization is necessary (Scott, 2015). In addition, these approaches suffer from a curse of dimensionality. Specifically, there can be a huge number of data bins/cubes if high-order interactions exist among the data attributes and are taken into account when the histogram is generated; this results in many bins with small or zero counts, which become obscured when noise is introduced. Hall et al. (2012) propose perturbing kernel density estimators with properly calibrated noise to yield differential privacy guarantees, from which synthetic data can be generated. Like histograms, kernel-based density estimators also suffer from the curse of dimensionality, a problem that is only exacerbated after factoring in differential privacy.

When the original sample is categorical, synthetic data can be generated through differentially private, low-dimensional contingency tables. While one could construct a full-dimensional contingency table across all the variables in the original data, it is highly unlikely

[55]Wasserman and Zhou (2010) also propose generating data from differentially private empirical cumulative density functions. This approach does not seem to be a viable choice in practice (Bowen & Liu, 2020). One difficulty lies in defining the set of all possible candidate cumulative density functions, the size of which increases rapidly with sample size or number of attributes, making the synthesis process computationally challenging and unrealistic for a large or high-dimensional dataset.

Copyright National Academy of Sciences. All rights reserved.

that high-order interactions among attributes would be meaningful either practically or statistically; in addition, a high-dimensional table would also have sparse or small cell counts. To determine the set of the low-dimensional contingency tables, one can choose an order (e.g., 2-way or 3-way) based on prior or domain knowledge or apply model selection procedures. After the set of tables is selected, one can apply various randomized mechanisms to sanitize the tables with differential privacy guarantees, from which synthetic data can then be generated (Abowd et al., 2022; Barak et al., 2007; Eugenio and Liu, 2019; Hardt et al., 2012; Hay et al. 2009; Nixon et al., 2022; Qardaji et al., 2014). There are several key considerations in this process, such as allocation of privacy budgets across the sanitization of different tables and ensuring count consistency across the tables (either through post processing or through constraint optimization, both of which are employed in the TopDown Algorithm by the U.S. Census Bureau to achieve disclosure avoidance in the 2020 Decennial Census) (Abowd et al. 2022).

Procedures for differentially private deep generative models have also been developed, such as differentially private generative adversarial networks (GANs) (Xie et al., 2018; Xu et al., 2018), GANs and PATE (Private Aggregation of Teacher Ensembles) combined to generate privacy-preserving synthetic data with differential privacy guarantees (Jordon et al., 2019), differentially private VAEs (Takahashi et al., 2020), and differentially private NFs (Lee et al., 2022; Su et al., 2023; Waites & Cummings, 2021). Due to the iterative nature of these algorithms, privacy loss is tracked over iterations and can use the accounting methods developed for differentially private deep learning (Abadi et al., 2016; Bu et al., 2020). Some of these methods have been applied to real data with some success. The downside of the deep generative models is that they often require a large amount of data to generate synthetic data of good utility for a given practically reasonable privacy budget, and the quality of the synthetic data can be sensitive to the specifications of algorithmic parameters used during iteration.

Distribution-dependent differentially private synthesis approaches are based on an assumed distribution or an appropriately defined probabilistic model given the original data (Bowen & Liu, 2020). For categorical data, a well-known model-based approach for differentially private data synthesis is PrivBayes (Zhang, 2017), which uses a low-dimensional distribution to approximate the full distribution of all variables in the original data by exploring the conditional independence among them rooted in Bayesian networks. The low-dimensional distribution is then perturbed to obtain a differentially private distribution from which synthetic

Copyright National Academy of Sciences. All rights reserved.

data can be generated. There are also distribution-dependent differentially private data synthesis approaches in the Bayesian framework, such as the Multinomial-Dirichlet synthesizer for categorical data (Abowd & Vilhuber, 2008), sanitization of sufficient statistics in an appropriate Bayesian model for a given sample dataset (Liu, 2022; Park et al., 2020), and differentially privately tempering of the posterior distribution (Foulds et al., 2016; Wang et al., 2015). Synthetic data can be generated by first drawing parameters from the differentially private posterior distribution and then samples from the assumed distribution for the original data, given the drawn parameters. Alternatives to sanitizing a posterior distribution include approaches that incorporate differential privacy in a posterior sampling algorithm that leverage inherent randomness to achieve privacy guarantees (e.g., MCMC; see Balle & Wang, 2018; Heikkila et al., 2019; Li et al., 2019; Yıldırım & Ermis, 2019) and in the analytical approximations of a posterior distribution (that is, differentially private variational distribution; see Jalko et al., 2016; Park et al., 2020; Sharma et al., 2019; Su et al., 2023).

Differential privacy can also be incorporated in sequential regression modelling. To achieve differential privacy guarantees, each regression model used during the synthesis process, which can be a parametric model, a non-parametric model, or a deep generative model, needs to be made private. Despite the flexibility, there are a few potential drawbacks for sequential regression models in the differential privacy setting for generating synthetic data. First, the total privacy budget is divided across all the models employed in the procedure. Within each model, the budget may be further divided, depending on the type of model. Therefore, if there are a large number of attributes, this approach could perform poorly in yielding useful generated synthetic data. Second, many existing regression procedures with differential privacy guarantees only generate privacy-preserving point estimates for the regression parameters or point-predicted values for the outcome without uncertainty quantification. For valid inference based on differentially private data, each privacy-preserving regression will output not just the parameter estimates but also information about the variability about the estimates.

As is the case for the multiple synthesis framework without differential privacy, selection of an appropriate synthesis model is critical; model misspecification would generate biased samples. If prior knowledge exists to help in selecting a suitable model, then no privacy budget would be spent on picking a model. If the model selection procedure uses the original data from

Copyright National Academy of Sciences. All rights reserved.

which synthetic data will be generated, then the data curator will have to allocate a certain portion of the total privacy budget to the model selection procedure.

Very recently, motivated by SIPP, researchers have started to investigate the specific problem of generating differentially private longitudinal data in a way that preserves temporal statistics, such as counts over fixed time windows and cumulative trends over time (Bun et al., 2023). However, this initial work only considers low-dimensional data, where each individual has a small number of binary attributes in each time unit.

Statistical Validity

Multiple synthetic sets of identical structure are often released as a way to propagate the uncertainty arising from the synthesis process. Methods have been developed to combine the results from multiple synthetic datasets to yield valid statistical inferences. For multiple synthetic datasets via the population synthesis, Raghunathan et. al. (2003) propose a mathematical formula to combine the estimated variance of the estimated parameter of interest within each synthetic dataset (within-set variance) and the variance of the estimated parameter across the multiple synthetic set (between-set variance) to yield a total variance estimate, based on which statistical inference of the parameter is obtained. For multiple partially synthetic datasets, Reiter (2003) proposes a different formula to obtain the total variance estimate given the within-set variance and between-set variance estimates from multiple sets.

As is the case for the data synthesis without differential privacy, when analyzing and obtaining valid statistical inferences from differentially private synthetic data, it is important to consider the uncertainty arising through the differentially private data synthesis procedure. Compared to data synthesis without differential privacy, the uncertainty comes from one additional source, namely the noise infusion mechanism employed to achieve differential privacy guarantees on top of the synthesis uncertainty and sampling error of the synthesis data. There are a couple of approaches to account for the uncertainty in the literature.

First, one may incorporate the differential privacy sanitization and data synthesis process directly when modelling and analyzing the synthetic data (Charest, 2011; Karwa et al., 2017; Nixon et al., 2022; Seeman et al., 2020). This approach can be demanding for data users either analytically or computationally. The second approach is to release multiple synthetic datasets,

Copyright National Academy of Sciences. All rights reserved.

similar to the multiple synthesis approach without differential privacy. Liu (2022) provides several schemes for multiple synthesis with differential privacy and an inferential rule for combining inference across multiple sets, which turns out to be the same rule used by Reiter (2003) for partial synthesis with differential privacy. Due to the extra source of variability from the usage of randomized mechanisms for achieving differential privacy guarantees, inferences from the differentially privately synthesized data are generally less precise than those from multiple synthesis approaches with differential privacy—a price paid for formal privacy guarantees. If the multiple synthesis approach is adopted, it is important to be careful regarding the choice of the number of datasets to synthesize and release in the differential privacy setting.

In multiple synthesis with differential privacy, the number of synthetic datasets is mainly driven by computational time and storage considerations, and a small number of datasets (e.g., 5 to 10) would be preferred in practice, as long as the number is large enough to capture the between-set variance and deliver valid inference. In contrast, in the differential privacy setting, the decision on the number of synthetic datasets is driven more by the utility of synthetic data at a prespecified privacy budget. Due to privacy loss composition across multiple releases, each synthetic set receives only a portion of the total budget. If the number of sets is too small, the released information may not be sufficient to capture the between-set variability; but a too-large number risks spreading the budget too thin across the sets, and each synthetic set would be so heavily perturbed that aggregating information across the multiple synthetic sets would not remedy the information loss. Tran et al. (2023) perform an empirical study on the budget allocation on their proposed differentially private synthesizer for the synthetic longitudinal business databases. The U.S. Census has done such privacy-loss and accuracy evaluations for the budget determination with their TopDown algorithm (Abowd et al. 2022).

STATISTICAL EVALUATION OF DATA UTILITY

A key consideration in adopting any data synthesis approach is the statistical validity of inferences constructed from synthetic data. This validity can be assessed by applying a set of specific analyses to a list of statistics, such as point estimates, interval estimates, and statistical-significance-of-hypothesis testing results, and for general utility without focusing on any specific analysis.

Copyright National Academy of Sciences. All rights reserved.

A straightforward metric for evaluating the similarity of a statistic calculated from the synthetic data and one derived from the original data is to calculate the absolute deviation, squared deviation, or maximum deviation of the former from the latter. This type of metric is often not inference-based, and it does not take into account the uncertainty around the statistic, but rather it focuses on how much the synthesis process deviates the statistic from its original value.

Another option is to use a propensity score analysis, in which the original and synthetic data are appended and compared, treating the original as "control" observations and the synthetic as "treated" observations. This approach then checks for balance in the covariate distributions across the two datasets. The empirical distributions of the propensity scores can be compared for the original and the synthetic records, separately. A similarity metric between the original and synthetic data can also be computed (Arnold & Neunhoeffer, 2021; Bowen & Liu, 2020; Sakshaug & Raghunathan, 2010; Snoke et al., 2018; Woo et. al 2009).

To assess the validity of statistical inferences, one can develop a large set of candidate analyses from the published literature or as suggested by a panel of SIPP users, and then compare the statistical significance of hypothesis testing results and interval estimates obtained from synthetic and original datasets. Karr et al. (2006) propose the confidence-interval overlap metric to compare interval estimates from the original and synthetic datasets. A value of 1 in this metric corresponds to a perfect match between the two confidence intervals (based on the synthetic and original data); a value of 0 means there is no overlap between the two confidence intervals, regardless of how far these two confidence intervals are separated. When one confidence interval is completely contained within the other, the value of the confidence-interval overlap is greater than 0.5.

Eugenio and Liu (2021) propose the "SSS" assessment. The first S refers to the *sign* of the point estimate, and the second and third S's refer to the *statistical significance* of the estimate. Based on the consistency in the sign and statistical significance for the estimates from the original and synthetic data, seven categories are formed, ranging from "the best," meaning matching in both sign and statistical significance, to "the worst," meaning matching in statistical significance but not in sign. Between the two extremes, there are 5 categories, repressing different types of mismatching on sign or statistical significance. Stanley and Totty (2021) also compare the statistical significance results of estimates based on original and synthetic data.

Copyright National Academy of Sciences. All rights reserved.

Bowen and Snoke (2021) conducted an in-depth analysis on the statistical utility of the generated synthetic data of differential privacy guarantees in the Differential Privacy Synthetic Data Challenge sponsored by National Institute of Standards and Technology Public Safety Communications Research (NIST PSCR) in 2019. There were three matches on three different real datasets in the competition in which different teams competed against one another.[56] Synthetic data were generated for different levels. The conclusion from this competition was that differentially private data synthesis algorithms that used preprocessing and budget allocation based on public information and subject matter utility criteria performed better. The overall best performer used a distribution-free approach, synthesizing data from a predetermined set of low-dimensional histograms/contingency tables (McKenna et al., 2021). Though this was a data challenge event, all three matches used real data, so the competition results and the subsequent analysis of different approaches are insightful in terms of the feasibility of differential privacy synthetic data generation in practice.

Validation and Verification Systems

The validity of the inferences rests on the synthesis model, and therefore a robust validation and verification system is important for assessing the quality of the inferences obtained from the synthetic data. Toward that end, an interactive and web-based system can be developed so that data users can query for measures to check the quality of their analysis results and to learn whether the conclusions suggested by the results are correct or not with a certain level of confidence (Vilhuber, 2019).

Figure 6-4 illustrates how a validation server works. In brief, data users perform analysis on synthetic data and obtain some results. Before publishing the results, they want to validate the quality of the results and gain confidence in their conclusions, so they send their code to the validation server, which runs the code on the confidential data. Releasing quality measures to queries submitted by data users to a validation system incurs disclosure risks due to the fact that information based on the confidential data is being shared with these users. To resolve this, the results sent back to the users can be made private through some privacy-protection mechanisms,

[56]The training datasets are large in size (236K, 314K, and 210K to 662K) with a mixture of 32, 32, and 98 categorical and numerical variables in each.

Copyright National Academy of Sciences. All rights reserved.

such as ones with differential privacy guarantees. After the users receive the privacy-protected results from the server, they can compare them with their own results based on the synthetic data. They can then make a decision about publishing either their synthetic data results or privacy-protected results from the server.

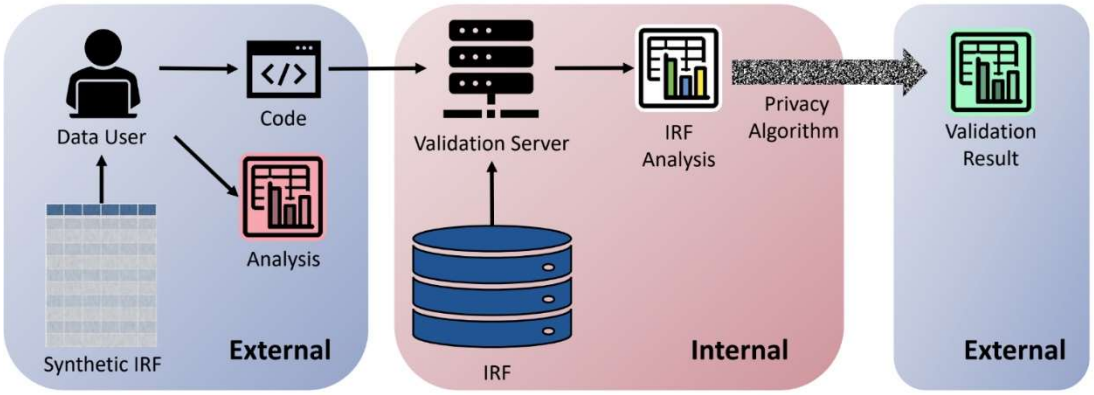


Figure 6-4 How a validation server works.

NOTE: IRF stands for "internal reference file"; *internal* refers to the agency that collected and provided the data; *external* refers to data users.

SOURCE: Gary Benedetto and Rolando Rodriquez, presentation to the panel, June 30, 2022.

The Urban Institute has developed a prototype of an automated privacy-preserving validation server that allows users to submit their code and run it on the confidential data, and which then returns privacy-protected noisy results with differential privacy guarantees (Taylor et al., 2021). The motivation behind the project is to speed up the validation process by automatically reviewing queries and generating privacy-protected results that are to be shared with users. Compared to manual clearance and manual review, this automated system would save time, manpower, and resources. Since the server is a prototype and "built to the minimum standards for testing," it currently runs only tabulation queries on the IRS Statistics of Income Public Use File.

Unfortunately, while the Urban Institute project may be a partial prototype, it is not yet completed and the SIPP data are considerably more complex. It may not be possible to develop a validation and verification system that accommodates every possible analysis of SIPP data, particularly considering the complexity of both the data and the data analyses that are commonly

Copyright National Academy of Sciences. All rights reserved.

pursued. Such a system is desirable from the user perspective and for limiting the burden on the Census Bureau, but developing a complete system is at best a long-term project. Creating SODA with easy access is imperative when a user who is unable to use the verification and validation system can avail him- or herself of SODA to perform the analysis and submit both the results for possible validation and verification.

In addition to validation servers, it is also important to provide users with clarity regarding the models, algorithms, or techniques employed for generating synthetic data so that users have some knowledge in advance about which relationships are likely to have been preserved and which are not. Extensive accuracy testing of the type described in the previous section will, ideally, be conducted by the Census Bureau prior to its release of a synthetic dataset to the general user community. This will save both Census and the user unnecessary effort and resources devoted to relationships that may not be preserved. The user community will want to be assured that particular relationships are preserved and to be able to trust that assurance.

Besides validation servers, there are also *verification* servers, which are related to but different from validation servers. Reiter (2023) defines a verification server as follows. Data users submit a query to the verification server on some statistical analysis. The server performs the analysis on both the confidential and synthetic data and the server calculates "analysis-specific measures of the fidelity of one to the other" (e.g., the overlap of the 95% confidence intervals for a parameter when computed on these two data sources [Karr et al., 2006]). The server then returns the value of the verification measure to the analyst. The analyst can decide, based on the feedback, whether the synthetic data have high utility for the analysis. Barrientos et al. (2018) apply the server integrated approach of synthetic data generation plus verification to longitudinal data on employees of the U.S. federal government, and are able to show that the integrated system performs as intended, allowing users to explore the synthetic data and verify their findings by querying the verification server.

In summary, with a validation server the agency runs the user's statistical analysis on the confidential data and returns to the user a privacy-protected version of the results. With a verification server, the agency runs a user's statistical analysis on both the synthetic and confidential data and returns values on verification measures, which may also be made privacy-preserving if the measures leaks privacy.

Copyright National Academy of Sciences. All rights reserved.

SIPP SYNTHETIC BETA (SIPP SSB)

The SIPP is one of the few major federal surveys that has attempted to produce a synthetic dataset for privacy protection. The creation of the SIPP Synthetic Beta (SSB) data file was not motivated by general privacy concerns about the public-use SIPP data files, but rather by the desire to add administrative data on earnings and Social Security benefits to the SIPP survey files. The addition of those data aimed to address well-known problems of misreporting of those variables in household surveys, including in the SIPP. However, administrative data on earnings are highly sensitive, and both legal restrictions and interagency agreements prevent the release of SSA and IRS data, so the Census Bureau chose to follow the synthetic data approach to make that information available to SIPP users (the data are also available in FSRDCs).[57]

There were two steps in the creation of the SSB. The first step was to merge administrative data on earnings and benefits from the Internal Revenue Service (IRS) and from the Social Security Administration (SSA) onto the internal SIPP files at the Census Bureau, using a protected identification key (PIK) to match the administrative records to the survey records. This file was dubbed the Gold Standard File. Second, using a sequential regression multiple synthesis approach (like the one discussed later in this chapter), the joint distribution of the administrative data and a relatively small number of SIPP survey variables (a little less than 100, out of the approximately 2,900 on the SIPP public-use file) were modeled. Random draws from the posterior of that distribution were made to create implicates for the user to analyze. Together with the administrative data, the SSB file has approximately 140 variables. The SSB file is fully synthetic in the sense that all variables in it are synthesized by the model.[58] The latest version of the SSB has survey information from nine SIPP panels, covering a period from 1994 to 2008 and containing the additional variables for administrative data on earnings and benefits, namely historical earnings that go back to the 1950s. It should be emphasized that the administrative data on earnings contain the entire lifetime history of earnings for any individual who participated in any of the SIPP panels, so those variables cover a far longer period than the survey data themselves.

[57]See Chapter 2 for additional discussion of the SSB.
[58]This means that the SSB does not have the potential problem of uncongeniality noted later in this chapter. There are no variables in the SSB file that are not in the data-generating model.

Copyright National Academy of Sciences. All rights reserved.

One of the important issues in any synthetic dataset, as we will discuss further in this chapter, is whether all relationships among the synthesized variables in the original file are preserved in the synthetic file. Preserving those relationships is likely to be difficult for a dataset with 141 variables, and especially so considering that those 141 variables are observed multiple times for each individual, because the SSB, like the SIPP itself, is a panel. For *all* relationships to be preserved, the joint distribution of the 141 variables at multiple points in time must be preserved, not just their cross-sectional distribution at each single point in time. Again, from the perspective of a user, if not all relationships have been preserved it is important to know which ones have and which have not.

However, while the general structure of the modelling effort has been published by the Census Bureau (Abowd et al., 2006; Benedetto et al., 2013; Benedetto et al., 2018), the details of the model—what variables are used and how they are used—have not been made available to users for confidentiality reasons. This makes it difficult for users to know which relationships in the SSB are likely to be accurate reflections of the Gold Standard File and which are not. The Census Bureau allows users—indeed, strongly encourages them—to submit their programs to the agency to run on the underlying Gold Standard File data, with the output returned to the user to check against the results they obtained with the SSB. This is an example of a validation server, as discussed above, and allows the user to ensure the accuracy of their results. However, as noted in Chapter 2, the statistics on requests for validation appear to suggest that some users do not submit their programs for verification; it is not required. Also as described in Chapter 2, to date there have only been just over 200 users of the SSB and, in fact, it has temporarily ceased operation as of this writing. Given the laborious involvement of the Census Bureau in running user programs for every validation request, it is unlikely that far greater numbers of users could be accommodated by Census Bureau staff to conduct more requests with their current staffing.

To assist users to know in advance which relationships have been preserved in the SSB and which have not, analysts at the Census Bureau have also conducted analyses to determine the accuracy of at least a few use types. Stanley and Totty (2021, 2023) conducted a number of comparisons of the SSB to the Gold Standard File, concentrating on the earnings data since they were the main variables motivating the creation of the SSB. Their study showed that the shapes of the earnings distributions in the two datasets were roughly the same, but not exactly, and there were some differences in the tails of the distributions. Median earnings were quite close in the

Copyright National Academy of Sciences. All rights reserved.

two datasets, but mean earnings were somewhat different. The correlation of earnings with demographic variables were roughly the same in the two datasets and tended to follow the same trends over time. However, some estimation models using quarter of birth differed in the datasets, and differences between the SSB and Gold Standard File were found when state-level data on the minimum wage (which are external to the SSB) were merged onto the SSB and used for a common statistical model used for the SIPP by researchers. The authors attributed this finding to the synthetic data model, which they said did not capture such relationships.

A second study, by Carr et al. (2023), compared the SSB to the Gold Standard File for an analysis of what is called earnings volatility, which measures how much individual earnings are or are not stable from one year to the next. The authors found that the levels of volatility in the SSB were quite different than those in the Gold Standard File for many (but not all) measures, and also found that both the cross-sectional variances and the covariances between earnings in different years were different. However, many of the relationships between demographic variables were similar, and the patterns over time were similar. Coupled with the Stanley-Totty studies just referenced, it appears possible that the SSB is not correctly capturing the relationship between an individual's earnings in different time periods of the SIPP panel (sometimes called "earnings dynamics") and, more generally, is not capturing the evolution of an individual's earnings over time.

Expanding on this experience, the Census Bureau can explore more use of synthetic data to release many variables that may be deemed disclosive. While the purpose of the SSB was to supplement SIPP data with external data, rather than to protect the SIPP survey data, experience with the SSB is instructive for how synthetic data might be approached. Several approaches may be considered. We outline possible strategies for SIPP data next.

CHALLENGES IN SIPP DATA SYNTHESIS AND UTILITY EVALUATION

As introduced in Chapter 2, SIPP is longitudinal (responses are taken from the same individual at multiple time point in the survey), high-dimensional (it reflects thousands of responses collected from the same individual), and multi-modal (containing different data types – numerical, categorical, relational). In addition, the SIPP data are collected using a complex survey design, with different sampling weights for different individuals and separate weights for different types of analyses; the final panel weights for SIPP 2020 range from 387 to 39,951 (see

Copyright National Academy of Sciences. All rights reserved.

Chapter 2). On top of the inherent complexity of the data, the SIPP data are also subject to high nonresponse rates on various variables, leading to a large portion of missing values in the collected data.

While there is a rich body of literature on data synthesis techniques, both with and without differential privacy, and some of these techniques which have been applied to releases of real data, many of the techniques deal with only one or a couple of the characteristics of the SIPP and none deal with all of them. In addition, many techniques, including the recent generative models rooted in deep neural networks, are developed with an explicit or implicit assumption of simple random sampling, do not have missing data, or have difficulty scaling up to the case when there are thousands of attributes in a dataset. If differential privacy is factored in, the usefulness of the synthetic data could further degrade.

Given the variety and the large number of response variables in the SIPP, it would be a daunting task to synthesize all the variables (quasi-identifiers and responses) in the data. If some of the collected response variables are unlikely to be released for either research or public use, they can be kept in-house but be conditioned on synthesis models for generating synthetic data for other variables. For the rest of the variables, it is important to consider the trade-off between disclosure risk and the utility of the synthesized data in each case.

As suggested above, uncongeniality between the synthesis model and the analyst's model is a common challenge in partial data synthesis approaches. If the synthesis model is not expansive enough, inferences based on the synthetic data for the analyst's model will be biased. For a partially synthetic SIPP file with synthesized sensitive variables but a large number of unsynthesized quasi-identifiers, depending on which attributes are released and what the analyst's models are, the uncongeniality issue can be serious. On the other hand, a synthesis model being more expansive increases the likelihood that the model will overfit the original data, which may be associated with higher disclosure risk. Finding the right balance between data utility and disclosure risk limitations will be important.

Finally, the variety of the types of analysis that can be performed on the rich SIPP data also makes it difficult to assess the utility of synthetic data. The particular set of analyses selected to compare synthetic data with the original SIPP will be important if one is to have confidence in the reliability and usability of the synthetic data. We suggest that utility be evaluated in light of the three unique contributions of SIPP data in comparison with other

Copyright National Academy of Sciences. All rights reserved.

datasets, as described in Chapter 2: its large number of variables describing characteristics of U.S. individuals and households, its high level of detail on social program participation, and its longitudinal nature, which allows it to follow individuals and households over time and chart the evolution of their socioeconomic status.

There are several alternative approaches to synthetic data files that the Census Bureau could consider. While a fully synthetic longitudinal SIPP would be difficult, as we have discussed above, a partially synthetic file that retains a large number of unsynthesized variables is worthy of consideration. The usefulness of such a file would depend on how many unsynthesized variables can be put on a public-use file, and that question can only be answered by running more re-identification analyses, as we recommended in earlier chapters. In addition, disclosure risk would have to be carefully assessed in the release of any such file. An alternative approach is to release synthetic data files separately from the public-use file, containing sensitive variables. This is the approach currently taken by the SSB. However, those synthetic files would necessarily also need to retain a number of the quasi-identifiers in the public-use file, although they could be synthesized (as they are, again, in the SSB).

It is also possible that the Census Bureau could construct and release multiple synthetic files, each containing a different set of synthesized synthetic variables. However, because these would each have to have a set of quasi-identifiers that would likely overlap one another as well as overlapping the public-use file, it would be necessary to assess the disclosure risk from an intruder combining all the synthetic files with the public-use file, which would be a challenging assessment to carry out.

Additionally, the use of a synthetic data approach may also be used as a viable method for providing data from outside of SIPP, similar to what was accomplished with the SSB.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 6-1: Data synthesis is a well-established framework for disclosure limitation, and many techniques have been developed for both fully and partially synthetic data generation, including some that use formal privacy guarantees. Data synthesis may be a useful approach for protecting data in a new tier of access.

Copyright National Academy of Sciences. All rights reserved.

Conclusion 6-2: The SIPP Synthetic Beta effort may be regarded as a pilot effort that has tested the reliability and usability of a small synthetic dataset drawn from SIPP. Its success to date has been mixed. Its use has been limited, and it has also demonstrated the need for an automated verification and validation server. The SIPP Synthetic Beta placed a considerable burden on the Census Bureau to manually verify results, and not all synthetic results were confirmed.

Conclusion 6-3: State-of-the-art machine learning frameworks, when combined with the increasing amounts of public data and information that are becoming available, are likely to increase the privacy risks associated with synthetic data. Traditional synthetic data generation may no longer be sufficient for privacy protection. Data synthesis with formal privacy guarantees can be more effective in frustrating privacy attacks.

Conclusion 6-4: Differentially private data synthesis provides an approach to generating synthetic data with formal privacy guarantees following a prespecified privacy budget. Compared to traditional data synthesis approaches, additional calibrated noise is introduced to meet the prespecified differential privacy requirement.

Conclusion 6-5: Inferences based on differentially private synthetic data are subject to all sources of variability, including the sampling error of the original data, uncertainty due to synthesis, and noise introduced by randomized algorithms for achieving differential privacy or other privacy criteria.

Conclusion 6-6: When releasing synthetic data, it is important to evaluate the similarity of inferences using general purpose measures to assess data utility and usability. Assessment of the validity of inferences based on synthetic data is multifaceted and requires comparison of point and interval estimates and hypothesis testing results, across a wide range of use cases.

Conclusion 6-7: It is important that the way synthetic datasets are constructed be transparent to the user and that the user be enabled to understand the relationships that are preserved and the relationships that are not preserved in order to have trust in the data. The user community will benefit by being informed in this way prior to the release of the synthetic dataset. Validation and verification servers can also be used to help build

Copyright National Academy of Sciences. All rights reserved.

trust by allowing the users themselves to determine which relationships are preserved and which are not.

Conclusion 6-8: Automated verification and validation servers would provide synthetic data users with a platform to check the quality of their analysis results based on synthetic data and to have confidence in their findings, while saving resources as compared to the use of verification and validation servers that are manually maintained.

Conclusion 6-9: There is currently no off-the-shelf approach for generating reliable synthetic data for a survey as complex as SIPP, with or without formal privacy metrics.

Recommendation 6-1: The Census Bureau should support research that develops new approaches and techniques for generating synthetic data and assessing data utility and disclosure risk, with or without differential privacy or other privacy guarantees, and should fund work that creates and updates tools and systems that evaluate synthetic data, disclosure risk, and usability and collects data users' feedback.

Recommendation 6-2: The Census Bureau should develop a systematic protocol and a set of metrics for assessing and measuring the validity of inferences constructed from synthetic data.

Recommendation 6-3: To support the use of synthetic data, the Census Bureau should develop a routine automated system to perform user-submitted analyses on the internal file and send the results back to the user after they have been treated for disclosure avoidance.

Recommendation 6-4. The Census Bureau should consider creating partially synthetic datasets to allow analyses of sensitive or disclosive SIPP survey data variables that are not on or may need to be dropped from the public-use file. These synthetic variables might be added to the public-use file in place of the original variables.

Copyright National Academy of Sciences. All rights reserved.

7
Disclosure Limitation Approaches: Flexible Table Generator and Remote Analysis Platforms

INTRODUCTION

Facing increasing demands to allow public-use files that have higher levels of granularity, as well as demands to expand the user base, many statistical agencies are exploring the use of remote flexible table generators and potential extensions to allow for remote analyses through a direct web link. Given the release of public-use file data, the aim is to provide internet platforms where more restricted data may be safely used. Table generators differ from the secure online data access discussed in Chapter 5 by not providing direct access to the microdata or allowing fully customized interaction with the data; instead they provide a tool for generating tables or some types of analyses (with constraints). Based on the literature review provided in Chapter 9, current SIPP users tend to perform complex analyses that could not readily be supported by a table generator at this time, though the development of an automated verification and validation server (per Recommendation 6-3) could be used to develop a tool for remote analysis of this type.

However, the complex structure of the SIPP data is itself a barrier to SIPP data use, and a simplified table generator might broaden the use of SIPP to a larger audience. For example, policymakers may want simple descriptive statistics of program participants, or researchers may want an initial look at the data before pursuing more complex analyses through secure online data access (SODA) or through a Federal Statistical Research Data Center (FSRDC). To begin, an initial product can be created as a simple tool with a simplistic interface, with a specific purpose and clarity about the specific value of that tool. Keeping track of users and the variables used in queries will help the Census Bureau gain an understanding of key uses and help it in guiding future development. If value is seen in offering the simple tool, it is possible to expand from there by surveying users about their interests. By starting small, being clear about the tool's

Copyright National Academy of Sciences. All rights reserved.

purpose, value, and what it will contain, an agency can "test the waters" on the usefulness of the tool.

Internet platforms that allow users to retrieve automatic outputs require applying perturbation methods in the outputs prior to their release. The panel envisions that the underlying microdata for this tool would contain the variables from the public-use file and a select set of variables that are not released in the public-use file. In contrast to releasing synthetic versions of these additional variables, here the perturbation is only applied on the requested outputs generated by the users, hence there is more utility in the data. The development of a remote table generator and analysis platform would serve the wider research and policy analyst community and could help provide evidence and inform subsequent grant or FSRDC applications using SIPP.

There are some working examples of remote flexible table generators, particularly for census outputs that can be used by the public or require a light registration process. Two examples are the European Union Census Hub and the Australian Bureau of Statistics Tablebuilder.[59] With more research and development, these flexible table generators could go a step further and provide statistical analyses such as exploratory analyses and regression models (e.g., the Australian Tablebuilder allows the calculation of means, medians, and confidence intervals). Some details of the confidentiality protection in the Australian Tablebuilder are discussed in Thompson, Broadfoot, and Elazar (2013). More research would be beneficial on handling complex statistical modelling, such as hierarchical and multilevel modelling, which can accommodate the analyses of SIPP longitudinal data.

TABLE GENERATOR PLATFORM

At its core, the flexible table generator with remote analysis extensions uses large multi-way hypercubes (crosstabulation) as input data, thus ensuring minimum building blocks for the tables and models, below which the data cannot be disaggregated. In particular, the generation of two outputs that differ by only one individual would break any potential confidentiality protection. The entry into the platform is through a menu-driven user interface and bespoke

[59]For the European Union hub see https://ec.europa.eu/CensusHub2. For the Australian Bureau of Statistics Tablebuilder see https://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder

Copyright National Academy of Sciences. All rights reserved.

software platforms where variables and their (coarsened) categories are predetermined and can be selected as a variable in the table or model. The software includes general rules-of-thumb regarding minimum size thresholds, sparsity thresholds, rows/columns that have only a few categories while the other categories are zero (attribute disclosures), number of dimensions allowed in a table or model, and more. When requesting the table or model, if these rules-of-thumb are not met, the user will receive a message that their request is denied and will be redirected to selecting a reduced number of categories of a variable or alternatively to include a larger sample size by expanding the subgroup selection. There are particular rules-of-thumb for remote analysis servers as well, for example that no single data point can be released. Therefore, all requested scatter plots are produced as sequential box plots, and outliers in the whiskers of box plots are suppressed. In addition, the minimum and maximum values would not be disseminated, and the underlying models may have robustness features added to them to down-weight outliers. Quantiles can be released but with appropriate perturbation added as described below.

The use of differential privacy in flexible table generators is well documented; see for example Rinott et al. (2018), where the exponential mechanism is utilized based on a Discrete Laplace Distribution. Since the SIPP is a probability sample and has survey weights, we can adjust for the weighted counts in the tables as shown in Shlomo et al. (2019). See Appendix C for more details.

Though the emphasis here is on a flexible table generator, the same framework may be extended. For continuous variables, such as sums, averages, quantiles, and correlations, we can use the same concept of the microdata keys to obtain the same perturbations. For more advanced modelling in the remote analysis server, one can add noise to the underlying estimating equations. However, more research is required on how the table generator can be extended for other analyses as well as on actual implementation in a differentially private setting. Implementation may require a registration process before policy analysts could access more restricted data compared to what would be available to the wider public on the internet. Future development of this strategy would require ongoing collaborations among the academic community and government researchers. SIPP users use sophisticated statistical models, and a table generator may not fulfill their needs, but such a system would be useful to perform some

Copyright National Academy of Sciences. All rights reserved.

exploratory or preliminary analyses before embarking on requests to access, for example through SODA or an FSRDC.

It could take years to fully develop these platforms, since it will take time for interface development and other IT solutions. The release of initial products with milestones outlined would help to provide access to restricted data through a work-in-progress tool. As an example, progress in the development of the flexible table generator and, more generally, of the remote analysis platform could follow a sequence like this:

1. Release of flexible table generator as a simple tool and initial product to gauge value and interest, using the public-use file data, with no noise introduced.
2. Survey of tool users to determine needs, audience for tool, common queries, comments.
3. Decisions on future development and next steps, for example in the following sequence:
    a. Flexible table generator, using the restricted data under the differential privacy framework.
    b. Remote analysis platform, using the public-use file data, with no noise introduced.
    c. Remote analysis platform, using the restricted data under the differential privacy framework.

CONCLUSION AND RECOMMENDATION

Recommendation 7-1: The Census Bureau should assess the demand for an initial flexible table generator as a simple tool, with a specific purpose, that is designed to gauge value and provide direction for further development, and proceed with the development of one if there appears to be sufficient demand.

Copyright National Academy of Sciences. All rights reserved.

8
Disclosure Limitation Approaches: Geography Variables

As discussed in earlier chapters, risks of disclosure increase when respondents can be broken down into very small subgroups. This can happen when one combines many different measures together but especially is a risk when a single measure is sufficient to create very small subgroups. Measures of geography are a notable source of risk because they are highly identifiable and they narrow down the sample and the overall population to much smaller subgroups. For example, through a re-identification study the Census Bureau identified disclosure risks from the geography measures used in the Current Population Survey, requiring changes to the measures (Benedetto et al., 2022).

Recognizing these risks, the Census Bureau limits the amount of geographic data available through SIPP (more strongly than it limits these data in the Current Population Survey). As is standard for confidential surveys, home addresses are automatically suppressed, as are the names of counties and specific metropolitan areas. In fact, while SIPP includes a measure indicating whether the respondent is in a metropolitan area, that measure is used only when a state has multiple metropolitan areas. SIPP does include state identifiers for all the states. However, as was discussed in Chapter 3, the inclusion of state identifiers can be highly disclosive.

The challenge from a data usability perspective is that geography variables can be highly valuable for research. For example, variations in state or local policies or programs create a type of natural experiment that researchers can use to estimate the impacts of those policies or programs. There also are naturally occurring differences that appear across states or regions that can be meaningful to researchers, such as the common differentiation in political dialogue between “red” states and “blue” states.

Different types of solutions can be developed depending on the specific needs of a research project. For example, some data users may only need to distinguish one state from another without actually identifying the specific states. Other researchers may be interested in

Copyright National Academy of Sciences. All rights reserved.

identifying geographic areas, but again without the need to identify specific states. Still other data users will need the ability to identify specific states. Thus, the panel considered three use-case scenarios in developing potential strategies for providing geographical variables and limiting disclosure.

USE 1: IDENTIFYING SPECIFIC GEOGRAPHIES IS SOMETIMES UNNECESSARY

Although geography is a natural variable used for producing small-area estimates, it is not strictly needed to produce the national-level regression analyses that are often sought by data users, in which the inferential task does not involve the geography itself. Instead, for these analyses it is sufficient to release public-use microdata with a "mask-id" (i.e., a "group-level" identifier) rather than a state-level identifier. Releasing a mask-id would allow researchers to incorporate this variable for analysis as either a fixed or random effect in a regression model. In other words, the inclusion of a mask-id variable as either a fixed or random effect would provide the same benefit as including a state-level (or other geographic-level) identifier but with the benefit of not necessarily being disclosive. It is envisioned that proceeding in this way will allow data users to produce national-level estimates for a subset of analyses associated with specific subpopulations.

Still, some states may be identifiable even with a mask-id. For example, the 10 largest states can each be identified by assuming that a number of respondents (or the sum of the sampling weights) in SIPP have the same rank ordering as the state population sizes. If the state-level id is determined to be a disclosure risk, then it may not be viable to simply associate a mask-id with each respondent. In this case, another potential path forward would be to consider partially synthetic data, where the state-level id (or mask-id) is synthesized for each record in the public-use file. Although there has been some research on generating synthetic geographies (Quick et al., 2015; Wang & Reiter, 2012), this is typically conducted to produce a synthetic location for each respondent. Instead, generating a synthetic state-level id (or mask-id) for each respondent would proceed from a latent class model and require additional research. In principle, this type of partial synthesis could also proceed using formal privacy methods. See Chapter 6 and the references therein for additional discussion on synthetic data methodology.

Copyright National Academy of Sciences. All rights reserved.

USE 2: MAKING SUBNATIONAL ESTIMATES

Sample surveys are designed to produce reliable estimates over large geographies and for populations and subpopulations with sufficient sample sizes. Nevertheless, there is an increasing demand for reliable estimates at more granular levels of geography and for smaller subpopulations. To meet this demand, modeling is often utilized as a means of producing estimates of sufficient accuracy that would not be possible based solely on a sample survey.

In many cases, small geographies or subpopulations have insufficient sample sizes (or no sample) to produce reliable estimates. In these cases, model-based approaches allow for the "borrowing of strength" by leveraging different sources of dependence and by incorporating auxiliary information. Therefore, estimates can be achieved for additional domains of interest, with the estimated quantities typically having increased precision. The Census Bureau has a rich history of running various small area estimation (SAE) programs. Arguably the two most notable of these programs are the Small Area Income and Poverty Estimates (SAIPE) and the Small Area Health Insurance Estimates (SAHIE). Both programs use Bayesian model-based methodology to provide estimates of increased precision at under-sampled geographies.

To produce a broad range of state-level estimates for various subpopulations, a robust model-based SAE program is desired. To achieve this goal the Census Bureau can proceed using either an area- or unit-level approach (Parker et al., 2023b; Rao & Molina, 2015). In either case, publishing estimates at a high enough level of aggregation could allow for dissemination without the need for extra noise infusion prior to release (Reiter, 2012) to reduce disclosure risk.

Area-level model-based approaches often proceed using a Bayesian hierarchical framework. A now standard formulation for the Bayesian hierarchical model conditionally expresses the model as having three components: the data model (or likelihood), the process model, and the parameter model (e.g., see Cressie & Wikle, 2015, and the references therein). The advantage to using a Bayesian hierarchical model in this context resides in its ability to accommodate a broad range of modelling tasks and to provide meaningful measures of uncertainty. The well known Fay-Herriot model (Fay & Herriot, 1979) is a special case of the multivariate spatial mixed effects model, where the process model's random effects are independent and identically distributed (see Appendix D for the model formulation). The Fay-

Copyright National Academy of Sciences. All rights reserved.

Herriot model is widely used and forms the basis for the model used in Small Area Income and Poverty Estimates[60] (Bell et al., 2016).

Importantly, in both the Gaussian and non-Gaussian cases, the Bayesian hierarchical model framework is general and provides the needed flexibility for customization to the SIPP. The main considerations in developing area-level models include whether to use multivariate or univariate modelling, to incorporate geographical (spatial) or temporal (longitudinal) dependence, to incorporate auxiliary information (e.g., administrative records and/or data from other surveys), and to deal with potentially non-Gaussian responses. Importantly, in the development of area-level models for specific data products, such as SIPP, many of these issues are not mutually exclusive.

Note that geographical (spatial) dependence may be defined in multiple ways (Nandy et al., 2023). A primary focus in this chapter is to look at states because state are often of interest in policy research (e.g., based on variations in policies across states) and because state is the most detailed clearly identifiable measure of geography incorporated in SIPP. However, other types of geographic dependence are of interest, including at a more detailed level (such as metropolitan areas) and at a more aggregated level (such as geographic regions). The Census Bureau has a definition of geographic region that it often applies,[61] and there often are commonalities within regions that help to distinguish them from one another (e.g., level of education, racial/ethnic composition, and poverty levels). If such measures are used to define neighbor relationships, then one has removed some identifying (state-level) information from the models and provided additional protection from disclosure.

Area-level models usually begin with the response variable consisting of the direct estimate (design-based estimate). Along with this estimate, the statistical agency (e.g., the Census Bureau) typically publishes an estimate of the sampling error variance. That is, in the Gaussian case, the response is given by the direct estimate, with the observation level (data model) variance usually assumed to be known and taken to be the sampling error variance. Consequently, in this setting the proposed model incorporates the survey design by construction.

As an example, consider state-level estimates of wealth. These could be produced using a straightforward Fay-Herriot model, where the response variable is a direct estimate (design-

[60]https://www.census.gov/programs-surveys/saipe.html

[61]https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Copyright National Academy of Sciences. All rights reserved.

based estimate) of state-level median household wealth regressed on the latent "true" state-level median wealth plus an independent error with known sampling error variance. In other words, a noisy version of true state-level median household wealth constitutes the response. At the next level in the model hierarchy, the "true" median household wealth is modelled conditional on state-level demographic variables and other wealth-related covariates. At this stage in the model hierarchy, a state-level random effect could be included to improve the model predictions.

The main goal for SAE within a statistical agency is to improve the precision of the estimates being disseminated and to provide estimates that may not otherwise be possible. As such, it is advantageous to leverage various sources of dependence and to incorporate auxiliary information. One significant source of dependence arises from multivariate relationships in the data. There are several ways to incorporate multivariate dependence depending on the specific problem being considered. For the Gaussian case, one can consider a model where the responses are assumed to be conditionally independent given the covariates (Wang et al., 2012). That is, the covariance matrix of the observations does not contain any sampling error covariance. Instead, the multivariate dependence between responses is incorporated in the process stage of the model hierarchy. When the errors in the process stage only contain multivariate dependence, this model can be viewed as a multivariate Fay-Herriot model (Benavent & Morales, 2016; Porter et al., 2015).

Spatio-temporal Models

Herein, we describe multivariate spatio-temporal models and note that Fay-Herriot models, univariate models and/or spatial-only models can be considered as special cases. In the Gaussian case, one can proceed using a multivariate spatio-temporal mixed-effects model (MSTM) (Bradley et al., 2015). The basic formulation is given in Appendix D. See Bradley et al. (2015) for additional details. Models for non-Gaussian data, such as data from the natural exponential family, can be efficiently accommodated using conjugate distribution theory (Bradley et al., 2020), data augmentation (Polson et al., 2013), and/or specialized software such as RStan, INLA, or NIMBLE, depending on the specific problem (de Valpine et al., 2017; Rue et al., 2009; Stan Development Team, 2023).

Revisiting the example of state-level median wealth income, a spatial or multivariate spatial model could be considered. In the case of a spatial model, the state-level random effect

Copyright National Academy of Sciences. All rights reserved.

can be replaced by a conditional autoregressive (CAR) model or intrinsic conditional autoregressive (ICAR) model (e.g., see Banerjee et al., 2014; Cressie & Wikle, 2015; and the references therein). Similarly, given an outcome correlated with state-level median wealth (also from SIPP), a multivariate or multivariate spatial model could be specified. The difference would reside in the specification of the fixed and random effects at the process level of the model hierarchy (e.g., see Bradley et al., 2015; and Cressie & Wikle, 2015, for additional details).

Again, in the case of a Gaussian response, typically the observation variance is considered fixed and known and taken to be the sampling error variance. However, this need not be the case. Instead, one can simultaneously model the mean and the variance (e.g., see Parker et al., 2023a, and the references therein). In particular, Parker et al. (2023a) provides a computationally efficient example of leveraging spatial correlation and covariate information in the model for the variance. Although there has been some research focused on shrinking both the mean and the variance, this research is relatively sparse by comparison.

When modelling non-Gaussian data using an area-level model, the question arises as to how to effectively account for the sampling design. One reason this problem arises is because the data model (likelihood) does not present a convenient form for incorporating the sampling error variance, in contrast to the Gaussian data model case. For the binomial likelihood, one can utilize the effective sample size and effective probability of success based on the design effect, the ratio of the variance under the survey design and under simple random sampling (Malec et al., 1999). Mohadjer et al. (2012) use a nonlinear linking model for modelling proportions that lead to unmatched sampling and linking models. For the Poisson likelihood, one can utilize the equidispersion property and treat both the sampling error variances and direct estimates as separate sources of data; see Bradley et al. (2016) for further detail.

Modelling at the Unit Level

Another framework for SAE considers modelling at the unit level (e.g., see Parker et al., 2023b and the references therein). These models rely directly on answers provided by individual survey respondents as the dependent variable and can resolve some of the disadvantages associated with area-level models. Specifically, by starting with data at the record level of the survey as the response variable (i.e., person-level or household-level in the case of SIPP), the method can be considered a “bottom-up” approach. In other words, in principle, predictions can

Copyright National Academy of Sciences. All rights reserved.

be made for any level of geography (or subpopulation) and aggregated up to any desired tabulation level, such as county, state, or national. In contrast to area-level modelling, there is no need for so-called benchmarking, where lower-level estimates are forced to aggregate to higher-level estimates as a means of achieving internal consistency among estimates at various tabulation levels (Bell et al., 2013). Additionally, because unit-level models directly leverage the entire dataset, rather than summary-level statistics (e.g., direct estimates), they are potentially more precise than estimates coming from an area-level model (Hidiroglou & You, 2016).

Even though there are several advantages to unit-level modelling, they also present several challenges. First, unit-level models are often non-Gaussian (e.g., binary or categorical) and, compared with area-level models, can be extremely high-dimensional. Taken together, this leads to a computationally intensive prediction problem. Perhaps more crucially, complex sample survey designs such as that in SIPP typically lead to informative samples (i.e., the probability of sample selection is correlated with the response). Importantly, if left unaddressed, the issue of informative sampling can lead to biased estimates.

In cases where the design is noninformative (e.g., a simple random sample), the distribution of the sampled responses is the same as the nonsampled responses after conditioning on covariates. As such, the basic unit-level model introduced by Battese et al. (1998) can be used (see Appendix D). This approach constitutes an unweighted model-based analysis.

Supposing that the survey design is informative, Gelman and Little (1997) describe a general approach to Bayesian unit-level modelling that conditions on the design variables. The idea is that if all the design variables are accounted for in the model, then the conditional distribution of the response, given the covariates, is independent of the probability of selection. Also, by using a Bayesian unit-level approach, the unsampled population can then be generated through the posterior predictive distribution. Consequently, a distribution can be obtained for any finite population quantity and, thus, estimates and appropriate measures of uncertainty can be produced.

In practice, conditioning only on the design variables may be impractical, because not all the design variables may end up being accounted for and the functional relationship is typically unknown. Therefore, other methods to account for informative sampling are desired, leading to models with survey weight adjustments.

Copyright National Academy of Sciences. All rights reserved.

One popular approach for handling informative sampling is the Bayesian pseudo-likelihood (Savitsky & Toth, 2016). This model extends the pseudo-likelihood of Binder (1983) and Skinner (1989) to the Bayesian setting. This extension essentially replaces the likelihood by an exponentially weighted likelihood (see Appendix D).

Cast within a Bayesian hierarchical framework, the pseudo-likelihood-based model results in a pseudo-posterior from which estimated model parameters can be obtained. Then, given the estimated model parameters, it is possible to generate estimates and measures of uncertainty for finite population quantities using post-stratification (Gelman & Little, 1997; Park et al., 2006).

Another popular approach to unit-level modelling proceeds by regressing on the survey weights. For example, Si et al. (2015) introduce a unit-level approach that addresses several issues encountered with unit-level modelling, namely lack of knowledge regarding the covariates, sampling weights, and population sizes associated with the nonsampled units. Using a multinomial distribution, the proposed methodology models the observed post-stratification cells conditional on the population size.

Incorporating Spatial Dependence and Correlated Random Effects

The unit-level models previously discussed offer several avenues for incorporating spatial dependence. The most straightforward approach for incorporating spatial dependence is to incorporate an incidence vector (i.e., a vector that equals 1 if the i_{th} unit is in area i and 0 otherwise) in the random-effect component of the underlying process model (Sun et al., 2022). This vector maps observations to specific geographies, providing an extra source of variation for units within the same geography. The specification above can also account for more complex spatial dependence using spatial basis functions in the random-effect or by using a conditional autoregressive or intrinsic conditional autoregressive model. See Parker et al. (2023b) and the references therein for additional detail. Area-level spatial dependence relies on the notion that units in neighboring geographic areas are correlated, whereas units in distant geographic areas are uncorrelated. For example, Vandendijck et al. (2016) propose an approach similar to Si et al. (2015) that incorporates spatial dependence using an ICAR model.

In the case of a Gaussian response, adding spatially correlated random effects is relatively straightforward. In addition, the resulting formulation is computationally feasible, even in

Copyright National Academy of Sciences. All rights reserved.

moderate to high-dimensional settings. In contrast, for the non-Gaussian case, estimation can be computationally intensive due to non-conjugacy in the model specification. For the Poisson, binomial, and multinomial cases, methods have been proposed to circumvent this issue (e.g., see Parker et al., 2020, 2023a).

Similar to the area-level modelling case, another issue to be considered when building small area statistical models is whether to specify multiple univariate models or, alternatively, to leverage multivariate dependence. In general, characterizing multivariate dependence can be challenging, especially when moving beyond bivariate relationships. Although there has been some research for multivariate modelling at the unit level (e.g., see Parker et al., 2022), model development will directly depend on the survey design associated with SIPP and the specific needs of the data user (i.e., the most frequent use cases). In particular, the survey design associated with SIPP produces several sources of temporal dependence. At the unit level, one way to incorporate temporal dependence is to have time-varying model parameters. Deciding which model parameters are allowed to vary over time would necessitate an exploratory analysis and (or) subject matter expertise.

Leveraging Auxiliary Data and Applying Disclosure Avoidance

In both area- and unit-level modelling, leveraging auxiliary data using data integration methodology can often lead to better estimates (e.g., see Bradley et al., 2016; and Lohr & Raghunathan, 2017). Sources of auxiliary data can include other surveys (e.g., the American Community Survey) or administrative records (Maples, 2017). In the case of integrating multiple surveys, several issues apply. At the area level, it is important to determine whether the tabulations are definitionally consistent and what the coverage of each tabulation is. At the unit level, in addition to definitional consistency, there may be issues associated with record linkage. Also at the unit level, administrative records may be used as a source of auxiliary data, though issues associated with coverage and record linkage may arise (Reiter, 2021).

Adding disclosure avoidance methodology to SAE is an important area of research. One viable path forward, as described previously, is to produce model-based predictions using the direct estimates on a large enough domain. Since this would constitute an indirect estimate that uses the entire dataset, in principle, the risk of disclosure would be significantly reduced relative to the direct estimator. This approach would lack the formal privacy guarantees associated with

Copyright National Academy of Sciences. All rights reserved.

differential privacy (Janicki et al., 2022), similar to many of the legacy disclosure avoidance methods applied to direct estimates such as top coding, swapping, and suppression, among others (see Chapter 4).

Implementing formal privacy methods in the SAE setting raises several research questions. The first question is at what stage the differential privacy noise is to be added. Unlike the case with the 2020 decennial census (Janicki et al., 2023), adding differential privacy noise to the direct estimates and then modelling the protected data results in a response variable with two sources of error (i.e., from differential privacy and sampling), thus making it more difficult to specify the statistical model. Similar issues arise from adding the differential privacy noise to model-based predictions. In addition, implementing differential privacy in either case requires specifying the privacy loss budget in the presence of longitudinal dependence and constitutes an open area of research.

For data coming from a complex survey, a promising direction is to generate differentially private synthetic data or use administrative records to generate a synthetic SIPP/administrative records dataset (e.g., see Chapter 6). Given the synthetic data, small area estimates can be produced. Conveying to data users the methods for accounting for the synthesis uncertainty of the synthetic data would be helpful when constructing small area estimates.

USE 3: IDENTIFYING SPECIFIC STATES OR LOCALITIES

A third situation occurs when there is a need to identify specific states or localities, such as when the data user wishes to measure the impact of a policy or program by making use of differences across states or localities (e.g., Dondero & Altman 2020; Ribar, 2005). In such situations, data users will wish to access the raw data on geography. Either Secure Online Data Access (SODA) or use of a Federal Statistics Research Data Center (FSRDC) will provide another path forward for producing small area estimates. In conjunction with the public-use files, SODA and FSRDCs can be thought of as a tiered system of access, where the specific "tier" depends on the desired analysis.

At a high level, SIPP access could be viewed as follows. Public-use files would contain "protected" data with access to limited variables. SODA would allow researchers access to a less restrictive version of SIPP that does not require any auxiliary data sources. Finally, the FSRDC would allow full access to the data and the possibility of linking additional data sources

Copyright National Academy of Sciences. All rights reserved.

(comingled data) as needed (e.g., linking SIPP with administrative records). Dissemination of small area estimates obtained through SODA or an FSRDC would require some form of disclosure review prior to release. For a comprehensive discussion regarding SODA and FSRDCs, see Chapter 5.

Recommendation 8-1: The Census Bureau should continue to pursue the development of a small area estimation program to meet SIPP users' need for geography-based analysis that preserves confidentiality and limits disclosure risk.

Copyright National Academy of Sciences. All rights reserved.

9

Maintaining Usability While Preserving Confidentiality: Potential Strategies

In chapters 4 to 8 we discussed several alternative methods for increasing privacy protection and reducing disclosure risk. In this chapter we explore how adoption of these methods might affect the usability of SIPP data and how reductions in usability could be mitigated through the provision of multi-tiered modes of access. We also explore more open two-way communications about data privacy and access between SIPP users and the Census Bureau. Finally, we discuss Title 13 and how its interpretation may be modernized to improve data access and usability.

As the Census Bureau evaluates potential privacy solutions and develops new modes of access to SIPP data, it will be important to evaluate those solutions in terms of their potential impact on the three dimensions of usability discussed in Chapter 4: *accuracy, feasibility,* and *accessibility*. Detailed discussions about various privacy protection methods are provided in the other chapters.

This chapter uses the following five potential modes of access as illustrations, arrayed here from the most to the least accessible:

1) Online tabular/analysis builder, a tool for online analysis of SIPP data, particularly for tabular outputs and statistical analyses, with limits on the outputs that can be generated (i.e., no minimums, maximums) and differential privacy protections applied to the results (see Chapter 7);
2) Public-use microdata, a publicly available microdata file that includes a limited number of variables, possibly along with synthesized variables for information that is too disclosive to release (see Chapters 2 and 3);

Copyright National Academy of Sciences. All rights reserved.

3) Synthetic data, a synthesized dataset with a gold standard verification or validation server (see Chapter 6);
4) Secure online data access, a secure virtual environment where approved users can access stand-alone SIPP data and obtain results following data disclosure review or infused with noise and correct inferences, but with lower start-up costs than when working through a Federal Statistical Research Data Center (FSRDC) (see Chapter 5); and
5) Federal Statistical Research Data Centers, the secure labs already in place, at which SIPP and administrative record linkages are available to users holding Special Sworn Status and who have approved proposals. Users may need to work in the lab rather than remotely if they are working with tax or other administrative data.

Now consider each dimension of usability in greater detail with illustrative examples of how they may be assessed across different modes of access.

ACCURACY

As introduced in Chapter 4, the core element of the most traditional, but also the narrowest, definition of usability—often called validity and sometimes called accuracy—builds on basic concepts of statistics. A statistic is any function of a set of observations from a dataset. Some statistics are simple, such as a mean, a variance, a correlation coefficient, or cells in a contingency table. But others are not so simple, such as the coefficients on variables included in a linear regression estimated by ordinary least squares. Yet others may be even more complex, such as the parameters in a generalized or nonlinear model, or the parameters in a sophisticated causal-effects design using propensity scoring, instrumental variable, or other methods. A statistic may be computed from a cross-sectional dataset, like one wave of SIPP, or from multiple waves, as would be the case if the statistic measured income mobility over time, for example.

Two important dimensions of the accuracy of a statistic from a particular dataset are unbiasedness and consistency, on the one hand, and reliability in terms of confidence intervals, on the other. Unbiasedness and consistency are related concepts, one relying on finite samples and one relying on asymptotically large samples. They signal whether, on average or with large datasets, a particular statistic is equal to its true value, either in the finite population or in an

Copyright National Academy of Sciences. All rights reserved.

imagined super-population. We use the term *reliability* to mean whether the uncertainty surrounding a statistic, measuring how far off it is from a true value (usually measured with confidence intervals) is small enough that one can be reasonably sure (at a particular "level" of confidence, say 5 percent or 10 percent) what the true value in the population is or is not. A closely related concept in statistics is what is called the *power* of a statistic, which is the ability of a particular statistic from a particular dataset to "detect" a particular value with a reasonable degree of confidence.

With these concepts, when considering the impact on usability of any data analysis system (DAS) method, the relevant comparison is between the original dataset and the released dataset. Assuming that a statistic from the original dataset can be used to obtain unbiased and consistent estimates of a parameter, the question is whether the released dataset will also allow such estimates to be obtained. In the extreme case in which a variable necessary for the computation of the statistic is not released at all, the answer is clearly no (note that this issue falls under the category of "feasibility" in our classification scheme in Figure 9-1 and will be covered in greater detail in the next section). In other cases, the released data may be so highly distorted from the original that while unbiased and consistent estimates are possible they can only be obtained with difficulty, and only if the user is given the details of the data analysis system method (to be able to compensate for it statistically). For reliability, the right comparison is between confidence intervals and power in the original data versus those intervals and power in the released dataset. A released dataset is effectively unusable if, by comparison with the original dataset, the confidence intervals widen so much and power is so greatly reduced that no reliable inferences can be made.

For any particular statistics obtainable from the original dataset, whether proper inferences can be made and validity can be obtained will depend on the mode of access of the data as well as the information in the released data. We listed five types of access at the beginning of this chapter. It is the combination of a particular model with a particular released dataset that determines whether valid and reliable estimates can be obtained. A good example of such an assessment is the study conducted by Jordan Stanley and Evan Totty (2023), which compared estimates generated from the SIPP Synthetic Beta dataset with the Gold Standard File, (which were discussed both in Chapter 2 and in Chapter 6). The authors found that simple statistics generated from the synthetic data (e.g., means, medians, trends, OLS coefficients)

Copyright National Academy of Sciences. All rights reserved.

closely replicated estimates based on gold-standard data, but analyses that relied on data features that were not explicitly modeled (e.g., quarter of birth in instrumental variable models) or more complex analyses (e.g., two-stage least squares or factor models) performed less well. Similar designs could be employed to compare original SIPP data (i.e., gold standard) with proposed public-use releases or synthetic datasets. This illustrates how usability interacts with the mode of access (which, in this case, is a synthetic dataset).

While the assessment of accuracy through comparisons with a gold standard is relatively straightforward in principle, the difficulty in applying this methodology to a high-dimensional dataset like SIPP is that there are thousands of different statistics that have been calculated and, looking to the future, could be computed for any topic area of interest. Any particular data analysis system method, resulting in a particular released dataset to be used in a particular mode, may have little effect on some uses of statistics produced from SIPP, only a modest effect on other uses, and a significant effect on still others, and may make the use impossible for others. In the sections below, we review the various ways SIPP has been used, aiming to illustrate the wide range of uses to consider when assessing the different dimensions of statistical usability from data analysis system methods.

Once assessments are made of the validity and reliability of SIPP-based estimates for a given data product, it will be important for the Census Bureau to clearly communicate its findings to the SIPP user community. It is possible that estimates based on some variables are more accurate than those based on other variables. For example, for a given synthetic data file, it is possible that estimates of repeated cross-sectional trends in poverty would be highly accurate while estimates of within-person income volatility were less accurate. Users who study income volatility may then especially depend on the verification process that uses the gold standard data, or may seek to access the data in other ways.

FEASIBILITY

Feasibility is a dimension of usability that refers to the degree to which analysts can use the data to investigate the range of scientific questions that the data were designed to answer and, thereby, to obtain valid and reliable estimates, as discussed in the previous section. In the case of SIPP, the key mission is to monitor and study population levels, disparities, trends, and within-person dynamics in public program use, income, poverty, hardship, wealth, and disability, and to

Copyright National Academy of Sciences. All rights reserved.

monitor and study their relationship with public policies. Of key importance is whether users have access to the variables required to conduct such analyses. A related consideration that impacts feasibility is whether users have access to individual data records and computer platforms that permit them to manipulate the data, including functions that enable users to recode, reshape, and merge data files and conduct statistical analyses.

Feasibility is contingent on the specific ways the data are being used. Rather than trying to evaluate feasibility for any possible way SIPP data could be used, it is important to prioritize uses that build on the key strengths of SIPP data, namely their substantive foci and longitudinal design. To that end, a survey of SIPP research literature identified five broad uses of SIPP data:

1) Analyses relying on unique SIPP content (whether for a descriptive or multivariate analysis)
2) Longitudinal analyses (including both repeated cross-sectional and within-panel)
3) Analyses relying on granular data and complex recoded data (necessary for some policy-relevant analyses)
4) Analyses of causal effects of public policies and other analyses that depend on geographic identifiers, and
5) Analyses relying on administrative record linkages.

These five uses are examined in more detail, in sequence, in the remainder of this section.

Next, consider the distribution of SIPP research across use categories. This information is important to consider when deciding which uses to prioritize, especially when it is necessary to make tradeoffs between usability and privacy. For this report, the top-50 most cited articles that used SIPP were examined, as measured through Scopus[62] (N = 46; 4 studies were dropped because they did not actually use SIPP data in their main analysis). Because frequently cited articles tend to be older than less frequently cited articles, we supplemented this list with every SIPP-based analysis published in the last 12 months (N = 22). Each article was classified according to how SIPP data were used. Note that articles could use the data in multiple ways. For

[62]Scopus performs a search on titles and abstracts, and would not detect studies that mention SIPP only in the main text. Google Scholar performs a full-text search, and thus detects additional studies; however, Google Scholar counts multiple versions of the same article separately. Google Scholar also incorporates a wider range of literature, while the Scopus review was limited to peer-reviewed journal articles. Both types of searches may identify publications in which SIPP was mentioned but not used as a data source.

Copyright National Academy of Sciences. All rights reserved.

example, a study could be both longitudinal and rely on unique SIPP content. Studies were coded as "descriptive only" if they did not use multivariate statistical analyses. The results are shown in Figure 9-1.

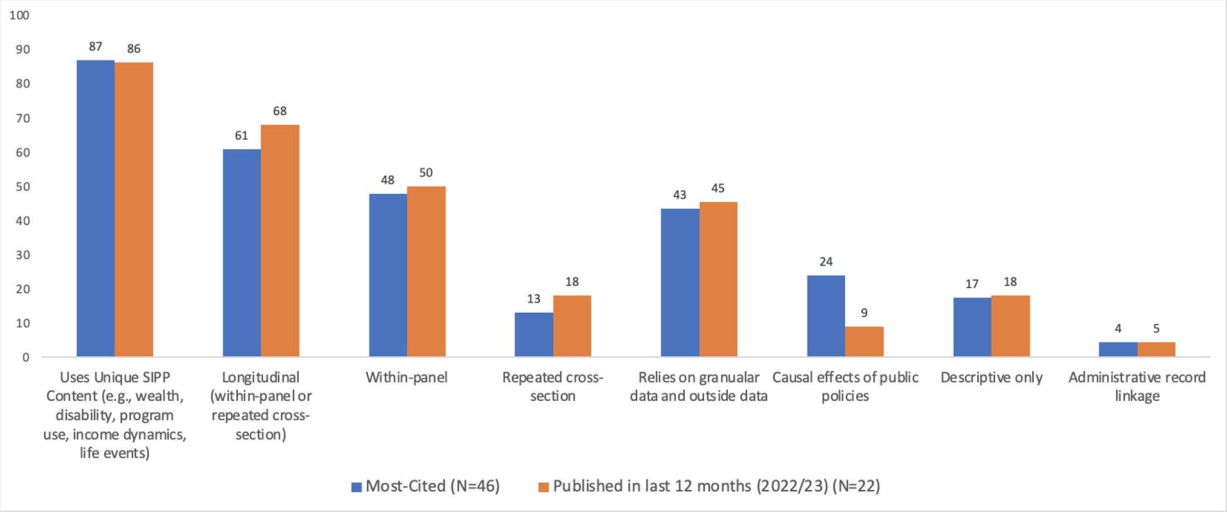


Figure 9-1 Uses of SIPP data in the most cited and recent studies (percentage).
NOTES: Each article can be included in multiple categories. Articles were drawn from a search using Scopus on February 7, 2023.

Use 1: Analyses Relying on Unique SIPP Content

SIPP provides uniquely detailed information about income and public program use, income (including poverty, hardship, and wealth), detailed relationships between all household members, and life events (including marriage, cohabitation, fertility, schooling, work, and migration). For example, SIPP has been used to obtain estimates of household clustering in public program use across multiple programs (Borjas & Hilton, 1996), trends in the use of means-tested public programs (Shaefer & Edin, 2013), and trends in children's coresidence with distant kin and unrelated household members (Pilkauskas & Cross, 2018). This content is a core strength, and its dissemination is central to SIPP's mission. Yet the rich detail may pose disclosure risks, particularly for individuals who are known to be SIPP respondents.

In the analysis of the most cited and recent articles, the key independent and dependent variables were also identified. As shown in Figure 9-2, roughly a quarter of the articles relied on data on wealth, disability and care, and (detailed) income or labor force measures as an independent or dependent variable (one exception is that only 14% of articles published in the

Copyright National Academy of Sciences. All rights reserved.

last 12 months relied on disability or care variables). Roughly 15% relied on public programs that use indicators, and between 5% and 10% relied on SIPP data on life events (e.g., the timing of marriages, births, educational degrees), hardships (with even more among articles on hardship published in the last 12 months), and immigration. Overall, nearly all studies surveyed (86% of the most-cited articles and 87% of recent articles) used variables in these categories as key independent or dependent variables.

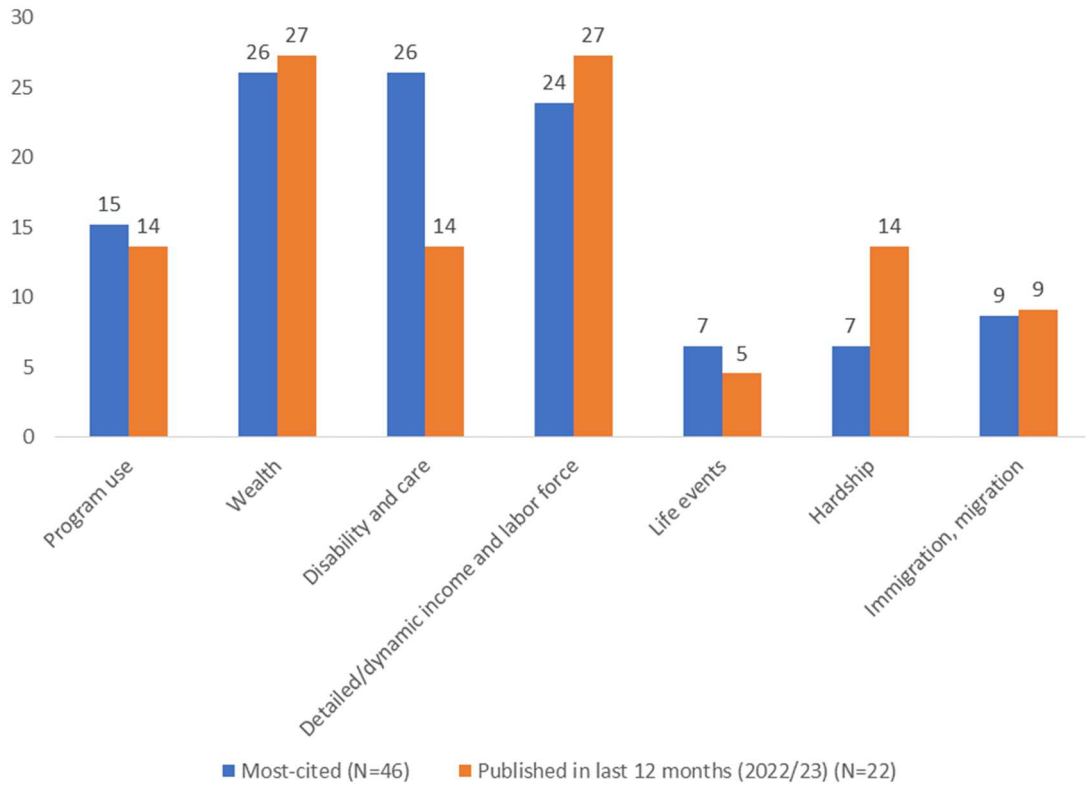


Figure 9-2 Unique SIPP content used in the most cited and recent studies (percentage).
NOTE: Each article can be included in multiple categories. Articles were drawn from a search using Scopus on February 7, 2023.

However, while the articles focused on a few key subject areas, their analysis spanned a larger range of areas. Based on responses to the call for information from SIPP users (see Appendix E), most responding SIPP users tend to use variables gathered across multiple subject areas. Though the call for information was not designed to produce nationally representative

Copyright National Academy of Sciences. All rights reserved.

results, two questions in the call are particularly relevant here. One question asked which components of the data had been used, and another asked in what fields the respondent had published based on SIPP. Every one of 26 listed modules (or topic areas) was used by at least three respondents, with four respondents adding in additional topic areas (Figure 9-3). The most commonly used modules were income, employment and earnings, demographics, family and household, education, and poverty. SIPP data users typically used multiple modules, with every respondent using at least two of the listed modules, and one using all 26. More than half used between one and 10 modules, but if a package of the top 10 modules were created, that would have met the needs of only six of the 41 users. No two respondents used exactly the same set of modules.

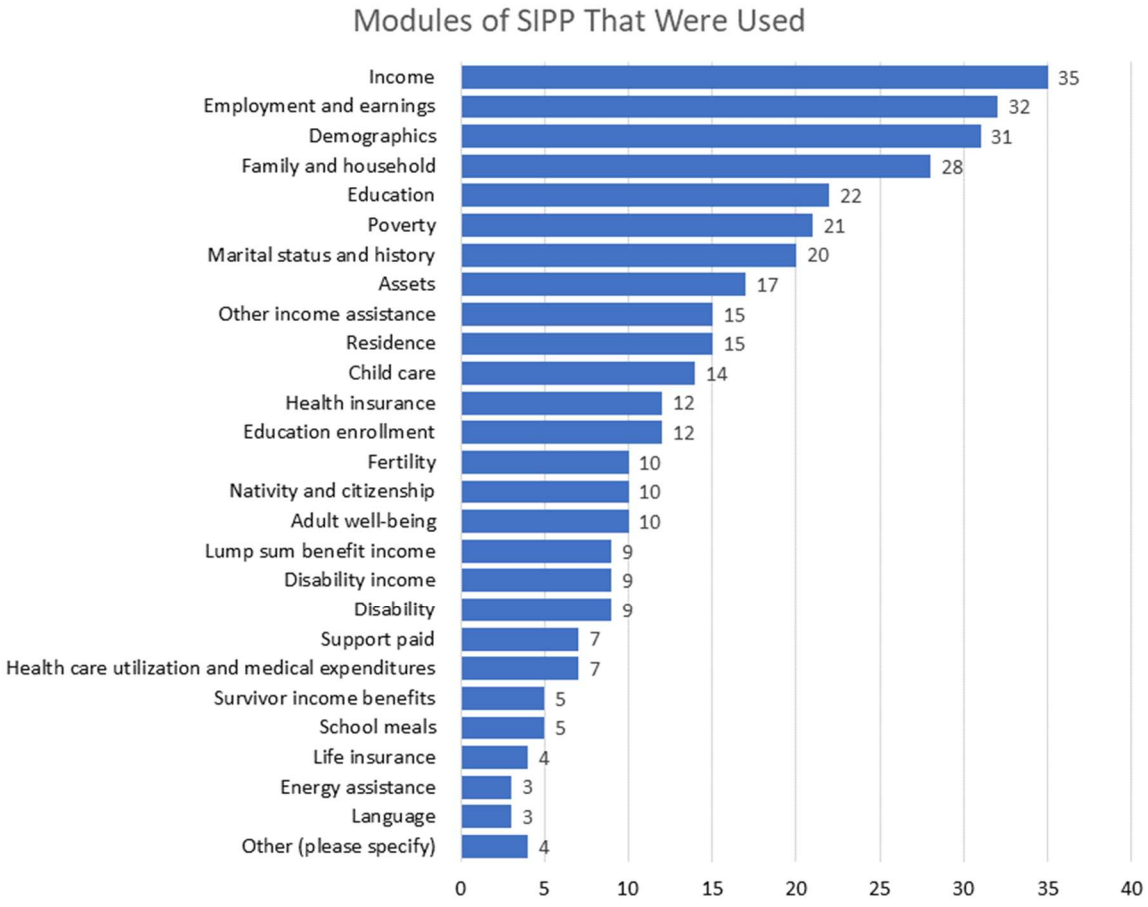


Figure 9-3 Number of respondents reporting use of various SIPP modules.
SOURCE: Call for information (see Appendix E).

Copyright National Academy of Sciences. All rights reserved.

This last point is important because one way to reduce disclosure risk might be to create separate, non-linkable files, each with a different subset of variables. For example, one file could include demographics, income, poverty, and program use, another file could include demographics and life history events, and so forth. Nevertheless, our findings suggest that this approach might not be feasible.

Use 2: Longitudinal Analysis

Another core strength of SIPP is its ability to monitor change. The longitudinal design of SIPP enables users to examine short-term change for individuals, families, and households for the various indicators noted above, including stability and instability in outcomes like employment, income, program participation, household composition, and family structure. A particularly valuable feature of SIPP that distinguishes it from another major longitudinal survey, the Current Population Survey (CPS), is that it follows individuals even if they move to new households. The longitudinal data design of SIPP supports statistical methods like fixed-effect models, difference-in-difference methods, event history models, and repeated measurement designs. Assembling multiple SIPP panels over time also enables researchers to monitor trends in outcomes and transitions for groups or populations. This allows the SIPP to be used as repeated cross-sections for trends analysis, such as trends in poverty, wealth, marriage/cohabitation, and so on.

Other data sources, such as CPS or the American Community Survey (ACS), are also suitable for repeated cross-sectional analysis, but SIPP's content allows for some use cases not possible in these other data sources. Furthermore, SIPP allows one to combine longitudinal and cross-section applications, such as trends in income instability or program dynamics. However, the detail of longitudinal data, particularly dates of events (e.g., births, marriages), may pose disclosure risks. As with data on program participation and income (noted above), longitudinal data may require protection through perturbation, that is, through swapping, masking, noise infusion, synthesis, or other techniques. It may be particularly challenging to create synthetic data that adequately represent within-person and within-household longitudinal relationships.

As shown in Figure 9-1, 61% of the most-cited SIPP studies and 68% of recent SIPP studies capitalized on the longitudinal features of SIPP. Of these, the majority—roughly half of

Copyright National Academy of Sciences. All rights reserved.

all studies we surveyed and about three-quarters of the longitudinal studies—examined within-person or within-household change over time.

Use 3: Analyses Relying on Granular Data and Complex Recoded Data

Data analysts often combine variables in creative ways to extract information that is not directly asked about in the SIPP questionnaire. For example, analysts overlaid respondents' marriage and birth histories to infer unmarried childbearing (Gibson-Davis, 2011) and multiple-partner fertility (Stykes & Guzzo, 2019). Another study combined information on the timing of the entry and exit of all household members as well as their relationships to the householder, and found high levels of household instability among Mexican immigrants due to the rapid entry and exit of friends and non-nuclear kin (Van Hook & Glick, 2007). Still others combine SIPP data with outside information to glean insight about people's lives. For example, Hall and Greenman (2015) linked SIPP data on occupation to Bureau of Labor and Statistics data on occupational fatalities and hazards. They found that undocumented immigrants are more likely to work in dangerous jobs, and were rewarded less in earnings for working in such jobs, than other groups.

Of the studies we surveyed, nearly half (43% of the most-cited studies and 45% of recent studies) used the SIPP data in a complex way as just described. This is important, because these kinds of analyses require access to detailed codes (e.g., timing of life events, occupational codes) that have disclosure risks. Such information may be masked or distorted in public data (depending on mode of access), and it is unclear whether synthesized data could account for complex interactions of multiple variables, particularly if measures are constructed with outside information.

These findings are consistent with results from the panel's call for information from SIPP users. As discussed in Chapter 2, some of the difficulties experienced by SIPP data users of the public-use files reflect their need for extensive and detailed data: eight respondents stopped their analysis or changed their approach because of a lack of detailed data, and four reported that the SIPP files lacked needed measures. Although those reporting difficulties were in the minority (24%), these responses have implications for whether data suppression or recoding would significantly damage data usability.

Use 4: Causal Effects of Public Policies

Copyright National Academy of Sciences. All rights reserved.

As mentioned before, SIPP is frequently used to analyze the impact of public policies. Across the studies we surveyed, about one-quarter (24%) of the most highly cited studies fell in this category. Fewer recent studies (9%) made causal inferences about policies (we suspect this difference is partially attributable to causal analyses being more likely to be cited). These analyses capitalize on SIPP's rich data on income and program use and on its ability to track changes for individual and family outcomes over time and across a variety of policy contexts. Such analyses often rely on state-level policy variables, whereby state level policies and policy changes (e.g., TANF or SNAP program rules, minimum wage) are attached to individuals, matching on state of residence. Sometimes, more detailed geographies are required, such as county or metropolitan area (e.g., Ribar, 2005). Policy effects are then estimated with difference-in-difference models, state or county fixed effects models, and instrumental variable approaches, many of which depend on the availability of geographic identifiers.

Despite the importance of this type of work, geographic identifiers are disclosure risks and often may not be released in raw form in public-use data files (as discussed in Chapter 8 relating to SIPP). Because of its reliance on geographic identifiers, public policy research could be particularly sensitive to masking, noise infusion, and synthetic data that distort geographic identifiers.

Use 5: Analyses Relying on Administrative Record Linkages

Finally, SIPP users have linked SIPP data with administrative records from the Social Security Administration and other agencies to supplement information collected in SIPP. This research tends to be highly innovative, yielding new findings that would not be possible with SIPP data alone, such as work on income (Meyer et al., 2021) and public program misreporting (Bruckmeier, Müller, & Riphahn, 2014), earnings growth across the life course (Kim et al., 2018), and the long-term self-sufficiency of former welfare recipients (Vaughan et al., 2021). Such analyses require access to administrative data and individual-level linkage variables (i.e., the Protected Identification Key, or PIK) that are available only at an FSRDC. For this reason, analyses that use administrative data are limited to researchers with access to FSRDCs, typically those at or near institutions with FSRDCs or who have time and money to travel to one.

Only a small share of articles among those surveyed relied on administrative record linkages (4% of the most cited and 5% of the recent articles). This may reflect the costs and

Copyright National Academy of Sciences. All rights reserved.

difficulties of obtaining access to and working in an FSRDC (a topic discussed later in the section on accessibility), and the fact that most of the published research could be accomplished with public-use data.

After identifying the most common uses of SIPP, it is important to evaluate how difficult it would be to carry out such analyses for the different modes of access. Such an exercise would help ensure that core SIPP strengths remain intact even as changes are made in mode of access. Table 9-1 illustrates how such an evaluation could be organized, with separate evaluations conducted within each cell. One would assess whether the necessary variables, response categories, and linkage variables would be available to conduct the analysis. To consider a simple example, a descriptive analysis of demographic trends or disparities in poverty levels—measures collected routinely in SIPP—may be feasible in any access mode. A specific example is provided in Box 9-1 of a descriptive study of visual impairment, which would be feasible to conduct using the simplest mode of access, an online tabular/analysis builder.

Copyright National Academy of Sciences. All rights reserved.

TABLE 9-1 Matrix for Evaluating Feasibility with the Context of Various Modes of Access

	Mode of access				
Uses	Online analysis builder	Public-use microdata	Synthetic data	Secure online data access	FSRDC
1. Analyses relying on unique SIPP content					
2. Longitudinal analysis					
3. Analysis relying on granular data and complex recodes					
4. Causal effects of public policies					
5. Analysis relying on administrative record linkage					

Copyright National Academy of Sciences. All rights reserved.

BOX 9-1

Illustration of Feasibility: Descriptive Analysis Using Unique SIPP Content

The economic burden of major adult visual disorders in the United States

This study used wave 5 of the 1996 SIPP panel (conducted in 1997) to estimate productivity losses (in labor force participation and wages) for visually impaired and blind persons ages 40 to 64. The authors used respondents' answers to the questions, "Do you have some difficulty in seeing words or letters?" to identify respondents with visual impairment and "Do you have severe difficulty in seeing words or letters?" to identify respondents with blindness. They then compared the mean labor force participation rates and average annual earnings for persons with blindness, visual impairment, and normal vision among those in the same age group. The authors used SIPP because it contained variables on vision impairment and blindness as well as measures of economic activity. This analysis was one piece of a larger study assessing the economic burden of visual disorders.

An online tabular/analysis tool is well-suited for studies such as this one. It allows users to define sample restrictions (i.e., adults ages 40 to 64) and compare mean values (i.e., proportion in the labor force and average annual earnings) by groups (i.e., by degrees of visual impairment).

SOURCE: Committee generated from Rein et al., 2006.

However, a more complex study may be feasible when using some but not other modes of access. Consider a potential study in which a researcher wants to estimate the impacts of working in a hazardous job. The researcher creates a measure of occupational hazard by merging occupational characteristics obtained from an outside source to SIPP respondents' records. This kind of analysis, relying on granular occupational codes and outside data, is quite common, so the Census Bureau will want to preserve the ability of researchers to conduct this type of study.

Copyright National Academy of Sciences. All rights reserved.

Yet this kind of analysis would not be feasible to conduct with an online tabular/analysis builder because of the complexity of the data manipulations that would need to occur. Additionally, it is possible that a review of disclosure risk would indicate that the release of granular occupational codes in a public-use file would pose too large a privacy risk. As a consequence, these codes may be suppressed from the public-use data, thereby making them inadequate for the analysis. Synthesized data may also be inadequate, because such data would not necessarily account for outside data on occupational characteristics; the relationships of interest are unlikely to have been modeled when the synthetic data were created. However, the study may be feasible to conduct through secure online data access (SODA). Another example of a feasibility analysis for a similarly complex longitudinal study on the effects of spouses' work on employment is provided in Box 9-2.

Overall, a feasibility analysis such as illustrated here will help the Census Bureau anticipate how changes in mode of access will impact the volume and type of research being conducted with SIPP data.

BOX 9-2

Illustration of Feasibility: Longitudinal Analysis with Household Relational Data

Reinforcing separate spheres: The effect of spousal overwork on men's and women's employment in dual-earner households (2010)

This study used the entire 1996 SIPP panel, using interviews conducted between 1995 and 2000, to examine the likelihood of quitting a job associated with spouses' work hours. The author conducted a multilevel multivariate logistic regression on a person-level longitudinal dataset that was linked to spousal characteristics as well as household and individual variables. The outcome variable—the probability of quitting—was derived by identifying employed individuals who responded that they were not in the labor force in the next wave. The key independent variable—spouse's work hours—was matched to the analytical dataset using SIPP's household and relationship identifiers.

Copyright National Academy of Sciences. All rights reserved.

An online tabular/analysis builder would be insufficient to conduct such an analysis, as it requires complicated recoding and merging as well as individual-level microdata to estimate the multivariate regression models. Furthermore, using a synthetic data file may be problematic unless it explicitly retains not only spousal employment characteristics within a wave but also employment transitions between waves. If this type of analysis is not possible using future SIPP public-use files, the most appropriate mode would be to access the restricted microdata through the SODA as described in Chapter 5.

SOURCE: Committee generated from Cha, 2010.

ACCESSIBILITY

Accessibility is a dimension of usability that refers to the cost and difficulty of accessing data. As accessibility declines, the number of SIPP users will also likely decline, particularly among those with fewer resources, reducing the potential impact of SIPP. An important task facing the Census Bureau is to determine how users would be impacted by making particular modes of access available or unavailable, how the impact would vary across different user groups, and how to mitigate any negative consequences.

In Table 9-2, three types of users are considered: (1) those with limited resources and expertise, such as undergraduate students and researchers without access to institutional review boards (IRBs); (2) those with moderate levels of resources and expertise, such as researchers at lower-resourced institutions, some nongovernment think tanks, or researchers-in-training such as graduate students; and (3) users with high levels of resources and expertise, such as researchers who have grant funds and work at FSRDC member institutions with the staffing to manage data use agreements and are close to an FSRDC.

Additional user categories may be required to recognize users with distinctive data needs. The survey of SIPP publications described earlier suggests that the bulk of SIPP users probably fall in the second and third categories above—that is, those with the greatest resources available. Nearly all the first authors were faculty members at universities or researchers in research organizations. Sixty-one percent of the first authors of the top-cited articles were either based at

Copyright National Academy of Sciences. All rights reserved.

institutions that hosted an FSRDC (15%) or worked in a city with an FSRDC (46%) (Figure 9-4). The first authors of recently published articles were not as proximately located: 41 percent were located near an FSRDC (18% in the same institution and 23% in the same city). Nevertheless, between 40% and 60% were not located near an FSRDC, which would likely pose challenges to users if the SIPP data were modified in such a way that they could no longer use it to conduct their work unless they did so in an FSRDC.

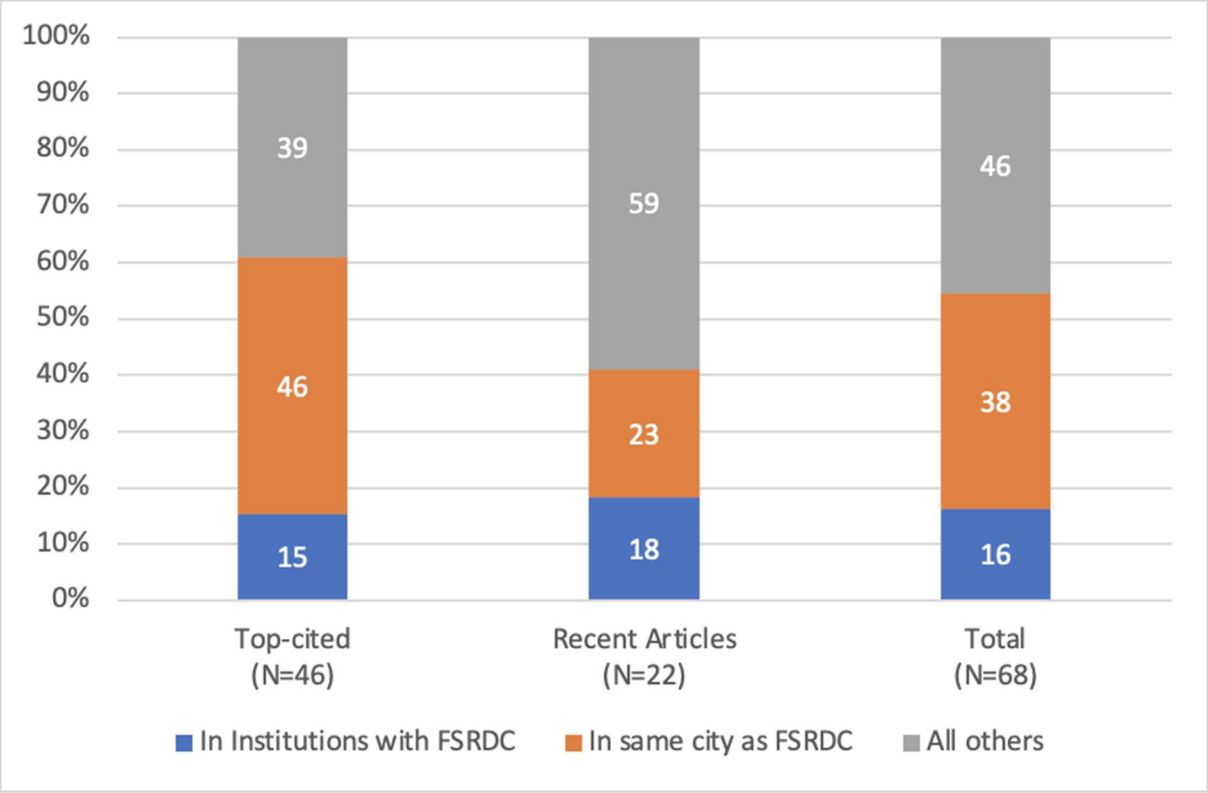


Figure 9-4 Share of first authors at institutions or in cities with Federal Statistical Research Data Centers (FSRDCs).

Moreover, it is critical to note that proximity to an FSRDC does not fully capture the barriers to working in an FSRDC. Regardless of their proximity to an FSRDC, most published data analyses that we surveyed were based on public-use data, and only 4% required access to restricted data. Our SIPP user survey yielded similar results. In that survey, we asked about three current modes of access: public-use microdata files, synthetic SIPP data, and restricted SIPP data available in the FSRDC. The respondents were most likely to have reported using the public-use file. Out of 65 respondents, 42 downloaded the public-use file, 18 accessed the synthetic data, 17

Copyright National Academy of Sciences. All rights reserved.

downloaded tables or reports, and only 7 (11%) accessed the data through an FSRDC. Two of the respondents had used all four sources, 24 used only one, and 24 used some combination of multiple sources. Overall, this suggests that even for data users who are located near an FSRDC, either there are considerable barriers to accessing them or no need to use them because the public data file was sufficient.

Table 9-2 Example of Evaluation of Accessibility by Mode of Access and User Type (1=low to 4 = high)

		Mode of Access				
User Type		Table maker	Public-use microdata	Synthetic data	Secure online data access	FSRDC
1	Low resources/expertise (e.g., undergraduate student)	4	3	1	1	1
2	Moderate resources/expertise (e.g., researcher-in-training; researcher at a think tank or R2 or R3 university)	4	4	3	2	1
3	High resources/expertise (e.g., researcher with grants at R1 university)	4	4	4	4	3

NOTE: The Carnegie Classification of Institutions of Higher Education divides institutions into multiple categories. Among doctorate-granting universities, R1 universities have the highest level of research activity, R2 universities have the next highest level, and R3 universities have the least. Other categories include baccalaureate colleges, master's colleges and universities, baccalaureate/associate's colleges, and associate's colleges.

After classifying users according to their data needs and ability to access restricted data, it is important to evaluate their ability to access SIPP data for each proposed mode of access. To illustrate how accessibility might be evaluated, Table 9-2 is populated with qualitative assessments of the ability of each user type to access SIPP data using the various access modes,

Copyright National Academy of Sciences. All rights reserved.

indicating a "high" (= 4), "moderate" (= 2 or 3), or "low" (= 1) likelihood of accessing the data. These evaluations are based on the professional experience of panel members; empirical estimates of access could be gathered through online user surveys.

It is unlikely that low-resourced users (e.g., undergraduate students) would be able to use any SIPP data source unless it were made available through an online tabular/analysis builder or public-use file. This group generally does not have the funding or institutional support to access restricted-use data, and only some have sophisticated data analysis and statistical skillsets. This group would stand to lose the most if public data were no longer available. At the other end of the scale, high-resourced users could probably access data regardless of access mode, although making SIPP data available only in a highly restrictive environment would slow down research or steer researchers away from SIPP and toward other data sources, such as the Panel Study of Income Dynamics (PSID), that are less difficult to access. In contrast to low- and high-resourced users, the moderately-resourced users would gain the most from the availability of synthetic data or secure online data access. They would use microdata and administratively linked SIPP data if they could but would also be stymied if the costs of accessing the data were too high. As discussed in the next section, the costs of accessing data through an FSRDC can be quite high, and prohibitively so for all but the high-resourced users. However, the costs of accessing SIPP data through SODA may be considerably lower for moderately resourced users. Many researchers in this group already access other data through SODA (e.g., PSID, NCES, Add Health, and HRS data), and they are accustomed to working within the confines of data use agreements.

SODA thus provides the best current prospect for maintaining usability among both low- and moderately-resourced SIPP users.

How the Introduction of SODA and Reinterpretation of Title 13 Would Affect Accessibility

Here a more detailed illustration is provided on how the addition of a new mode of access—namely, SODA—could enhance accessibility for users. Such an addition would likely entail a reinterpretation of Title 13, particularly if SODA users were not required to obtain Special Sworn Status or demonstrate a direct benefit to the Census Bureau (an issue we discuss

Copyright National Academy of Sciences. All rights reserved.

later in this chapter). Note that Chapter 5 further details how SODA can expand access to restricted data.

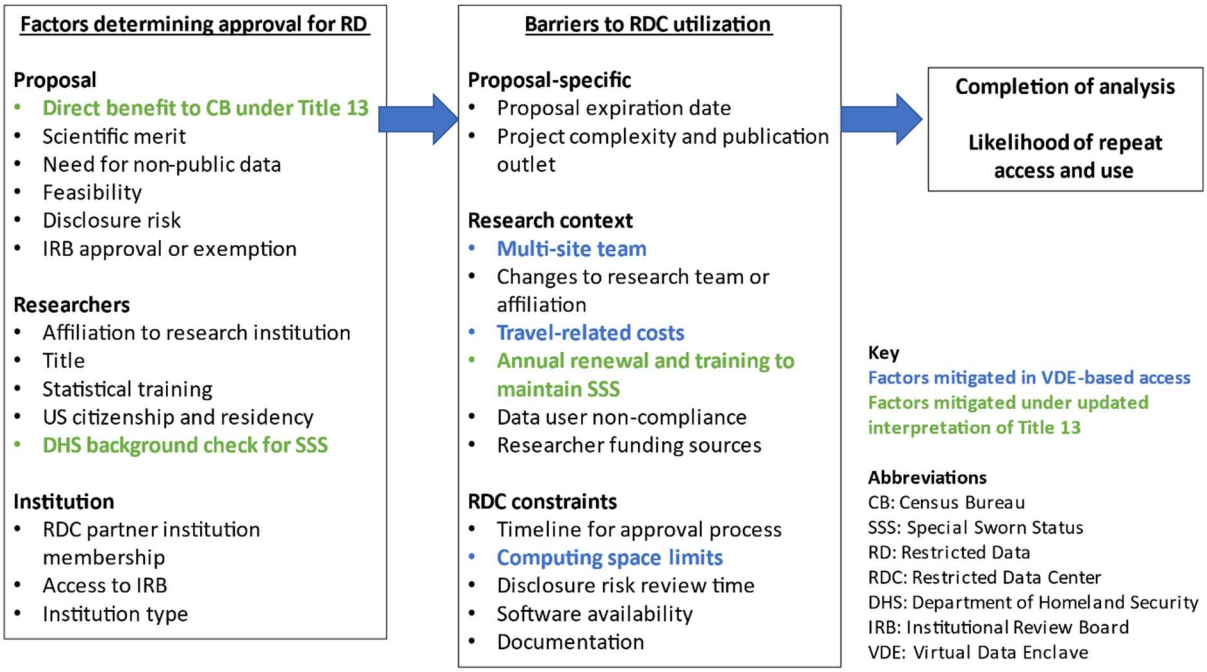


Figure 9-5 Determinants and barriers to accessing restricted Census data through the current FSRDC system.

NOTE: Green text indicates barriers to access that would be mitigated through a reinterpretation of Title 13, and blue text indicates barriers that would be mitigated through SODA.

For researchers who wish to use the restricted SIPP data, obtaining approval through the current FSRDC system requires a multistage process that can take several months. The proposal must demonstrate that the research project's predominant purpose is to benefit Census Bureau programs, and this criterion carries the greatest weight during the approval process over other considerations, including scientific merit, need for restricted data, and feasibility. Approval or exemption from an institutional review board is also required.

Applicants must also obtain and annually renew a Special Sworn Status. Non-US citizen researchers must have lived in the United States for longer than 36 out of the prior 60 months and be affiliated with a US-based institution. Clearances for non-US citizen applicants take longer to complete. The Department of Homeland Security administers Special Sworn Status applications and researchers must submit fingerprints, complete online training, submit

Copyright National Academy of Sciences. All rights reserved.

residential, travel, employment, and education histories, provide references, and complete an interview. Doctoral students who are applying for restricted data must have a faculty advisor who also has a current Special Sworn Status.

The applicant's institution can facilitate approval to access restricted data. Researchers who are not affiliated with institutions that can enter Data Use Agreements or have institutional review boards are at a disadvantage when preparing their proposals. Those who are not based at one of the 50 FSRDC partner organizations must pay an additional fee.

The structure of the current FSRDC creates unequal barriers to accessing the restricted data even for proposals and researchers that are cleared for FSRDC access. Applications, approvals, and fees are specific to a single RDC location, and multisite research teams can incur additional costs to gain access to multiple RDCs. Researchers must renew their Special Sworn Status annually, and changes in affiliation, title, or members can cause interruptions. Travel to physical RDC locations can become costly, especially for researchers who are not close to an RDC or do not have funding dedicated for data access expenses.

However, as described in Chapter 5, expanding access through SODA or by reinterpreting Title 13 (a possibility we discuss at the end of this chapter) to enable access to formerly public-use data without the need for Special Sworn Status or a long and detailed proposal would help researchers avoid the additional administrative and cost burdens of conducting research in an FSRDC. This is illustrated in Figure 9-5, where the barriers to access that would be mitigated if data were provided via SODA are shown in blue, and other barriers to access that would be mitigated through a reinterpretation of Title 13 are shown in green.

To conclude this section on usability, more research on the user community might be conducted to assess the number and types of users who would be most impacted if certain modes of access were eliminated (e.g.., a public-use file) or if new modes of access were made available (e.g., SODA). However, our preliminary research suggests the following points.

First, to maintain the value of SIPP data (and, accordingly, the size of the current SIPP user community), it will be important for the Census Bureau to continue to produce a public-use microdata file if the disclosure risks are sufficiently low to do so (see Chapter 3 for more discussion on this topic). Most SIPP research is currently conducted on the public-use file, and a large share of SIPP users do not have easy access to an FSRDC.

Copyright National Academy of Sciences. All rights reserved.

Second, to the degree that the public-use SIPP microdata become less useful due to efforts to increase privacy protections, the Census Bureau should consider providing mid-tiered access to restricted SIPP data through SODA. Current SIPP users tend to be sophisticated data users. Most of the analyses performed on SIPP are conducted by faculty members, graduate students, and professional researchers. This sophistication is reflected in the types of data analysis they perform, most of it being multivariate analyses that involve complex data recodes, longitudinal analysis, and incorporation of outside data sources. For many SIPP users, certain modes of data access will be inadequate for their work, such as an online tabular/analysis builder, or public data that suppresses variables like state, occupation, or industry.

This last conclusion was further backed by input from SIPP users as described in Chapter 2, who reported using both tabulations and statistical modelling when working with the SIPP data, and most of whom reported using data for longitudinal and cross-sectional analyses. Only 7% said their research needs could largely be met through the availability of a comprehensive set of standardized tables. By contrast, roughly three-fourths of users said that using standardized tables would meet their research needs only to a small extent (31%) or to no extent at all (43%). At the same time, the analysis of SIPP publications indicates that roughly half of the first authors are *not* located near an FSRDC. A simple reflection about the steep requirements of working in an FSRDC, combined with the panel member's professional experience, suggests that most researchers would consider accessing an FSRDC burdensome.

Third, to increase the size of the SIPP user community and expand the uses of SIPP data beyond its current scope, the Census Bureau might consider investing in highly accessible modes of access that do not require specialized skills in data analysis, such as an online tabular\analysis builder. This particular mode of access is described in detail in Chapter 7. A review of SIPP publications revealed that about 18 percent of studies are based solely on descriptive analyses, all of which could be conducted with an online data analysis tool (see Box 9-1 for a description of such a study). The availability of a simple online analysis tool has the potential to expand the use and value of the SIPP data for the purpose of descriptive analysis for people who have not historically used SIPP data and who may not have the skills to analyze microdata. Nevertheless, it is not at all clear that "if you build it, they will come." More research would help identify the number and types of people who currently do not use SIPP but who would benefit from such a

Copyright National Academy of Sciences. All rights reserved.

tool. Still further work would be involved to reach out to this group and provide training in the value of SIPP and the use of the online analysis builder.

TITLE 13 AND DATA ACCESS

A key question is whether the full set of restrictions to data access applied in FSRDCs must necessarily be applied in other tiers. This section evaluates that topic and concludes that the Census Bureau has much more flexibility in how data may be released under Title 13 than has been commonly viewed.

Title 13 defines the role and functions of the Census Bureau, and it contains provisions designed to both direct and authorize the Census Bureau to protect the confidentiality of its data. The title's provisions are thus very important when considering disclosure avoidance for SIPP. The key sections regarding confidentiality are 9(a), 23(c), and 214. Section 9(a) limits the uses of the data, prohibits the release of identifying information, and limits access to the data to sworn officers. Section 23(c) provides a vehicle for outside individuals to access the data. Section 214 specifies the penalties for unauthorized release of Title 13 data (the penalties were later increased under CIPSEA and the Evidence Act to $250,000).[63]

> §9(a). Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998 or section 2(f) of the Census of Agriculture Act of 1997— (1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or (2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or (3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

[63] 44 USC 3572(f); /https://www.govinfo.gov/content/pkg/PLAW-115publ435/pdf/PLAW-115publ435.pdf

Copyright National Academy of Sciences. All rights reserved.

> §23(c). The Secretary may utilize temporary staff, including employees of Federal, State, or local agencies or instrumentalities, and employees of private organizations to assist the Bureau in performing the work authorized by this title, but only if such temporary staff is sworn to observe the limitations imposed by section 9 of this title.
> §214. Whoever, being or having been an employee or staff member referred to in subchapter II of chapter 1 of this title, having taken and subscribed the oath of office, or having sworn to observe the limitations imposed by section 9 of this title, or whoever, being or having been a census liaison within the meaning of section 16 of this title, publishes or communicates any information, the disclosure of which is prohibited under the provisions of section 9 of this title, and which comes into his possession by reason of his being employed (or otherwise providing services) under the provisions of this title, shall be fined not more than $5,000 or imprisoned not more than 5 years, or both.

This wording allows much broader access to data than has often been recognized:

- It allows the Census Bureau to determine which data should be protected, only saying that establishments and individuals should not be identified. The Census Bureau once treated that definition as applied to information such as names and addresses, but has been broadening that definition.
- It states that the data should be used for the statistical purposes for which they are supplied.
- It leaves unspecified the procedures for determining which people and projects are authorized to access the data.
- It leaves the mode and rules for data access unspecified.

To clarify and implement Title 13, the Census Bureau has adopted two policies—DS002 and DS006. Policy DS002 states that a project must deliver a benefit to the Census Bureau's programs and activities, and that the benefit must be the predominant purpose for conducting the project.[64] The policy lists 13 criteria for determining if the Census Bureau is benefited.[65] At least

[64]www2.census.gov/foia/ds_policies/ds002.pdf
[65]www2.census.gov/foia/ds_policies/ds006.pdf

Copyright National Academy of Sciences. All rights reserved.

one of the 13 criteria must be satisfied: the first four apply generally to Title 13 data, while if Federal Tax Information is to be used then at least one of the remaining nine criteria must be satisfied. In summary, the 13 criteria require that the project benefit the Census Bureau by seeking to accomplish one of the following:

1. Evaluating the Census Bureau's data or reports
2. Analyzing trends that affect the Census Bureau's programs
3. Increasing the utility of the Census Bureau data
4. Facilitating data collection, processing, or dissemination
5. Helping understand or improve the quality of data
6. Leading to new or improved methodology
7. Enhancing the data
8. Identifying limitations or improving the sampling frames or classification schemes
9. Identifying shortcomings of current data collection programs or documenting new data needs
10. Constructing, verifying, or improving the sampling frame
11. Preparing population estimates
12. Addressing non-response, or
13. Developing statistical weights.

These criteria add considerable specificity to the language contained in Title 13 ("assist the Bureau in performing the work authorized by this title") and generally focus on improving the Census Bureau's operations or the data file itself. However, some of the criteria could be given a broader meaning. For example, one might argue that examining the interrelationships among variables (e.g., documenting that it is important to adjust for differences in race/ethnicity) ultimately increases the utility of the data (Criterion 3) and helps to understand the data (Criterion 5). However, a lack of clarity makes it difficult for potential data users to know whether a data request is likely to be accepted, potentially preventing the submission of what would be considered acceptable proposals. The criteria also tend to imbed the assumption that the Census Bureau's purposes are only to collect and distribute the data, not to encourage use of the data or support policy or social science research.

Copyright National Academy of Sciences. All rights reserved.

Policy DS006 defines how non-employees may access Title 13 data. It sets three criteria for determining if a project qualifies, three individual/organization criteria for determining who may receive Special Sworn Status, and four criteria for determining if a project can take place at a non-Census Bureau facility. In sum, the project must require access to Census Bureau confidential data, benefit the Census Bureau's Title 13 Programs, and be viable. The individual and organization must have a good track record for handling sensitive or confidential data, have no identified conflict of interest, and pass a background investigation. Access to the data off-site depends on the type of project proposed (internal, joint, reimbursable, external, or oversight). External projects must be carried out at a Census Bureau facility (including FSRDCs), while other projects may be granted an exemption if they provide a technical or logistical advantage, meet required security models, have legal or regulatory functional separation of the data collected for statistical purposes (for governmental agencies or organizational units), and obtain Data Stewardship Executive Policy Committee (DSEP) approval prior to any major commitment of resources.

Again, these criteria add considerable specificity to the goal that temporary staff must be "sworn to observe the limitations imposed by section 9 of this title." For example, Title 13 does not say that such researchers must be prequalified, let alone how they should be prequalified.

Because many of the specifics are contained in policies rather than in Title 13 itself, it is possible to change many of the requirements without changes in legislation. In fact, the rules have been updated in the past; they were originally established in 2002,[66] while DS002 was last signed in 2018, and DS006 was updated in 2009 (in this case, only to update organizational names).

One aspect of change, and a key focus of this report, is that the definition of personally identifying data is evolving. It once focused on names and addresses, but there now is a recognition that many data elements present disclosure risks to the degree that similar data can be found in and matched with other databases. The federal government's priorities also have changed, giving a high priority to increasing access to data and to using tiered access as a tool for supporting that goal. OMB's guidance on the Information Quality Act states:

[66]www.governmentattic.org/6docs/CensusBureauDataStewardship_2002-2010.pdf

Copyright National Academy of Sciences. All rights reserved.

Implementation Update 3.4: Agencies should prioritize increased access to the data and analytic frameworks (e.g., models) used to generate influential information. All data disclosures must be consistent with statutory, regulatory, and policy requirements for protections of privacy and confidentiality, proprietary data, and confidential business information.

Implementation Update 3.S: Agencies should explore methods that provide wider access to datasets while reducing the risk of disclosure of personally identifiable information. In particular, tiered access offers promising ways to make data widely available while protecting privacy. Implementation of such approaches must be consistent with principles for ethical governance, which include employing sound data security practices, protecting individual privacy, maintaining promised confidentiality, and ensuring appropriate access and use.

To ensure reproducibility, the *Guidelines* set an expectation of access to data underlying influential information, subject to "compelling interests such as privacy, trade secrets, intellectual property, and other confidentiality protections." Since the 2002 *Guidelines,* the technology for allowing protected access to data has progressed significantly. New approaches to secure data access using cutting-edge technologies reduce the risk of re-identification and therefore may mitigate certain privacy risks associated with providing such access. Risk reduction techniques include creating multiple versions of a single dataset with varying levels of specificity and protection (sometimes referred to as "tiered access").

The virtue of tiered access is that data users who wish to conduct activities with a statistical purpose without first obtaining special authorization have access to the versions of the data in the least restricted tiers, allowing them to conduct research while protecting confidentiality. Such approaches to increasing access to data for statistical purposes could be considered by more federal agencies, thereby allowing stakeholders to replicate analyses and explore the sensitivity of the conclusions to alternative assumptions while accessing only the data they need. As agencies consider adding intermediate tiers between fully open and fully closed, they must

Copyright National Academy of Sciences. All rights reserved.

build in sufficient controls to monitor who is accessing the data and allow access only for authorized purposes. (Vought, 2019a)

Tiered access serves a dual function. First, it allows increased access to data that otherwise might not be available, helping to keep SIPP data usage from being undercut if some data in the public-use file are found to be disclosive. Second, and at the same time, tiered access protects data security through data minimization, "giving access to the least amount of data needed to complete an approved project" (Commission on Evidence-Based Policymaking, 2017, p. 38). The federal government's move toward increasing access to data, while also respecting confidentiality, is also expressed in the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) (Vought, 2019b). From this perspective, the largely internal approach to defining "benefit" in Policy DS002 is no longer appropriate.

The Census Bureau might consider modernizing its interpretation of Title 13 in several ways.

- First, Title 13 does not require that data access benefit the Census Bureau specifically, instead focusing on accomplishing the *work* of the Census Bureau. Since part of the work of the Census Bureau is to support research and open access to data, the criteria specified in Policy DS002 are fine in themselves, but additional and more broadly-based criteria might be added. When the Census Bureau determines that a survey such as SIPP accomplishes its purposes, then supporting research based on the survey also accomplishes its purposes.
- Second, the purposes of the Census Bureau are also defined by Title 44 (also known as the Evidence Act), which applies to all federal statistical agencies. Thus, expanding data access for evidence building might also be considered as a benefit to the Census Bureau.
- Third, while Title 13 requires that users be sworn to protect confidentiality, it does not impose conditions on who may be sworn. DS006 requires a background investigation, which itself involves additional criteria (e.g., it is very difficult for a non-U.S. citizen to be authorized). These requirements are both excessive and time-consuming. The process of preparing a justification and application package itself is onerous. Conducting some kind of review is appropriate, but the Census Bureau

Copyright National Academy of Sciences. All rights reserved.

might examine what other agencies and organizations are doing and create a simpler and more expedited process.

- Fourth, rather than offering a largely dichotomous choice between public-use data and FSRDCs, the Census Bureau should consider allowing for multiple levels of security depending on the sensitivity of the data. Some types of data, such as names and addresses, deserve a very high level of security, and blended administrative data often do as well, since these may have a separate set of restrictions that are imposed by another agency. Other data might be deemed less sensitive, requiring more constraints than a public-use file but fewer constraints than an FSRDC. The Census Bureau has already moved in this direction by offering the synthetic beta file and by increasing virtual access to Title 13 data. To the degree that re-identification studies indicate that some of the public-use data present a disclosure risk, these data might be moved to an intermediate security level rather than being available only through FSRDCs.

The panel is not taking a position on whether the conditions for access to FSRDCs should be revised, other than supporting the Census Bureau's recent move to allow virtual access to the data. Based on the logic presented above, it would be possible to change the conditions for access to FSRDCs, since these conditions are based on policy decisions on how to implement Title 13 rather deriving from Title 13 itself. However, FSRDCs would contain the most sensitive SIPP data, which deserve stronger levels of access, and FSRDCs often are used for blending data from other sources, meaning that the restrictions imposed by other agencies also are a factor.

CONCLUSIONS AND RECOMMENDATIONS

Conclusion 9-1: The research studies based on SIPP use many variables covering many different topics. Selecting only a limited set of modules for public release will not satisfy the needs of many SIPP users.

Conclusion 9-2: A formal evaluation of the tradeoff between risk reduction and usability would help inform the selection of disclosure limitation approaches for different tiers of access to SIPP data.

Copyright National Academy of Sciences. All rights reserved.

Conclusion 9-3: SIPP users and potential users have a variety of needs, and a single mode of access is unlikely to suit all of them.

- *Most SIPP data users' needs cannot be met by providing standardized tables, while accessing SIPP data only through an FSRDC would be overly burdensome for many SIPP data users and create inequities in access.*
- *An online tabular/analysis builder for accessing SIPP data has some potential for expanding the SIPP user community to include those who may not have the analytical skills or resources to work with the public-use microdata files.*
- *Public-use files provide users with the level of control that they need in working with the data.*
- *A mid-tier mode of access—such as secure online data access—would provide for those situations in which desired data are not available on a public-use file.*

Conclusion 9-4: SIPP respondents and potential respondents would benefit from a better understanding of the disclosure limitation methods used to protect their information and how these methods safeguard confidentiality.

Conclusion 9-5: SIPP data users would benefit from additional communications regarding changes to data access and disclosure protection, how to use the SIPP data, how to compute standard errors in the presence of added noise to protect data, and the types of analyses that can be conducted using various tiers of access. Forms of communication could include documentation, workshops and webinars, and sample code. Carrying out statistical inferences on confidentialized data will require releasing details of the disclosure limitation methods.

Conclusion 9-6: Title 13 requires that confidentiality be protected but does not specify the means of protection. The current use of Special Sworn Status and FSRDCs is based on the policy interpretations of Title 13, not on Title 13 itself.

Recommendation 9-1: The Census Bureau should conduct regular assessments of the validity and reliability of estimates generated from SIPP data products treated for disclosure protection and communicate the results to SIPP users.

Copyright National Academy of Sciences. All rights reserved.

Recommendation 9-2: When considering which access modes to prioritize, the Census Bureau should evaluate feasibility for the most common SIPP uses and those that exploit the unique characteristics of SIPP and could not be obtained from other datasets.

Recommendation 9-3: The Census Bureau should seek to continue providing public-use files for SIPP users, assuming that appropriate disclosure avoidance techniques can be adopted and that quasi-identified and sensitive variables are treated for disclosure avoidance. The variables to be included should depend on the results of disclosure risk studies.

Recommendation 9-4: Given the differences in user needs and approaches, the Census Bureau should offer multiple tiers of access, with approaches for confidentiality protection applied.

Recommendation 9-5: The Census Bureau should modernize its interpretation of Title 13 consistent with changes in technology, policy guidance, and legislation (i.e., the Evidence Act and the Information Quality Act). Doing this should enable the development and operationalization of tiered data access in which (1) the types and levels of protection and access vary across tiers; (2) the requirement of benefit is redefined to include evidence building, the productive use of Census Bureau–developed data, and other statistical purposes; and (3) the eligibility criteria for accessing the data in these tiers are broadened.

Copyright National Academy of Sciences. All rights reserved.

10
Conclusions and Recommendations

While personal data were once scattered in paper forms across multiple locations, now large amounts of data can be harvested online and combined, making it possible to develop powerful databases that can be highly comprehensive. Several commercial organizations make a practice of collecting and selling such information, and data obtained by federal agencies are also aggregated and used for a variety of other reasons, including research and criminal activities. In this context, those responsible for performing surveys face increasing challenges in protecting the identity of survey respondents and maintaining the confidentiality of their data while allowing the data to be used for their intended purposes.

SIPP AND THE RISK OF DISCLOSURE AVOIDANCE

The nature of SIPP data greatly complicates approaches to disclosure avoidance. For example, SIPP collects much more data than the decennial census, with over 3,000 variables released. Some SIPP data are highly personal (such as financial information), and the data are also highly detailed, including continuous data among the financial measures. Further, the data do not just describe individuals but link each individual as part of a larger household, and they are longitudinal as well, with four years of monthly data. The measurement of substantial changes, such as a change in jobs or loss of a job, a change in marital status, or a change in the number of children, could provide key information that would be helpful in identifying a respondent.

A recent re-identification study conducted by the Census Bureau provides encouraging findings about the level of disclosure risk that SIPP poses. However, the panel views that study as too limited to provide a definitive measure of disclosure risk. It focused only on two federal databases (IRS and Social Security data); though those are among the best databases, they lack information contained in commercial and state databases that also would be useful for

Copyright National Academy of Sciences. All rights reserved.

identifying SIPP respondents. Further, the study took only a limited look at how the concatenation of data from multiple people in the same household might influence disclosure risk, and it took a limited view of how measures of change over time might affect disclosure risk. Thus, there is not yet an accurate assessment of the level of risk in the release of SIPP data.

Recommendation 3-1: The Census Bureau needs to establish a systematic and on-going program for assessing the disclosure risk of SIPP data products, especially microdata public-use files. This program should include both empirical and analytical approaches to assessing the risks of re-identification attacks, comparisons with a variety of external databases, probabilistic models, and the risks present in both cross-sectional and longitudinal data. All forms of data releases should be evaluated.

Recommendation 4-2: Based on the results of the disclosure risk assessment exercises recommended in Chapter 3, the Census Bureau should evaluate what disclosure avoidance methods are needed to achieve acceptable levels of disclosure risk for each mode of access.

SIPP USE AND USABILITY

SIPP is widely used, including by academics, policy researchers, and multiple departments within the federal government. Its use spans multiple disciplines, including economics, political science, sociology, anthropology, and demography, and multiple policy areas, such as employment, disability, health care, and retirement. A Scopus search identified 711 journal articles using SIPP data, although this is likely to be an underestimate because Scopus checks only the title and abstract, not the full text. A Google Scholar search (which is a full-text search) produced more than 20,000 results, though this is likely to be an overcount because Google Scholar counts different versions of an article separately, and texts may mention SIPP without using SIPP data.

To assess usability, the panel chose the following key components: enabling researchers to make valid and reliable inferences, having sufficiently comprehensive data to support the desired analyses, providing access in a way that is reasonably timely and does not impose hardship, and limiting the costs or difficulties of using the data. Built within these elements is

Copyright National Academy of Sciences. All rights reserved.

also the concept of equity: if data access depends on what resources are available to researchers, then there is a lack of equity in access. Providing access through a public-use file comes the closest to meeting these criteria, and appears overwhelmingly to be the most widely used means of access to SIPP data. In the panel's opinion, diminishing that capacity would both undercut valuable research and lessen the value of SIPP to researchers and policymakers, possibly threatening the survival of SIPP as an ongoing survey. By contrast, the primary other means of accessing the bulk of the data is through FSRDCs,[67] and obtaining access through FSRDCs involves the need to get appropriate approvals, financial costs for using the facilities (even for online use), and a background check. The Census Bureau recently expanded virtual access to Title 13 data at FSRDCs, helping to lower the logistical barriers, but the approval process still remains time-consuming and difficult.

As discussed in Chapter 1, previous National Academies reports also stated the importance of releasing a public-use file containing microdata. See recommendations 5 (National Research Council, 1989), 6-3 (National Research Council, 1993), 3-6 (National Research Council, 2009), and 9-1 (NASEM, 2018).

Recommendation 9-3: The Census Bureau should seek to continue providing public-use files for SIPP users, assuming that appropriate disclosure avoidance techniques can be adopted and disclosive and sensitive variables are treated for disclosure avoidance. The variables to be included should depend on the results of disclosure risk studies.

The panel also considered how best to maintain usability if later re-identification studies indicate that the public-use file presents too much disclosure risk. In the panel's opinion, one part of that answer is to continue producing a public-use data file, even if some data must be removed or otherwise altered to protect respondents' confidentiality. For those researchers who need more data than can be provided in the public-use file, the panel concluded it is important to provide one or more alternative modes of access to the data. Those modes are discussed later in this chapter, but following are the recommendations relating to usability criteria that should be applied to those alternative tiers.

[67]Some SIPP data are also available through the synthetic beta file, but it has only a limited number of variables.

Copyright National Academy of Sciences. All rights reserved.

Recommendation 9-1: The Census Bureau should conduct regular assessments of the validity and reliability of estimates generated from SIPP data products and communicate the results to SIPP users.

Recommendation 9-2: When considering which access modes to prioritize, the Census Bureau should evaluate feasibility for the most common SIPP uses and those which make use of the unique characteristics of SIPP and could not be obtained from other datasets.

Recommendation 9-4: Given the differences in user needs and approaches, the Census Bureau should offer multiple tiers of access, with approaches for confidentiality protection applied.

Noting that the SIPP public-use files are difficult for many to use, the panel also considered whether a new tool—a table generator—should be considered. Such a tool ultimately would not meet the needs of most current SIPP data users, who tend to perform complex analyses. However, if needed measures are dropped from the public-use file for disclosure avoidance purposes, a table generator may provide a useful tool for preliminary planning and for preparing research proposals, providing easier access than the SODA. Additionally, given the complexity of working with the SIPP public-use file, the availability of a table generator might create a new audience for the data and thus increase the total number of SIPP users. The Census Bureau's past experience with the American Factfinder system and an abandoned longitudinal tool for SIPP, along with its current experience with query systems for the Decennial Census and American Community Survey, may be helpful in creating a table generator for SIPP. However, the complex structure of the SIPP data, combined with the highly customized programming often performed by SIPP data users, make the development of a table generator more difficult than would normally be the case.

Recommendation 7-1: The Census Bureau should assess the demand for an initial flexible table generator as a simple tool, with a specific purpose, that is designed to gauge value and provide direction for further development, and proceed with the development of one if there appears to be sufficient demand.

Copyright National Academy of Sciences. All rights reserved.

CURRENT SIPP APPROACHES TO DISCLOSURE AVOIDANCE

SIPP has two primary approaches to protecting respondents' confidentiality. The first, labeled as traditional or classical disclosure limitation, involves modifying the data by suppressing some variables; top-coding, bottom-coding, or recoding data to combine relatively unique responses with responses that are more common; and adding noise to the data (in SIPP's case by modifying the birth month). (Another traditional approach is to swap particular data values among selected respondents; however, swapping is not currently used by SIPP.) The second approach to protecting confidentiality in SIPP is to control researcher's access to the data by offering the original (unmodified) data in a controlled environment (FSRDCs) in which both the researchers and the research topic must be approved, and in which analytical results cannot be removed from the secure environment without approval by the Census Bureau's Disclosure Review Board. A third approach, offering online access to a synthetic data file, is now used as a means to allow the blending of administrative data with selected data from SIPP, but this could be generalized as a tool for disclosure avoidance for the SIPP data as well.

In the view of the panel, this largely dichotomous choice between a public-use file and FSRDCs is inadequate, for the following reasons.

- *The public-use file already has important limitations.* For example, it suppresses data on country of birth and immigration status, and it seems likely that more extensive re-identification studies would find disclosure risks both from highly granular data (such as state identifiers) and the provision of data on changes over time and on household combinations; addressing these risks will limit the value of the public-use data further. Given current trends, the amount of data available from external sources that might be matched with SIPP data can be expected to increase, creating new disclosure risks that are not present today, and resulting in new limitations on the public-use file.
- *FSRDCs are difficult and costly to use*. For that reason, they would not be a practical option for many researchers. Further, if a large number of current public-use file users were to shift to FSRDCs, this would be likely to create a burden on the Census Bureau in terms of reviewing the data for public release, along with possibly requiring changes to the infrastructure for making the data available.

Copyright National Academy of Sciences. All rights reserved.

Protecting SIPP data from disclosure while supporting current levels of research will require creating one or more alternative approaches that provide more access to the data than will be available in the public-use file but are less onerous than FSRDCs (which still will be important for protecting the most sensitive data and for supporting the blending of data from other sources).

The remaining sections of this chapter provide a roadmap for future actions.

TIERS OF ACCESS AS A TOOL FOR PROTECTING PRIVACY

In the panel's view, a solution can be found by providing additional tiers of access. There are a variety of models for how data are being safely released in restricted-use environments without requiring FSRDCs. The Census Bureau's SIPP synthetic beta file is one example, though its primary purpose is to blend in administrative data, not to expand the data available from SIPP (which in fact are highly limited). The National Center for Education Statistics, Bureau of Labor Statistics, U.S. Department of Agriculture, National Center for Health Statistics, and National Center for Science and Engineering Statistics all are releasing at least some restricted-use data in ways that are less onerous than the use of FSRDCs. Their approaches require signed user agreements on how the data will be used; they also establish restrictions on where and how the data may be stored and accessed and on what data may be released. Tiered access is a way of both limiting access (not giving researchers more data than they need) and promoting access (providing data that are too disclosive to be in a public-use file, but in a more controlled environment). Creating two additional tiers would be valuable.

- *A remote table generator would provide users with ways of producing their own tables.* At the same time, it would limit what data could be output (e.g., cells with too few observations) and add noise to protect confidentiality. Most current users would probably not be completely satisfied with a remote table generator because they often engage in complex analyses that such a generator would not support. However, they may still find a table generator useful for preparing research proposals. Additionally, the availability of a remote table generator may result in new SIPP data users among those who find the current files too difficult to work with. Potentially, the table generator could evolve into a remote analysis system with more capabilities, such as regression analysis.

Copyright National Academy of Sciences. All rights reserved.

- *Secure online data access (SODA) would provide a way of making data available in a controlled environment*. This would be useful for those situations in which data are not available (or not at sufficient granularity) in the public-use file. It could be designed to prevent the printing of output, requiring that data users first go through a disclosure review process. It also typically involves user agreements in which users accept limitations on what data can be released and acknowledge penalties for the improper release of data. Note that this is different from the online access that the Census Bureau currently provides through FSRDCs. As compared with FSRDCs, SODA would offer simpler, faster, and less costly processes for applying for access, getting approvals for access, and disclosure review for public data release.

As discussed in Chapter 1, the creation of a secure remote access that does not require visiting an FSRDC was recommended in a previous report: Recommendation 3-8 in 2009 (National Research Council, 2009).

Recommendation 9-4: Given the differences in user needs and approaches, the Census Bureau should offer several tiers of access, with different approaches for confidentiality protection applied.

Recommendation 4-1: The Census Bureau should modernize and strengthen disclosure limitation methods applied to SIPP to satisfy standards on both data utility and disclosure risk while moving toward methodological transparency. In particular, the Census Bureau should continually assess how the science and technology for privacy are advancing and consider implementing new tools when advantageous.

Recommendation 4-3: The Census Bureau should evaluate using additional modes of access to provide products with more disclosure protection than public-use files offer but less onerous access requirements than FSRDCs impose.

Recommendation 5-1: If disclosure risk assessment studies find that the current public-use file does not provide adequate disclosure avoidance, the Census Bureau should consider secure online data access as a mode likely to support both access and security.

Copyright National Academy of Sciences. All rights reserved.

In the short term, SODA seems to offer the most practical alternative; it would require a change in the infrastructure but not special software or database development. Further, in the current proposed budget for FY 2024 the Census Bureau is already proposing to "implement a robust virtual access program,"[68] and there also are outside organizations that could help to provide the infrastructure needed, having the experience in doing so for other federal agencies.

ADDITIONAL TOOLS TO SUPPORT DISCLOSURE AVOIDANCE

This section addresses four specific tools/approaches for disclosure avoidance: synthetic data, differential privacy, SAE, and disclosure review procedures.

Synthetic Data

Synthetic data consist of surrogate data that have been created through statistical modelling to preserve the statistical properties of the original data. When synthetic data are well-designed, users can perform analyses on the synthetic data in the same way they had been using the original data, making valid inferences while protecting confidentiality. Synthetic data are also a useful tool for blending other data with SIPP, providing a means of benefitting from external data while keeping the external data separate. Data synthesis is a valuable tool with considerable potential to make SIPP data both accessible and safe from disclosure. However, important limitations should be noted:

- There is no off-the-shelf software currently available for creating a synthetic data file with the size and complexity of the SIPP data. Thus, it is only practical to synthesize a select number of variables—namely those that are most sensitive or that are based on blended data.
- It is possible that the synthetic models will simulate the data so accurately that the resulting data may contain disclosure risks. Thus, a synthetic data file should itself be examined through tools such as re-identification studies.
- The validity of inferences from synthetic data depends on the modelling process used to create these data, and how well that process supports a particular type of analysis.

[68] https://www.commerce.gov/sites/default/files/2023-03/Census-FY2024-Congressional-Budget-Submission.pdf

Copyright National Academy of Sciences. All rights reserved.

Thus, data users need guidance on when to use synthetic data and may need ways to validate their findings based on the original data.

Recommendation 6-2: The Census Bureau should develop a systematic protocol and a set of metrics for assessing and measuring the validity of inferences constructed from synthetic data.

Recommendation 6-3: To support the use of synthetic data, the Census Bureau should develop a routine automated system to perform user-submitted analyses on the internal file and send the results back to the user after they have been treated for disclosure avoidance.

Recommendation 6-4: The Census Bureau should consider creating partially synthetic datasets to allow analyses of sensitive or disclosive SIPP survey data variables that are not on or may need to be dropped from the public-use file. These synthetic variables might be added to the public-use file in place of the original variables.

Differential Privacy

Differential privacy is both a metric for measuring disclosure risk and a framework for limiting disclosure risk under this metric. With differential privacy, disclosure limitation is achieved by injecting noise into statistical computations in a way that obscures the effect of each individual data point while still allowing statistical signals about the population to come through. While differential privacy has a rich and rapidly advancing academic literature and a number of large-scale deployments in industry and government (most notably for disclosure avoidance in the 2020 U.S. Decennial Census), the statistical tools have not yet advanced to the point that they could create a differentially private synthetic dataset for the size and complexity of SIPP. In the short run, differential privacy may be most useful as a component of a table generator; in the long term, one might work toward the creation of a differentially private public-use data file.

Recommendation 4-1: The Census Bureau should modernize and strengthen disclosure limitation methods applied to SIPP to satisfy standards on both data utility and disclosure risk while moving toward methodological transparency. In particular,

Copyright National Academy of Sciences. All rights reserved.

the Census Bureau should continually assess how the science and technology for privacy are advancing and consider implementing new tools when advantageous.

Small Area Estimation

Geography-based variables present a particular challenge because of their high granularity and thus potential disclosiveness, combined with their substantial analytic value for policy studies such as through comparisons of policies across states. SAE is a model-based approach to combining survey data with other information to create more accurate, and less individually identifiable, estimates for small areas (typically small geographic areas.) Chapter 8 presents three levels of data availability depending on what users may need from such data.

Recommendation 8-1: The Census Bureau should continue to pursue the development of a small area estimation program to meet the needs of SIPP users for geography-based analysis that preserves confidentiality and limits disclosure risk.

Disclosure Review Board

The Census Bureau has a Disclosure Review Board that reviews all data products based on confidential survey data to make sure data are not released until they satisfy all agency-confidentiality requirements. This is the final step in confidentiality protection. This can be a labor-intensive process, and the efforts to continue to protect SIPP data as confidentiality research progresses, developing new data products and new data-protection approaches, are likely to make this effort more burdensome and time consuming. There are potential ways for limiting this burden:

- The Census Bureau might design a table generator that automatically implements disclosure protections to protect the data so that they can be released.
- The Census Bureau may consider the relatively quick review procedures that have been developed by some external organizations.

The Census Bureau also encounters a burden when seeking to validate the accuracy of analyses based on synthetic data, and that burden has the potential to increase if some synthetic data are added to the public-use file. Again, there are potential ways for limiting this burden:

Copyright National Academy of Sciences. All rights reserved.

- Potentially, the Census Bureau could provide sufficient guidelines on the validity of synthetic data in various situations so that some categories of investigation would not need validation against the original data.
- During the validation process (when used), the Census Bureau may be able to confirm that the synthetic data fall within acceptable parameters, without replacing the synthetic data estimates with estimates from the original data.

TITLE 13 AND ITS REQUIREMENTS FOR DISCLOSURE AVOIDANCE

One ambiguity with regard to creating secure online data access is the role of Title 13 in establishing conditions for access. Here it is important to distinguish between Title 13 and the regulations that are used to implement Title 13. Title 13 itself is relatively ambiguous, prohibiting the release of identifying information and requiring that those who are given access to confidential information must be sworn to observe the limitations it specifies. The regulations that interpret Title 13 impose far stricter requirements than seem to be required by Title 13 itself, and greater than those often used by federal agencies for protecting confidentiality.

Further, the Census Bureau now faces a different context from the period when the regulations interpreting Title 13 were first formulated, with both the Information Quality Act and the Evidence Act making access to data an important priority. This new context is relevant when interpreting Title 13. We believe that it is possible to create tiered data access that is consistent with Title 13 but that accounts for differing levels of sensitivity for different data and that promotes evidence building as an appropriate goal for obtaining access. Such a reinterpretation might simplify what criteria both people and projects must meet to obtain access, what procedures are used to examine whether people and projects meet those criteria, and what levels of data protection are required depending on the specific data involved.

Recommendation 9-5: The Census Bureau should modernize its interpretation of Title 13 consistent with changes in technology, policy guidance, and legislation (i.e., the Evidence Act and the Information Quality Act). Doing this should enable the development and operationalization of tiered data access in which (1) the types and levels of protection and access vary across tiers; (2) the requirement of benefit is redefined to include evidence building, the productive use of Census Bureau-

Copyright National Academy of Sciences. All rights reserved.

developed data, and other statistical purposes; and (3) the types of individuals who are eligible to access the data in these tiers are broadened.

THE BENEFITS OF ENHANCED COMMUNICATION

As discussed in Chapter 2, the SIPP data files are complex files to work with, and data users often have expressed difficulty in working with them. Data users' needs for clear communication have been noted frequently in previous CNSTAT reports, as listed in Chapter 1. The Census Bureau's general communications concerning SIPP are outside of the charge given to the panel, and thus are not discussed here. However, the use of disclosure avoidance techniques magnifies users' needs for clear and comprehensive training for the following reasons.

First, both data users and the general public have shown little awareness of the Census Bureau's efforts to protect confidentiality. Some users of the decennial census data appear not to realize that the Census Bureau has long been altering data to protect confidentiality, leading to complaints after the adoption of differential privacy that would have also applied to earlier decennial censuses as well (Boyd & Sarathy, 2022). A recent survey of economists shows that researchers often are unfamiliar with current approaches; for example, 54 percent had never heard of the concept of differential privacy (Williams et al., 2023). The general public is unlikely to be interested in the details of disclosure avoidance, but potential respondents will wish to know that the responses they give will be kept confidential. To avoid misperceptions of disclosure avoidance techniques used by SIPP, it is important to document both past and new procedures.

Conclusion 10-1: Many current users do not understand privacy protection procedures and how they affect the reliability of their statistical output or the methodological procedures necessary to make valid inference.

Second, to the extent that new approaches are taken, data users will need information on how to analyze and interpret the data.[69] Differential privacy, if used, requires different

[69]The need for users to be able to conduct analyses which recognize and incorporate the statistical disclosure limitation (SDL) methods embedded in the data applies to traditional SDL methods as well. It is not clear whether users fully recognize the methods that the Census Bureau currently uses.

Copyright National Academy of Sciences. All rights reserved.

approaches for computing standard errors, and data users will need both training and tools to perform the appropriate calculations. Most disclosure avoidance methods will require users to use new software to obtain unbiased and reliable estimates. Depending on the complexity of the disclosure avoidance method, this software could be difficult to use and not readily available. Most of the existing software "packages," which are used by the vast majority of users of SIPP, will no longer be usable, yet it is unrealistic to think that all users will start programming new statistical procedures in Fortran, R, Python, Matlab, or whatever their favorite programming language is. In some very simple cases, where noise infusion is minimal or where synthetic data are produced using simple models, software is easily available (e.g., to use multiple imputation or correction for errors-in-variable). However, for the vast majority of methods that are likely to be used, the necessary programming will be much more complex, and data users will need to be trained in the use of these methods.

As discussed in Chapter 6, users of synthetic data need to have communicated to them the nature of the statistical generating model, including which relationships the model captures and which it cannot, to enhance their understanding of what relationships are likely to be accurate and reliable and which are not. Also, users will need to be trained on appropriate methods for analyzing synthetic data to construct valid confidence intervals.

Third, if different levels of data are offered through different tiers of access, then data users need information to determine which tiers will be most useful for their analysis. For example, the accuracy of estimates based on synthetic data depends on how well the models used for synthesizing the data align with data users' planned analyses. Thus, data users need information about the models that were used and the implications for their analyses.

Communication should not be in only one direction, that is, only from the Census Bureau to data users. When the panel began its work, Census had no regular process in place for determining the needs and satisfaction of SIPP data users other than providing an email contact. For that reason, the panel conducted its own data collection. Since that time, SIPP conducted a two-day data conference in February 2023, part of which was devoted to allowing time for SIPP data users to ask questions and express their needs. The panel commends Census for holding the conference, and particularly for allowing for user dialogue.

Communication is also important at all stages of data preparation and release. Understanding data users' needs will help the Census Bureau to know how to prepare access

Copyright National Academy of Sciences. All rights reserved.

modes and tools, and how best to prepare documentation that will support data users in making use of protected data. In turn, when SIPP data users are offered transparency and consultation regarding disclosure avoidance approaches, they will be better able to plan and conduct their research while mitigating tensions over disclosure avoidance approaches.

The importance of communication has been a longstanding concern, as noted in previous National Academies reports. See recommendations 6-4 (National Research Council, 1993), 9-3 (NASEM, 2018) and 9-5 (NASEM, 2018).

> **Recommendation 3-2: The Census Bureau should find ways to communicate to researchers and the general public the findings of its assessments of re-identification attacks and disclosure risk so that data users will understand the need for disclosure avoidance techniques.**
>
> **Recommendation 10-1: The SIPP program/Census /Bureau should work to enhance communications and engagement with SIPP users. The Census Bureau should maintain regular contact with SIPP users through user groups, periodic surveys, and regular workshops. The Census Bureau should also solicit user input and use it in its ongoing planning for the survey.**

One way of systematically seeking user feedback is through the creation of a technical advisory group. For example, the Census Bureau might request that the American Statistical Association create a committee to partner with it; this would be a simpler and less cumbersome approach than creating a group that would fall under the auspices of the Federal Advisory Committee Act (FACA). Given that privacy-preserving techniques are rapidly evolving, such a group could also be helpful in providing guidance on new potential approaches or risks, and in explaining the need for disclosure avoidance to data users.

While the Census Bureau currently has no active SIPP technical users group with which it regularly consults, at various times it has engaged with stakeholders about the quality and future of SIPP. The 1993 CNSTAT report documents a few technical advisory committees, the OMB interagency SIPP committee, SSRC committee, ASA/SRM Working Group on Technical Aspects of SIPP, and the Association of Public Data Users (APDU) SIPP Users group (NASEM, 2018). During the late 1990s and early 2000s, Census engaged SIPP users to draft SIPP Quality

Copyright National Academy of Sciences. All rights reserved.

Profiles and User Guides.[70] During the re-engineering of SIPP in mid-2000s, Census continually engaged with stakeholders.[71] In conjunction with the release of the 2009 CNSTAT SIPP report, Census reconstituted the American Statistical Association's Survey Research Methods (ASA/SRM) Working Group on Technical Aspects of the Survey of Income and Program Participation (SIPP), which provided it with technical assistance on the re-engineering until the new 2014 CNSTAT panel was created. In recent years, Census has sponsored user trainings, workshops and conferences.[72]

Recommendation 10-2: The SIPP program/Census Bureau should consider creating a standing technical consulting group that can provide guidance, accountability, and justification on disclosure avoidance approaches while considering the needs of data users.

Finally, the Census Bureau is not the only federal agency dealing with privacy issues. Census sometimes has taken the lead in developing new approaches, as when it provided a synthetic version of the SIPP data and when it adopted differential privacy for the 2020 decennial census, but establishing a close working relationship with other federal agencies could help in creating a systematic body of research on disclosure avoidance, as well as possibly creating a consensus on best approaches. The National Secure Data Service (NSDS) and America's Datahub Consortium could be helpful in this regard. While the Census Bureau has the greatest funding of all the principal federal statistical agencies, it is far from alone. Other agencies are also dealing with statistical disclosure issues, the sharing of data, and the sponsorship of research in disclosure avoidance. Collaborating with these other agencies could help in setting federal statistical standards, creating mechanisms for data access, and sponsoring research on disclosure avoidance.

[70]https://www.census.gov/library/working-papers/1998/demo/SEHSD-WP1998-11.html

[71]https://www.census.gov/programs-surveys/sipp/about/re-engineered-sipp.html

[72]For the 2023 conference, see https://www.census.gov/programs-surveys/sipp/events/2023-sipp-conference.html

Copyright National Academy of Sciences. All rights reserved.

LINKING SIPP DATA WITH ALTERNATIVE DATA SOURCES

The utility of SIPP as a program may be enhanced by linking the survey data with administrative data sources. While the Census Bureau has some freedom in how to interpret Title 13, the issue is more complicated when data from other agencies are considered. The restrictions placed on IRS data are especially tight, and the Census Bureau must respect the limitations placed on it by the IRS and other agencies. Thus, data merges have been restricted to data access through FSRDCs, the creation and use of the SIPP synthetic beta, and internal use of such databases for SIPP error checking. Nevertheless, examples where researchers might wish to link SIPP with alternative data sources include these:

- To improve data quality, such as when administrative data are used to correct respondent errors or omissions
- To supplement SIPP with data that otherwise would not be available on SIPP, and
- Conversely, to supplement external data with data available on SIPP.

Depending on what linking data are available and what requirements are placed on the data by the organizations providing them, the linking might involve either directly inserting data taken from external sources or using the external data to impute data or correct for response errors (e.g., using statistical models). Currently, to support the linking of SIPP (or other survey) data with administrative data, the Census Bureau uses the name, date of birth, sex, and address to find a Social Security number (SSN), which it then replaces with a protected identification key (PIK).[73] This process produces a match rate of approximately 90% with SIPP data, while administrative record variables are imputed when the PIK is missing.

As suggested by the examples in Chapter 2, there are many potential uses of linked data, and there are many potential sources for linking. Linkages are the most powerful and privacy concerns are the greatest when there are direct linkages. Since SIPP uses statistical sampling and generally does not intentionally overlap its samples with other surveys, linkages work best when the external data comprise a census rather than a sample. Some examples include tax data, Social Security, unemployment insurance, Medicare, student loans, participation in SNAP (available for selected states), and other databases on state-provided assistance.

[73]B. Gurrentz, Model-based imputation & administrative records in SIPP processing (presentation to panel on June 30, 2022).

Copyright National Academy of Sciences. All rights reserved.

If data users wish to link the data themselves using the PIK, currently they would need to work through an internal Census Bureau project or a joint statistical project and then pay for the linking process. Potentially, depending on the nature of the data sharing agreement for accessing the data to be linked, the Census Bureau might create a way in which data linkages could also be provided through secure online data access. Sometimes, however, there may not be adequate data for linking individuals across databases, the providers of the external data may set conditions on how the data may be used, or there may be other privacy concerns, requiring the use of imputed or synthetic data rather than inserting administrative data directly.

However, a different approach, which would enhance SIPP usability and perhaps increase the number of SIPP users, is for the Census Bureau (or its designee) to perform the linkages and provide a synthetic data file. This is what the Census Bureau currently does with the SIPP synthetic beta (SSB), and the approach might be generalized to make more data accessible. Such data could even be included in the public-use file, or alternatively be provided through secure online data access. Given the time and resources required to create such data, it would be important to first prioritize which variables would be most useful to researchers and policymakers, and only then to synthesize just those variables and augment the SIPP survey data. In generating these synthetic variables, it is important to maintain some key variables in the SIPP survey data. The importance of linking SIPP with other data has also been noted in previous National Academies reports, as discussed in Chapter 1. See recommendations 4 (National Research Council, 1989), 3-5 and 3-7 (National Research Council, 2009), and 5-2 (NASEM, 2018).

FUTURE RESEARCH TO ADDRESS CURRENT KNOWLEDGE GAPS

Disclosure avoidance is a complex and rapidly changing field, as illustrated by the ever-expanding scientific literature on differential privacy. The Census Bureau is already conducting and funding research on disclosure avoidance, and continued research is vital. In that respect, the Census Bureau should work with researchers outside the federal government to get their perspectives and benefit from their capabilities.

Recommendation 6-1: The Census Bureau should support research that develops new approaches and techniques for generating synthetic data and assessing data utility

Copyright National Academy of Sciences. All rights reserved.

and disclosure risk, with or without differential privacy or other privacy guarantees, and should fund work that creates and updates tools and systems that evaluate synthetic data, disclosure risk, and usability and collect data users' feedback.

Recommendation 10-2: The SIPP program/Census Bureau should consider creating a standing technical consulting group that can provide guidance, accountability, and justification on disclosure avoidance approaches, while considering the needs of data users.

Recommendation 3-3: The Census Bureau should find ways to partner with and involve external researchers/experts in its risk assessment research program. Such involvement will expand the bureau's capacity and will allow it to tap into state-of-the-art developments in the area of disclosure risk assessment.

CENSUS BUREAU RESOURCES FOR ADDRESSING DISCLOSURE AVOIDANCE

The panel recognizes that applying the full panoply of tools discussed in this report will take time, resources, and research. There is currently no off-the-shelf tool for creating either a satisfactory synthetic data file or a SIPP public-use file with differential privacy, especially given the large number of variables and the presence of longitudinal data. The Census Bureau has already experimented with a table generator for longitudinal SIPP data and decided it was too difficult to maintain. SIPP data will be available in the Microdata Access Tool soon, according to the Census Bureau, though it will only include pre-2014 SIPP data.[74] Creating a full implementation of such tools at the current time would at best greatly delay the release of the SIPP data and, at worst, may not yet be doable. The challenges do not stop with the creation of a database, either, but include how it is maintained. The validity of inferences from synthetic data varies depending on what models were used to generate the synthetic data and how well the data support a particular planned data analysis, meaning that validation against the original data often will be needed. It is to be hoped that Census will be able to publish data that will indicate which kinds of analyses are most reliable and which are most problematic, helping to reduce the number of times validation is needed. Still, validation of synthetic data findings already is

[74]Email from Holly Fee at the U.S. Census Bureau, July 31, 2023.

Copyright National Academy of Sciences. All rights reserved.

burdensome on the Census Bureau staff, and it would not be practical for Census staff if the number of users and the types of analyses they conduct were to increase due to the number of synthetic data users increasing greatly. The Census Bureau might find that the best option is to partner with another organization, one that has the capacity for supporting SODA.

Conclusion 10-2: Census might consider using outside contractors or grantees to provide some of the infrastructure required for developing and sustaining a table generator/remote analysis system, producing synthetic data, providing validation, providing secure online data access, and/or performing disclosure review.

A ROADMAP FOR ADDRESSING DISCLOSURE AVOIDANCE IN SIPP

Conducting the research and developing the tools discussed here will take substantial time, while the timely release of SIPP data remains important. Therefore, the panel recommends phasing in additional protections and access over time. The Census Bureau should continue offering a public-use file, omitting items that are determined to have too high a disclosure risk, or replacing them with synthetic estimates. While a long-term solution may be the production of a fully synthetic public-use file, in the short run the most practical solution is probably to produce a somewhat less comprehensive public-use file, combined with a way to provide restricted access to additional SIPP data without imposing all the requirements currently required for FSRDC access.

The Census Bureau might also explore even shorter-term changes as a way of temporizing, given that it will take time to implement additional tiers of access, and given the potential for disclosure risks identified in this report. One approach is to strengthen traditional statistical disclosure limitation methods through more coarsening of the data, additional top codes, additional geography suppression, and similar methods. Based on the Census Bureau's re-identification exercise and on our analysis of data granularity, age and geography might pose the greatest threats; these are the two most readily available measures on external datasets with high granularity. For example, age might be recoded into a largely categorical measure, while still retaining key age thresholds for program eligibility, and state identifiers might be replaced with measures of geographic region. However, these increased protections come at the cost of reduced data utility, and any adoption of increased statistical disclosure limitation methods should seek to

Copyright National Academy of Sciences. All rights reserved.

minimize the loss of that utility. Another option that has been adopted in a number of European countries is a registration process for those who download the public-use file, requiring users to give identifying information and state the intended purpose of the data analysis. Alternatively or in addition, users of public-use SIPP files could be required to sign a user agreement that requires them to agree to specific uses of the data and to not disclose any data that reasonably seems confidential. User agreements are common for restricted-use files, but they can also be applied to public-use files; for example, note how the National Center for Education Statistics imposes some conditions on the use of public-use files, as discussed in Chapter 4. These options do not represent a permanent solution, and it would still be possible for a sophisticated intruder to attempt reidentification. Still, they would enhance current protections and make the general user community aware of these issues and sensitive to the problems. It is important that any user agreements or enforcement actions be implemented very carefully so as not to have a chilling effect on privacy-related and other types of research.

Multiple National Academies reports have noted the importance of the timely release of data. Given the current state of research and technology for tools such as synthetic data and differential privacy, the Census Bureau should continue to release public-use files using current technology (which may include a limited number of synthetic variables) and with appropriate protections of confidentiality (which may require moving some data to be available through SODA). See Chapter 1 for recommendations in previous reports: i.e., recommendations 5 (National Research Council, 1989), 3-6 (National Research Council, 2009; also see conclusion 4-7), and 9-1 (NASEM, 2018).

Considering which steps might be taken quickly and which will require more time, following is a roadmap with suggested milestones for implementing the changes recommended in this report.

Milestone 1

- Write a risk assessment plan, including plans on approaches to conduct risk assessments, when, how often, and how to conduct them, and who should conduct them, including support from external researchers/experts.
- Write a communication plan for disclosure risk results, which may include plans for seminars for data users, providing example text for written communication.

Copyright National Academy of Sciences. All rights reserved.

- Assess the likely use of a simple and flexible table generator using the public-use file, and whether it would broaden the user base.
- Begin planning for establishing secure online data access, including an evaluation of what is needed to satisfy Title 13, the Evidence Act, and the Information Quality Act.

Milestone 2

- Implement an on-going risk assessment program for SIPP, one that includes consideration of its longitudinal structure and determine what changes may be needed to the public-use file to protect confidentiality.
- Continue publishing the SIPP public-use file, modified as necessary based on the on-going risk assessment, including partial synthesis of a limited set of variables deemed too disclosive on an experimental basis.
- Depending on the results of the assessment conducted under Milestone 1, prepare the design of a flexible table generator.
- Plan for and begin developing secure online data access (SODA) to provide access to data/synthetic estimates that are omitted in the public-use files.
- Consider how to work with external partners to deal with issues such as risk assessment, hosting SODA, and performing disclosure review.
- Begin developing geography-level model-based estimates and evaluate them using data utility metrics; consider a differentially private framework.
- Investigate multi-tiered modes of access, including the public-use file, synthetic data, SAE, secure online data access, and remote analysis platform.
- Coordinate with other federal agencies, including the Social Security Administration and the Internal Revenue Service, to determine what data can be shared (including synthetic data) and through what means.

Milestone 3

- Continue refining the SIPP public-use file, as necessary, based on the on-going risk assessment. (Note: this should be a continuing activity throughout the release of public-use files.)

Copyright National Academy of Sciences. All rights reserved.

- Implement a flexible table generator that uses the differential privacy framework if sufficient demand for it has been established.
- Explore the feasibility, design, and usefulness of a partial synthetic file of a larger set of variables, including the possible incorporation of formal privacy, and evaluate its accuracy with data utility metrics.
- Implement SODA.

Milestone 4

- Obtain feedback from data users on the experimental partially synthetic dataset if one has been deemed useful.
- Expand the flexible table generator to become a remote analysis platform based on the public-use file, possibly including regression modelling, if work on it in Milestone 3 has continued.

Milestone 5

- Add differential privacy or some other formal privacy framework to the remote analysis platform, infusing noise under the framework and laying the foundation for automated verification and validation servers.

Milestone 6

- Publicly release a partially synthetic file with an automated verification and validation server.
- Given the creation of a verification and validation server, decide whether to adjust the remote analysis platform so the partial synthetic file can be used to provide consistency between data products, or else continue with the formal privacy version.

Copyright National Academy of Sciences. All rights reserved.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In *Proceedings of the 2016 Association for Computing Machinery Special Interest Group on Security, Audit, and Control (ACM SIGSAC) Conference on Computer and Communications Security* (pp. 308-318).

Abowd, J. M. (2021). Declaration of John Abowd, *State of Alabama v. United States Department of Commerce,* Case No. 3:21-CV-211-RAH-ECM-KCN.

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown algorithm. *Harvard Data Science Review* (Special Issue 2). https://doi.org/10.1162/99608f92.529e3cb9

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review* 109(1), 171–202.

Abowd, J., Stinson, M., & Benedetto, G. (2006). *Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project*. U.S. Census Bureau.

Abowd, J. M., & Vilhuber, L. (2008). How protective are synthetic data? In *Proceedings of the Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference*, Istanbul, Turkey, September 24-26, 2008. (pp. 239-246). Berlin & Heidelberg: Springer.

An, D., & Little, R. J. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*) *170*(4), 923–940.

Arnold, C., & Neunhoeffer, M. (2021). Really useful synthetic data: A framework to evaluate the quality of differentially private synthetic data. Presented at the Workshop on Economics of Privacy and Data Labor, 37th International Conference on Machine Learning, Vienna, Austria, 2020. https://doi.org/10.48550/arXiv.2004.07740

Balle, B., Barthe, G., & Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergencies. *Advances in Neural Information Processing Systems 31*. https://papers.nips.cc/paper_files/paper/2018/file/3b5020bb891119b9f5130f1fea9bd773-Paper.pdf

Balle, B., & Wang, Y. X. (2018). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. Presented at the 35th Conference of the International Conference on Machine Learning https://doi.org/10.48550/arXiv.1805.06530

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modelling and analysis for spatial data*. CRC Press.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., & Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the 26th Association for Computing Machinery Symposium on Principles of Database Systems,* 273–282. https://dl.acm.org/doi/10.1145/1265530.1265569

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., de Figueiredo, J. M., Machanavajjhala, A., & DeLong, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *The Annals of Applied Statistics 12*(2), 1124–1156. https://www.jstor.org/stable/26542565

Bart Stykes, J., & Guzzo, K. B. (2019). Multiple-partner fertility: Variation across measurement approaches. *Analytical Family Demography,* Springer International Publishing, 215–239.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*(401), 28–36.

Beimel, A., Brenner, H., Kasiviswanathan, S.P., & Nissim, K. (2014.) Bounds on the sample complexity for private learning and private data release. *Machine Learning 94*, 401–437. https://doi.org/10.1007/s10994-013-5404-1

Copyright National Academy of Sciences. All rights reserved.

Bell, W. R., Basel, W. W., & Maples J. J. (2016). An overview of the U. S. Census Bureau's Small Area Income and Poverty Estimates Program. In M. Pratesi [Ed.], *Analysis of poverty data by small area estimation* (pp. 349–378). New York: Wiley.

Bell, W. R., Datta, G. S., & Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika 100*(1), 189–202.

Benavent, R., & Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Computational Statistics & Data Analysis 94*, 372–390.

Benedetto, G., Linse, K., & Parker, E. (2022). *Improving disclosure avoidance procedures for the Current Population Survey public use file.* Presented to the Federal Economic Statistics Advisory Committee, June 10, 2022. https://apps.bea.gov/fesac/meetings/2022-06-10/Paper-Current-Population-Survey-PUF-Disclosure-Avoidance-Proposal-to%20FESAC-061022.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://www.census.gov/library/working-papers/2018/adrm/SIPP-Synthetic-Beta.html

Benedetto, G., Stinson, M., & Abowd, J. M. (2013). *The creation and use of the SIPP synthetic beta*. U.S. Census Bureau.

Bennett, N., & King, M. D. (2022, April). *Student debt and its co-occurrence with other types of debt.* Presented at the annual meeting of the Population Association of America. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2022/demo/sehsd-wp2022-09.pdf

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review 51*(3), 279–292.

Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: The SuLQ framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '05)* (pp. 128–138). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1065167.1065184

Blum, A., Ligett, K., & Roth, A. (2013). A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM) 60*(2), 1–25.

Boch, S. J., Taylor, D. M., Danielson, M. L. Chisolm, D. J., & Kelleher, K. J. (2020). Home is where the health is: Housing quality and adult health outcomes in the Survey of Income and Program Participation. *Preventive Medicine 132*(105995) https://doi.org/10.1016/j.ypmed.2020.105990

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., Shaw, T., Stapleton, L. M., Wooley, M. E., & Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness 12*(4), 616–647.

Borjas, G. J., & Hilton, L. (1996). Immigration and the welfare state: Immigrant participation in means-tested entitlement programs. *The Quarterly Journal of Economics* 111(2), 575–604.

Bowen, C. M. (2022). *Protecting your privacy in a data-driven world*. Chapman and Hall/CRC.

Bowen, C.M., & Liu, F. (2020). Comparative study of differentially private synthesis methods. *Statistical Science*, *35*(2), 280–307, May 2020. https://doi.org/10.1214/19-STS742.

Bowen, C.M., Liu, F., & Su, B. (2021). Differentially private data release via statistical election to partition sequentially. *METRON* 79, 1–31.

Bowen, C. M., & Snoke, J. (2021). Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *Journal of Privacy and Confidentiality 11*(1), http://dx.doi.org/10.29012/jpc.748

Boyd, D., & Sarathy, J. (2022). Differential perspectives: Epistemic disconnects surrounding the US Census Bureau's use of differential privacy. *Harvard Data Science Review* (Forthcoming).

Copyright National Academy of Sciences. All rights reserved.

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer House-Hold Dynamics. *Annals of Applied Statistics 9*, 1761–1791.

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2016). Multivariate spatio-temporal survey fusion with application to the American Community Survey and Local Area Unemployment Statistics. *Stat*, *5*(1), 224–233 https://onlinelibrary.wiley.com/doi/full/10.1002/sta4.120

Bradley, J. R., Holan, S. H., & Wikle. C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association 115*, 2037–2052.

Bradley, J. R., Wikle, C. K., & Holan, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association 111*(514), 472–487.

Bruckmeier, K., Müller, G., & Riphahn, R. T. (2014). Who misreports welfare receipt in surveys? *Applied Economics Letters 21*(12), 812–816.

Bu, Z., Dong, J., Long, Q., & Su, W. J. (2020). Deep learning with gaussian differential privacy. *Harvard Data Science Review 2(3).* https://doi.org/10.1162/99608f92.cfc5dd25

Bun, M., Drechsler, J., Gaboardi, M., McMillan, A., & Sarathy, J. (2022). Controlling privacy loss in sampling schemes: An analysis of stratified and cluster sampling. 3rd Symposium on Foundations of Responsible Computing (FORC 2022). *Leibniz International Proceedings in Informatics*. Schloss Dagstuhl -- Leibniz-Zentrum https://doi.org/10.4230/LIPIcs.FORC.2022.1

Bun, M., Gaborardi, M., Neunhoeffer, M., & Zhang, W. (2023). Continual release of differentially private synthetic data. https://doi.org/10.48550/arXiv.2306.07884

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology 172*(9), 1070–1076.

Capps, R., Bachmeier, J. D., & Van Hook, J. (2018). Estimating the characteristics of unauthorized immigrants using U.S. census data: Combined sample multiple imputation. *Annals of the American Academy of Political and Social Science 677*(1), 165–179. https://doi.org/10.1177/0002716218767383

Carless, W. 2020. Census launches online after last-minute software switch. The Center for Investigative Reporting, March 11, 2020. https://revealnews.org/article/census-launches-online-tomorrow-after-last-minute-software-switch/

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramèr, F. (2022). *Membership inference attacks from first principles* [Symposium presentation]. 2022 IEEE Symposium on Security and Privacy, San Francisco (pp. 1897–1914). doi: 10.1109/SP46214.2022.9833649

Carlini. N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023) Extracting training data from diffusion models. https://doi.org/10.48550/arXiv.2301.13188

Carlini, N., Tram, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. *30th USENIX Security Symposium* (USENIX Security 21). https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Carman, K. G., Liu, J., & White, C. (2020). Accounting for the burden and redistribution of health care costs: Who uses care and who pays for it. *Health Services Research 55*(2), 224–231.

Carr, M. D., & Weimers, E. E. (2020). *The role of education in long-run earnings inequality and mobility.* Russell Sage Foundation and the Washington Center for Equitable Growth.

Carr, M. D., Wiemers, E. A., & Moffit, R. A. (2023, May 4–5). *Using synthetic data to estimate earnings dynamics: Evidence from the SIPP GSF and SIPP SSB* [Conference session]. NBER conference on data privacy protection and applied research: Methods, approaches, and their consequences, Cambridge, MA.

Copyright National Academy of Sciences. All rights reserved.

Cha, Y. (2010). Reinforcing separate spheres: The effect of spousal overwork on men's and women's employment in dual-earner households. *American Sociological Review 75*(2), 303–329.

Chan, T. H. H., Shi, E., & Song. D. (2011, November). Private and continual release of statistics. *ACM Transactions on Information and System Security* 14(3), Article 26. https://doi.org/10.1145/2043621.2043626

Charest, A. S. (2011). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality 2*(2).

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*. https://jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.pdf

Clemens, J., & Wither, M. (2019). The minimum wage and the Great Recession: Evidence of effects on the employment and income trajectories of low-skilled workers. *Journal of Public Economics 170*, 53–67. https://www.sciencedirect.com/science/article/abs/pii/S0047272719300052

Cohen, A. (2022). Attacks on deidentification's defenses. In *Proceedings of the 31ste USENIX Security Symposium*. USENIX: The Advanced Computing Systems Association. https://www.usenix.org/system/files/sec22-cohen.pdf

Cohen, A., & Nissim, K. (2020). Linear program reconstruction in practice. *Journal of Privacy and Confidentiality* 10(1). https://doi.org/10.29012/jpc.711

Commission on Evidence-Based Policymaking. (2017). *The promise of evidence-based policymaking*. Commission on Evidence-Based Policymaking, Washington, DC. www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Couzin, J. (2008, September 5). Whole-genome data not anonymous, challenging assumptions. *Science 321*(5894), 1278–1278. https://www.science.org/doi/full/10.1126/science.321.5894.1278

Creamer, J., Shrider, E., & Edwards, A. (2020). *Estimated 17.8% of adults ages 25–34 lived in their parents' household last year*. Washington, DC: U.S. Bureau of the Census. https://www.census.gov/library/stories/2020/09/more-young-adults-lived-with-their-parents-in-2019.html

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Crystal, S., Shea, D. G., & Reyes, A. M. (2017). Cumulative advantage, cumulative disadvantage, and evolving patterns of late-life inequality. *Gerontologist 57*(5), 910–920. https://doi.org/10.1093/geront/gnw056

Daily, D. (2022). *Disclosure avoidance protections for the American Community Survey*. Washington, DC: U.S. Bureau of the Census. https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference 6*(1), 73–85. https://www.sciencedirect.com/science/article/pii/0378375882900581

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics 26*(2), 403-413.

Desfontaines, D., & Pejo, B. (2020). SoK: Differential privacies. *Proceedings on privacy enhancing technologies*. https://doi.org/10.3929/ethz-b-000451916

Differential Privacy Team. (2017, December). Learning with privacy at scale. Machine Learning Research, Apple Computer. https://machinelearning.apple.com/research/learning-with-privacy-at-scale

Ding, B., Kulkalmi, J., & Yekhanin, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems 30* (NIPS 2017). https://papers.nips.cc/paper_files/paper/2017/hash/253614bbac999b38b5b60cae531c4969-Abstract.html

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '03).* New York, NY: ACM. Doi: 10.1145/773153.773173

Copyright National Academy of Sciences. All rights reserved.

Domingo-Ferrer, J., Mateo-Sanz, J., & Torra, V. (2001, June). Comparing SDC methods for microdata on the basis of information loss and disclosure risk. Proceedings of ETK-NTTS 2001. Office for Official Publications of the European Communities. https://www.researchgate.net/publication/229034399_Comparing_SDC_Methods_for_Microdata_on_the_Basis_of_Information_Loss_and_Disclosure#fullTextFileContent

Dondero, M., & Altman, C. E. (2020). Immigrant policies as health policies: State immigrant policy climates and health provider visits among U.S. immigrants, *SSM – Population Health* 10, 100559. https://doi.org/10.1016/j.ssmph.2020.100559

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.

Duchi, J. C., Jordan, M. I., & Wainwright. M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association* 113(521), 182–201.

Duncan, G. T., Elliot, M., & Salazar-González, J-J. (2011). *Statistical Confidentiality.* New York: Springer.

Duncan, G. T., & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association 81*(393), 10–18.

Duncan, G. T., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics* 7(2), 207–217.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), (pp. 1–12). Berlin & Heidelberg: Springer.

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. 265–84). Berlin: Springer-Verlag. https://doi.org/10.1007/11681878_14

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2017). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality 7*(3), 17–51.

Dwork, C., Naor, M., Pitassi, T., & Rothblum, G. N. (2010). Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on theory of computing (STOC '10)* (pp. 715–724). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1806689.1806787

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and its Applications*. doi: 10.1146/annurev-statistics-060116-054123

Eggleston, J., & Reeder, L. (2018). Does encouraging record use for financial assets improve data accuracy? Evidence from administrative data. *Public Opinion Quarterly* 82(4), 686–706.

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making 11*(46).

El Emam, K., Mosquera, L., & Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research 22*(11), e23139.

Elamir, E., & Skinner, C. J. (2006). Record-level measures of disclosure risk for survey micro-data. *Journal of Official Statistics 22*, 525–539.

Elliot, M. J., Manning, A. M., & Ford, R. W. (2002). A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 5*(10), 493–509.

Elliot, M., Maning, A. M., Mayes, K., Gurd, J., & Bane, M. (2005). SUDA: A program for detecting special uniques. Presented to the Joint United Nations Economic Commission for Europe (UNECE)/European Commission Statistical Office of the European Communities (EUROSTAT) work session on statistical data confidentiality, November 9-11, 2005. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.44.e.pdf

Erlingsson, U., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*

Copyright National Academy of Sciences. All rights reserved.

(CCS '14) (1054–1067). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2660267.2660348

Eugenio, E. C., & Liu, F. (2021, July). Construction of differentially private empirical distributions from low-order marginals set through solving linear equations with l2 regularization. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 3* (pp. 949-966). Cham: Springer International Publishing.

Favreault, M. (2018). *How Might Earnings Patterns and Interactions Among Certain Provisions in OASDI Solvency Packages Affect Financing and Distributional Goals?* (Working Paper No. 2). Center for Retirement Research at Boston College. https://crr.bc.edu/working-papers-social-security/how-might-earnings-patterns-and-interactions-among-certain-provisions-in-oasdi-solvency-packages-affect-financing-and-distributional-goals/

Fay, R. E. III, & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association 64*(328), 1183–1210.

Fienberg, S. E., & McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, *21*(2), 309–323. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/data-swapping-variations-on-a-theme-by-dalenius-and-reiss.pdf

Fienberg, S. E., & Slavković, A. B. (2005). Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. *Data Mining and Knowledge Discovery 11*, 155–180. https://link.springer.com/article/10.1007/s10618-005-0010-x

Foote, A. D., Machanavajjhala, A., & McKinney, K. (2019). Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (PSEO). *Journal of Privacy and Confidentiality 9*(2). https://doi.org/10.29012/jpc.722

Foulds, J., Geumlek, J., Welling, M., & Chaudhuri, K. (2016). *On the theory and practice of privacy-preserving Bayesian data analysis.* Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence, Association for Uncertainty in Artificial Intelligence. https://doi.org/10.48550/arXiv.1603.07294

Francis, P. (2022). A note on the misinterpretation of the US Census re-identification attack. In Domingo-Ferrer, J., and Laurent, M. (Eds.), *Privacy in Statistical Databases*. PSD 2022. Lecture Notes in Computer Science, vol. 13463. Springer, Cham. https://doi.org/10.1007/978-3-031-13945-1_21

Fry, R., Passel, J. S., & Cohn, D. (2020). *A Majority of Young Adults in the U.S. Live with Their Parents for the First Time Since the Great Depression*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/09/04/a-majority-of-young-adults-in-the-u-s-live-with-their-parents-for-the-first-time-since-the-great-depression/

Garfinkel, S., Abowd, J. & Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM 62*(3), 46–53. doi: 10.1145/3287287.

Gelman, A., & Little, T. C. (1997). Post-stratification into many categories using hierarchical logistic regression. *Survey Methodology 23*, 127.

Gibson-Davis, C. (2011). Mothers but not wives: The increasing lag between nonmarital births and marriage. *Journal of Marriage and Family 73*(1), 264–278.

Gittleman, M., Klee, M. A., & Kleiner, M. M. (2018). Analyzing the labor market outcomes of occupational licensing. *Industrial Relations 57*(1), 57–100. https://onlinelibrary.wiley.com/doi/10.1111/irel.12200

Gong, R., & Meng, X-L. (2020). Congenial differential privacy under mandated disclosure. *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, 59–70.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Bing Xu, B., Warde-Farley, D., Ozair, S., Aaron Courville, A., & Yoshua Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems 27*. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Gumber, C., & Sullivan, B. (2022, July). Occupation, earnings, and job characteristics. *Current Population Reports.* https://www.census.gov/content/dam/Census/library/publications/2022/demo/p70-178.pdf

Copyright National Academy of Sciences. All rights reserved.

Hall, M., & Greenman, E. (2015). The occupational cost of being illegal in the United States: Legal status, job hazards, and compensating differentials. *International Migration Review 49*(2), 406–442.

Hall, R., Rinaldo, A., & Wasserman, L. (2012). Differential privacy for functions and functional data. *Journal of Machine Learning Research* 14(1). https://dl.acm.org/doi/pdf/10.5555/2567709.2502603

Hardt, M., Ligett, K., & McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 25.

Hardt, M., & Rothblum, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. *2010 IEEE 51st Annual Symposium on Foundations of Computer Science.* Las Vegas, NV: IEEE. doi:10.1109/FOCS.2010.85.

Hawes, M. (2021, May). *Understanding the 20202 Census Disclosure Avoidance System: Simulated reconstruction-abetted re-identification attack on the 2010 Census. Research and Methodology Directorate, Census Bureau.* https://www2.census.gov/about/training-workshops/2021/2021-05-07-das-presentation.pdf.

Hay, M., Rastogi, V., Miklau, G., & Suciu, D. (2009). Boosting the accuracy of differentially-private histograms through consistency. *Proceedings of the VLDB Endowment 3(1)*. https://doi.org/10.48550/arXiv.0904.0942

Heikkila, M., Jalko, J., Dikmen, O., & Honkela, A. (2019). Differentially private Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems 32*, 4113–4123.

Hidiroglou, M. A., & You, Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodology 42*(1), 41–61.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems 33*, 6840-6851.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoSGenet 4*(8), e1000167. doi: 10.1371/journal.pgen.1000167

Hotz, V. J., Bollinger, C, R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, D., & Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences 119*(31), e2104906119.

Hryshko, D., John, C., & McCue, K. (2017). Trends in earnings inequality and earnings instability among U.S. couples: How important is assortative matching? *Labour Economics 48* 168–182. https://doi.org/10.1016/j.labeco.2017.08.006

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., & de Wolf, P. D. (2012). *Statistical Disclosure Control*. Wiley.

Irvin, C. & Czajka, J. (2010). *Simulation of Medicaid and SCHIP Eligibility: Implications of Findings from 10 States - Final Report*. Mathematica Policy Research, Inc. https://aspe.hhs.gov/reports/simulation-medicaid-schip-eligibility-implications-findings-10-states-final-report

Jain, P., Raskhodnikova, S., Sivakumar, S., & Smith, A. (2022). The price of differential privacy under continual observation. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2112.00828

Jälkö, J., Dikmen, O., & Honkela, A. (2016). Differentially private variational inference for non-conjugate models. https://doi.org/10.48550/arXiv.1610.08749

Jälkö, J., Prediger, L., Honkela, A., & Kaski, S. (2022). DVPIm: Differentially private variational inference improved. https://doi.org/10.48550/arXiv.2210.15961

Janicki, R., Holan, S. H., Irimata, K. M., Livsey, J., & Raim, A. (2023). Spatial change of support models for differentially private decennial census counts of persons by detailed race and ethnicity. *Journal of Statistical Theory and Practice 17*(2), 31.

Copyright National Academy of Sciences. All rights reserved.

Janicki, R., Raim, A. M., Holan, S. H., & Maples, J. J. (2022). Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations. *The Annals of Applied Statistics 16*(1), 144–168.

Jarmin, R. S. (2021). *Disclosure avoidance for the 2020 Census: An introduction*. Washington, DC: U.S. Census Bureau. https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf

Jordon, J., Yoon, J., & Van Der Schaar, M. (2019, May). *PATE-GAN: Generating synthetic data with differential privacy guarantees.* Presentation at International Conference on Learning Representations, May 6-9, New Orleans.

Kamath, G., & Ullman, J. (2020). *A Primer on Private Statistics*. https://doi.org/10.48550/arXiv.2005.00010

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), 224–232.

Karwa, V., Krivitsky, P. N., & Slavković, A. B. (2017). Sharing social network data: Differentially private estimation of exponential family random-graph models. *Journal of the Royal Statistical Society, Series C (Applied Statistics*), 481–500.

Kasiviswanathan, S. K., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing 40*(3). https://doi.org/10.1137/090756090

Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., & Zhuravlev, P. (2022). *Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census*. U.S. Census Bureau Working Paper No. CED-WP-2022-004. Washington, DC: U.S. Census Bureau.

Kim, C., Tamborini, C. R., & Sakamoto, A. (2018). The sources of life chances: Does education, class category, occupation, or short-term earnings predict 20-year long-term earnings? *Sociological Science 5*, 206–233. https://sociologicalscience.com/articles-v5-9-206/

King, M., & Giefer, K. G. (2021, June). *Most children receiving SNAP get at least one other social safety net benefit*. U.S. Census Bureau. https://www.census.gov/library/stories/2021/06/most-children-receiving-snap-get-at-least-one-other-social-safety-net-benefit.html

Kingma, D. P., & Welling, M. (2022). Auto-encoding variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

Klein, M., Moura, R., & Sinha, B. (2019). *Multivariate Normal Inference Based on Singly Imputed Synthetic Data Under Plug-In Sampling*. Research Report Series (Statistics #2019-06), U.S. Census Bureau, Washington, D.C. https://www.census.gov/content/dam/Census/library/working-papers/2019/adrm/RRS2019-06.pdf

Klein, M., & Sinha, B. (2015). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models. *Sankhya B,* 77, 293–311. https://doi.org/10.1007/s13571-015-0100-8

Krenzke, T., Li, J., & Li, L. (2014). An evaluation of the impact of missing data on disclosure risk measures. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 548–557). American Statistical Association. http://www.asasrms.org/Proceedings/y2014/files/311082_86754.pdf

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* 9, 313–313.

Lee, J., Kim, M., Jeong, Y., & Ro, Y. (2022, June). Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence 36*(7), 7345–7353.

Leftin, J., Smith, J., Cnnyngham, K., & Trippe, C. (2014). *Creation of the 2011 MATH SIPP+ Microsimulation Model and Database*. Mathematica Technical Working Paper, Mathematica, February 28, 2014. https://www.mathematica.org/publications/creation-of-the-2011-math-sipp-microsimulation-model-and-database

Li, B., Chen, C., Liu, H., & Carin, L. (2019). On connecting stochastic gradient MCMC and differential privacy. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 89:557-566. Presented in Naha, Okinawa, Japan, April 16-18, 2019.

Copyright National Academy of Sciences. All rights reserved.

Li, J., Li, L., Krenzke, T., & Chang, W. Y. (2021). An approach to estimate the re-identification risk in longitudinal survey microdata. Presented at Joint Statistical Meetings 2021, *American Statistical Association,* virtual conference, Aug. 8, 2021.

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

Little, R. J., Liu, F., & Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman & X. Meng (Eds.), *Applied Bayesian Modelling and Causal Inference From Incomplete-Data Perspectives* (pp. 141–152). Wiley. http://dx.doi.org/10.1002/0470090456.ch13

Liu, F. (2022). Model-based differentially private data synthesis and statistical inference in multiply synthetic differentially private data. *Transactions on Data Privacy 15*(3), 141–175.

Liu F., & Zhao, X. (2022, November 8). Disclosure risk from homogeneity attack in differentially privately sanitized frequency distribution. *IEEE Transactions on Dependable and Secure Computing.* doi: 10.1109/TDSC.2022.3220592

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293–312.

Lu, Y., Wang, J. S. -H., & Han, W.-J. (2017). Women's short-term employment trajectories following birth: Patterns, determinants, and variations by race/ethnicity and nativity. *Demography 54*(1), 93–118.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke J., & Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *2008 IEEE 24th International Conference on Data Engineering* (pp. 277–286). New York: IEEE. doi: 10.1109/ICDE.2008.4497436

Malec, D., Davis, W. W., & Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine 18*(23), 3189–3200.

Maples, J. J. (2017). Improving small area estimates of disability: Combining the American Community Survey with the Survey of Income and Program Participation. *Journal of the Royal Statistical Society* (*Series A: Statistics in Society*) *180*(4), 1211–1227.

McClure, D., & Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy 5*(3), 535–552.

McKenna, L. (2019). *A History of the Survey of Income and Program Participation and disclosure Avoidance*. Washington, DC: U.S. Census Bureau, https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-McKenna-sipp%20and%20da.pdf

McKenna, R., Miklau, G., & Sheldon, D. (2021). *Winning the NIST contest: A scalable and general approach to differentially private data*. *Journal of Privacy and Confidentiality 11(3)*. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/778/727

McKernan, S.-M., Ratcliffe, C., & Braga, B. (2021). The effect of the US safety net on material hardship over two decades. *Journal of Public Economics 197,* 104403. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8153365/

Medalia, C., Meyer, B. D., O'Hara, A. B., & Wu, D. (2019). Linking survey and administrative data to measure income, inequality, and mobility. *International Journal of Population Data Science 4*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8142965/

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9(4), 538–558.

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., & Wilkins, A. (2020). Facebook privacy-protected full URLs data set, https://doi.org/10.7910/DVN/TDOAPG, HarvardDataverse, V10.

Meyer, B. D., Wu, D., Mooers, V., & Mdealia, C. (2018). *The Use and Misuse of Income Data and the Rarity of Extreme Poverty in the United States.* Institute for Research on Poverty Summer Research Workshop. https://www.russellsage.org/sites/default/files/use%20and%20misuse%20of%20income%20data.pdf

Copyright National Academy of Sciences. All rights reserved.

Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., & Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics 66*(1), 55–64.

Monte, L. M. (2017, March). Multiple partner fertility research brief. *Current Population Reports.* https://www.census.gov/content/dam/Census/library/publications/2017/demo/p70br-146.pdf

Nandy, S., Holan, S. H., & Schweinberger, M. (2023). A Socio-Demographic Latent Space Approach to Spatial Data When Geography Is Important but Not All-Important. https://doi.org/10.48550/arXiv.2304.03331

Narayanan, A., & Felten, F. W. (2014). No silver bullet: De-identification still doesn't work. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy* (pp. 111–125) (Oakland, CA, USA). doi: 10.1109/SP.2008.33

Narayanan, A., & Shmatikov, V. (2010, June). Myths and fallacies of "Personally Identifiable Information."*Communications of the ACM 53*(6), 24–26. https://doi.org/10.1145/1743546.1743558

National Academies of Sciences, Engineering, and Medicine (NASEM). (2018). *The 2014 Redesign of the Survey of Income and Program Participation: An Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24864

National Research Council. (1989). *The Survey of Income and Program Participation: An Interim Assessment*. Washington, DC: The National Academies Press.

National Research Council. (1993). *The future of the Survey of Income and Program Participation*. Washington, DC: The National Academies Press. https://doi.org/10.17226/2072

National Research Council. (2009). *Reengineering the Survey of Income and Program Participation.* Washington, DC: The National Academies Press. https://doi.org/10.17226/12715

Near, J. P., & Xi He, X. (2021). Differential privacy for databases. *Foundations and Trends in Databases 11*(2), 109–225. http://dx.doi.org/10.1561/1900000066

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford, CA: Stanford University Press.

Nixon, M. P., Barrientos, A. F., Reiter, J. P., & Slavković, A. (2022). A latent class modeling approach for generating synthetic data and making posterior inferences from differentially private counts. *Journal of Privacy and Confidentiality 12*(1). https://doi.org/10.29012/jpc.768

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software 74*(11), 1–26.

Office of Management and Budget (OMB). (2013). Open data policy—Managing information as an asset. Memorandum No. M-13-13. Washington, DC.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modelling and inference. *The Journal of Machine Learning Research 22*(1), 2617–2680.

Park, M., Foulds, J., Chaudhuri, K., & Welling, M. (2020). Variational Bayes in private settings (VIPS). *Journal of Artificial Intelligence Research 68*, 109–157.

Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In Jeffrey E. Cohen (Ed.), *Public Opinion in State Politics* (Redwood City, CA: Stanford University Press), pp. 209–28.

Parker, P. A., Holan, S. H., & Janicki, R. (2020). Conjugate Bayesian unit-level modeling of count data under informative sampling designs. *Stat 9*(1), e267. https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.267

Parker, P. A., Holan, S. H., & Janicki, R. (2023a). Conjugate modeling approaches for small area estimation with heteroscedastic structure. *Journal of Survey Statistics and Methodology* 00, 1–20. *https://doi.org/10.1093/jssam/smad002*

Copyright National Academy of Sciences. All rights reserved.

Parker, P. A., Janicki, R., & Holan, S. H. (2023b, forthcoming). A comprehensive overview of unit level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology 11, 829-857*. https://doi.org/10.1093/jssam/smad020

Pilkauskas, N. V., & Cross, C. (2018). Beyond the nuclear family: Trends in children living in shared households. *Demography 55*(6), 2283–2297. https://doi.org/10.1007/s13524-018-0719-y

Pistner, M., Slavković, A., & Vilhuber, L. (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In J. Domingo-Ferrer & F. Montes (Eds.), *Privacy in Statistical Databases*. Springer-Verlag. https://doi.org/10.1007/978-3-319-99771-17

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Porter, A. T., Wikle, C. K., & Holan, S. H. (2015). Small area estimation via multivariate Fay–Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics 57*(1), 15–29.

Potok, N., & Hart, N. (2022). *A Blueprint for Implementing the National Secure Data Service: Initial Governance and Administrative Priorities for the National Science Foundation*. Washington, D.C.: Data Foundation.

Pozen, D. E. (2005, December). The mosaic theory, national security, and the Freedom of Information Act. *The Yale Law Journal*, 628–679.

PPTTT (Privacy Preserving Techniques Task Team). (2023). *UN Handbook on Privacy-Preserving Computation Techniques* (D. W. Archer & D. Bogdanov, Eds.) BigData UN Global Working Group. https://unstats.un.org/bigdata/task-teams/privacy/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf

Qardaji, W., Yang, W., & Li, N. (2014, June). Priview: Practical differentially private release of marginal contingency tables. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery), pp. 1435–1446.

Quick, H., Holan, S. H., Wikle, C. K., & Reiter, J. P. (2015). Bayesian marked point process modelling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, *14*, 439–451.

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 9, e53275.

Raab, G. M., Nowok, B., & Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality, 7*(3), 67–97. https://doi.org/10.29012/jpc.v7i3.407

Raghunathan, T. E. (2016). *Missing data analysis in practice*. CRC Interdisciplinary Statistics Series. Chapman & Hall.

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

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85–96.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19*(1), 1.

Raghunathan, T. E., Solenberger, P., Berglund, P., & Van Hoewyk, J. (2016). *IVEware: Imputation and variance estimation software* (version 0.3). Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan. https://src.isr.umich.edu/wp-content/uploads/iveware_manual_revised.pdf

Raghunathan, T. E., Solenberger, P. W., & Van Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software.* Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Rao, J. N., & Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Copyright National Academy of Sciences. All rights reserved.

Rein, D. B., Zhang, P., Wirth, K. E., Lee, P. P., Hoerger, T. J., McCall, N., Klein, R., Tielsch, J. M., Vijan, S., & Saadine, J. (2006). The economic burden of major adult visual disorders in the United States. *Archives of Ophthalmology 124*(12), 1754–1760.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology 29*, 181–188. http://www2.stat.duke.edu/~jerry/Papers/sm03.pdf

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, *168*, 185–205.

Reiter, J. P. (2005b). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics 21*(3), 441.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly 76*(1), 163–181.

Reiter, J. P. (2021). Assessing uncertainty when using linked administrative records. In A.Y. Chun, M.D. Larsen, G. Durrant and J.P. Reiter (Eds.), *Administrative Records for Survey Methodology* (Wiley), pp. 139–153. https://doi.org/10.1002/9781119272076.ch6

Reiter, J. P. (2023) *Synthetic data: A look back and a look forward. Transactions on Data Privacy 16(1),* 15-24.

Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality 1.1*, 99–110.

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis 53*(4), 1475–1482.

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, 1530–1538.

Ribar, D. C. (2005). Transitions from welfare and the employment prospects of low-skill workers. *Southern Economic Journal*, *71*(3), 514–533. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2325-8012.2005.tb00655.x

Rinott, Y., O'Keefe, C., Shlomo, N., & Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Sciences 33*(3), 358–385.

Rocher, L., Hendrickx, J.M., & De Montjoye, Y. (2019). Estimating the success of re-identifications in incomplete datasets using generative models, *Nature Communications 10*(1), *3069*.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Ruggles, S., Flood, S., Sobek, M., Brockman, D., Cooper, G., Richards, S., & Schouweiler, M. (2023). IPUMS USA: Version 13.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D010.V13.0

Ruggles, S., & Magnuson, D. (2020, June). Census technology, politics, and institutional change, 1790-2020**.** 2020. *Journal of American History, 107*(1), 19–51. doi: 10.1093/jahist/jaaa007

Sabia, J. J., & Nielsen, R. B. (2015). Minimum wages, poverty, and material hardship: New evidence from the SIPP. *Review of Economics of the Household 13*(1), 95–134. Sakshaug, Y. W. & Raghunathan, T. E. (2010). Synthetic data for small area estimation. In *International Conference on Privacy in Statistical Databases* (pp. 162–173). Springer.

Savitsky, T. D., & Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, *10*(1), 1677–1708.

Scherer, Z. (2022). *Leave Usage Following a First Birth Among Men in the United States: Evidence From New Nationally Representative Data*. SEHSD Working paper No. 2033-05, U.S. Census Bureau, Washington, D.C. https://www.census.gov/content/dam/Census/library/working-papers/2022/demo/sehsd-wp2022-05.pdf

Copyright National Academy of Sciences. All rights reserved.

Scherer, Z., & Mayol-Gracia, Y. (2020, November). Transitions in parental presence among children: 2017. *Current Population Reports.* https://www.census.gov/content/dam/Census/library/publications/2020/demo/p70-169.pdf

Scherpf, E., Newman, C., & Prell, M. (2015). *Improving the Assessment of SNAP Targeting Using Administrative Records*, ERR-186, U.S. Department of Agriculture, Economic Research Service, May.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Seeman, J., Slavković, A., & Reimherr, M. (2020). *Private Posterior Inference Consistent with Public Information: A Case Study in Small Area Estimation From Synthetic Census Data*. Lecture Notes in Computer Science book series (LNISA, volume 12276). https://link.springer.com/chapter/10.1007/978-3-030-57521-2_23

Shaefer, H. L., & Edin, K. (2013). Rising extreme poverty in the United States and the response of federal means-tested transfer programs. *Social Service Review*, *87*(2), 250-268.

Sharma, M., Hutchinson, M., Swaroop, S., Honkela, A., & Turner, R. E. (2019). Differentially private federated variational inference. Privacy in Machine Learning Workshop (PriML 2019) at the 33rd Conference in Neural Information and Processing Systems (NeurIPS) https://doi.org/10.48550/arXiv.1911.10563

Shlomo, N. (2010). Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality 2*(1), 73–91.

Shlomo, N. (2020). Integrating differential privacy in the statistical disclosure control tool-kit for synthetic data production. In J. Domingo-Ferrer & K. Muralidhar (Eds.), *Privacy in Statistical Databases* (Spring LNCS 12276), pp. 271–280.

Shlomo, N., & De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics 24*(2), 1–26.

Shlomo, N., Krenzke, T., & Li, J. (2019). Confidentiality protection approaches for survey weighted frequency tables. *Transactions on Data Privacy 12*(3), 145–168.

Shlomo, N., & Skinner, C. (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions paper. *Journal of the Royal Statistical Society, Series A: Statistics in Society. 185*(4), *1644–1662* . http://dx.doi.org/10.1111/rssa.12902

Shlomo, N., & Young, C. (2008). Invariant post-tabular protection of census frequency counts. In J. Domingo-Ferrer & Y. Saygin (Eds.), *PSD'2008 privacy in statistical databases* (Springer LNCS 5261), pp. 77–89

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA), pp. 3–18. doi: 10.1109/SP.2017.41.

Si, Y., Pillai, N. S., Gelman, A., et al. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis 10*, 3, 605–625.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (Chichester: Wiley), pp. 80–84.

Skinner, C.J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B, 64*, 855–867.

Skinner, C.J. & Holmes, D. (1998) Estimating the re-identification risk per record in microdata. *Journal of Official Statistics 14*, 361–372.

Skinner, C. J. & Shlomo, N. (2008). Assessing identification risk in survey micro-data using log linear models. *Journal of American Statistical Association 103*(483), 989–1001.

Slavković, A. B., & Fienberg, S. E. (2004). Bounds for cell entries in two-way tables given conditional relative frequencies. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in Statistical Databases—PSD 2004*), Lecture Notes in Computer Science, vol 3050 (Berlin, Heidelberg: Springer), pp. 30–43.

Copyright National Academy of Sciences. All rights reserved.

Slavković, A., & Seeman, J. (2023). Statistical data privacy: A song of privacy and utility. *Annual Reviews of Statistics and Its Application*, *10*, 189–218. https://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-033121-112921

Smith, K., Williams, A. R., & Mudrazija, S. (2021, October 21). *Modeling Income in the Near Term*, Urban Institute Research Report, Urban Institute, Washington, DC. https://www.urban.org/sites/default/files/publication/104958/modeling-income-in-the-near-term.pdf

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavković, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(3), 663–688.

Snoke, J., & Slavković, A. (2018). pMSE Mechanism: Differentially private synthetic data with maximal distributional similarity. *Privacy in Statistical Databases*. https://www.semanticscholar.org/reader/3433dd5f43ace6f53f96f6a76af901b56d9a5ce2

Sohn, H. (2015). Health insurance and risk of divorce: Does having your own insurance matter? *Family Relations 77,* 982–995. https://doi.org/10.1111/jomf.12195

Sohn, H. (2017). Racial and ethnic disparities in health insurance coverage: Dynamics of gaining and losing coverage over the life-course. *Population Research and Policy Review 36*, 181–201.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems 32.* https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf

Stanley, J., & Totty, E. (2021). *A Penny Synthesized Is a Penny Earned? An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta*. Washington, DC: U.S. Census Bureau. https://www2.census.gov/adrm/CED/Papers/CY21/2021-006-StanleyTotty.pdf

Stanley, J., & Totty, E. (2023, May 4–5). A penny synthesized is a penny earned? An exploratory analysis of accuracy in the SIPP Synthetic Beta. (Conference paper). Delivered at the NBER Conference on Data Privacy Protection and Applied Research: Methods, Approaches, and Their Consequences, Cambridge, MA. https://conference.nber.org/conf_papers/f175215.pdf

Stykes, J. B., & Guzzo. K. B. (2019). Multiple-partner fertility: Variation across measurement approaches. In Schoen, R. (Ed.) *Analytical Family Demography*. The Springer Series on Demographic Methods and Population Analysis, vol 47 (Springer, Cham.), pp. 215–239. https://doi.org/10.1007/978-3-319-93227-9_10

Su, B., Wang, Y., Schiavazzi, D. E., & Liu, F. (2023). Differentially private normalizing flows for density estimation, data synthesis, and variational inference with application to electronic health records. https://doi.org/10.48550/arXiv.2302.05787

Sun, A., Parker, P. A., & Holan, S. H. (2022). Analysis of household pulse survey public-use microdata via unit-level models for informative sampling. *Stats 5*(1), 139–153.

Takahashi, T., Takagi, S., Ono, H., & Komatsu, T. (2020). Differentially private variational autoencoders with term-wise gradient aggregation. https://doi.org/10.48550/arXiv.2006.11204

Tamborini, C. R., & Villarreal, A. (2021). Immigrants' employment stability over the Great Recession and its aftermath. *Demography 58*(5), 1867–1895.

Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., & Miklau, G. (2022). Benchmarking differentially private synthetic data generation algorithms. Presented at the Third AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22), Association for the Advancement of Artificial Intelligence. https://doi.org/10.48550/arXiv.2112.09238

Taylor, S., MacDonald, G., Ueyama, K., & Bowen, C. M. (2021). *A Privacy-Preserving Validation Server Prototype*. Technical paper, Urban Institute, Washington, DC.

Tersine, A. G. (2022, September 7). Source and accuracy statement for the Survey of Income and Program Participation calendar year 2021 data collection public use files (S&A-26). (Memorandum). U.S. Bureau of the Census. https://www2.census.gov/programs-surveys/sipp/tech-documentation/source-accuracy-statements/2021/sipp-2021-SA-07-SEP22.pdf

Copyright National Academy of Sciences. All rights reserved.

Thompson, G., Broadfoot, S., & Elazar, D.L. (2013, October 28–30). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics (Conference session). Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

Thompson, D., & King, M.D. (2022, February). Income sources of older households: 2017. *Current Population Reports.* https://www.census.gov/content/dam/Census/library/publications/2022/demo/p70br-177.pdf

Tournier, A. J., & de Montjoye, Y. (2022). Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. *Science Advances 8*(3), https://doi.org/10.1126/sciadv.abl6464

Tran, T., Reimherr, M., & Slavković, A. (2023). Differentially private synthetic heavy-tailed data. https://doi.org/10.48550/arXiv.2309.02416

Tudor, C., Cornish, G., & Spicer, K. (2014). Intruder testing on the 2011 UK Census: Providing practical evidence for disclosure protection. *Journal of Privacy and Confidentiality, 5*(2). https://doi.org/10.29012/jpc.v5i2.632

Ullman, J., & Vadhan, S. (2020). PCPs and the hardness of generating synthetic data. *Journal of Cryptology* 33(4): 2078-2112.

U.S. Bureau of the Census (Census Bureau). (2021). *2020 Survey of Income and Program Participation Users' Guide*. Washington, DC. https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2020_SIPP_Users_Guide_OCT21.pdf

Van den Burg, G. J. J., & Williams, C. K. I. (2021). On memorization in probabilistic deep generative models. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021*). https://proceedings.neurips.cc/paper_files/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4-Paper.pdf

Van Hook, J., & Glick, J. E. (2007). Immigration and living arrangements: Moving beyond economic need versus acculturation. *Demography 44.2,* 225–249.

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., & Hens, N. (2016). Model-based inference for small area estimation with sampling weights. *Spatial Statistics 18*, 455–473.

Vaughan, D. R., Haley, B., & Dajani, A. (2021). Ten years later: Self-sufficiency of welfare mothers before the Great Recession, *Poverty & Public Policy 13*(3), 1–40. https://onlinelibrary.wiley.com/doi/abs/10.1002/pop4.308

Vietri, G., Archambeau, C., Aydore, S., Brown, W., Kearns, M., Roth, A., Siva, A., Tang, S., & Wu, S. Z. (2022). Private synthetic data for multitask learning and marginal queries. *Proceedings, Advances in Neural Information Process Systems 35*. https://proceedings.neurips.cc//paper_files/paper/2022/hash/7428310c0f97f1c6bb2ef1be99c1ec2a-Abstract-Conference.html

Vilhuber, L. (2019). Utility of two synthetic data sets mediated through a validation server: Experience with the Cornell Synthetic Data Server. Presentations by the Labor Dynamics Institute, Cornell University. https://ecommons.cornell.edu/handle/1813/69624

Vought, R. (2019a). Office of Management and Budget Memorandum M-19-15, April 24, 2019. https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf

Vought, R. (2019b). Office of Management and Budget Memorandum M-19-23, July 10, 2019. https://www.whitehouse.gov/wp-content/uploads/2019/07/m-19-23.pdf

Vu, D., & Slavković. A. (2009). Differential privacy for clinical trial data: Preliminary evaluations (Conference session). In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW '09* (Washington, DC: IEEE Computer Society), pp. 138–143.

Waites, C., & Cummings, R. (2021, July). Differentially private normalizing flows for privacy-preserving density estimation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery), pp. 1000–1009. https://dl.acm.org/doi/10.1145/3461702.3462625

Wang, Y. X., Fienberg, S., & Smola, A. (2015, June). Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *International Conference on Machine Learning* (PMLR), pp. 2493–2502.

Copyright National Academy of Sciences. All rights reserved.

Wang, J. C., Holan, S. H., Nandram, B., Barboza, W., Toto, C., & Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics 17*, 84–106.

Wang, H., & Reiter, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *The Annals of Applied Statistics 6*(1), 229.

Warren, L., & Tettenhorst, A. (2022, September). Poverty dynamics: 2017–2019. *Current Population Reports*. https://www.census.gov/content/dam/Census/library/publications/2022/demo/p70br-179.pdf

Wasserman, L., & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association 105*(489), 375–389.

Williams, A. R., Snoke, J., Bowen, C. M., & Barrientos, A. F., (2023). Disclosing economists' privacy perspectives: A survey of American Economic Association members on differential privacy and data fitness for use standards. (Conference session). Presentation at NBER conference: Data privacy protection and the conduct of applied research: Methods, approaches and their consequences (Cambridge, MA), May 4–5. https://conference.nber.org/conf_papers/f178417.pdf

Wilson, R.J., Zhang, C. Y., Lam, W., Desfontaines, D., Simmons-Marengo, D., & Gipson, B. (2020). Differentially private SQL with bounded user contribution. *Proceedings on Privacy Enhancing Technology* (2), pp. 230–250. doi: https://doi.org/10.2478/popets-2020-0025

Winkler, W. E. (2006). *Overview of Record Linkage and Current Research Directions*. Statistical Research Division report. Washington, DC: U.S. Bureau of the Census. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

Woo, M. J., Reiter, Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality 1*(1).

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., & Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law 209*. http://www.jetlaw.org/wp-content/uploads/2018/12/4_Wood_Final.pdf.

Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. https://doi.org/10.48550/arXiv.1802.06739.

Xiong, H. (2008). *The US Occupational Mobility from 1988 to 2003: Evidence from SIPP*. Mimeo, University of Toronto, Canada.

Xu, J., Ren, X., Lin, J., & Sun, X., (2018). DP-GAN: Diversity-promoting generative adversarial network for generating informative and diversified text. 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Special Interest Group on Linguistic Data & Corpus-based Approach to Natural Language Processing, Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1802.01345

Yancey, W.E., Winkler, W.E., & Creecy, R.H. (2002) Disclosure risk assessment in perturbative micro-data protection. In J. Domingo-Ferrer (Ed.), *Inference Control in Statistical Databases* (Ed.), 135–151. New York: Springer.

Yıldırım, S., & Ermis, B. (2019). Exact MCMC with differentially private moves. *Statistics and Computing 29*(5), 947–963.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., & Mironov, I. (2021). Opacus: User-friendly differential privacy library in PyTorch. Presented at Privacy in Machine Learning workshop, NeurIPS 2021, online, Dec. 6-14, 2021. Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2109.12298

Yu, M., He, Y., & Raghunathan, T. E. (2022). A semiparametric multiple imputation approach to fully synthetic data for complex surveys, *Journal of Survey Statistics and Methodology 10*(3), 618–641. https://doi.org/10.1093/jssam/smac016

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). Privbayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS) 42*(4), 1–41.

Copyright National Academy of Sciences. All rights reserved.

Zhang, Y., & Van Hook, J. (2009). Marital dissolution among interracial couples. *Journal of Marriage and Family 71*(1), 95–107.

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics 125*(103977).

References to Data Files, Software Packages, and Data Production Codes

Stan Development Team. (2023). RStan: the R interface to Stan. R package version 2.21. 8, https://mc-stan.org/

University of Michigan. https://www.src.isr.umich.edu/software/iveware/

Urban Institute. (https://www.urban.org/research/publication/privacy-preserving-validation-server-prototype)

U.S. Bureau of the Census (Census Bureau). (2022). https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

U.S. Bureau of the Census (Census Bureau). (2023). https://www.census.gov/topics/research/guidance/restricted-use-microdata/standard-application-process.html

Copyright National Academy of Sciences. All rights reserved.

Appendix A
Technical Details on Measuring Disclosure Risk

Denote F_k as the population size and f_k the sample size in cells *k* defined by cross-classifying the *K* key variables. The set of sample uniques is defined as: $SU = \{k: f_k = 1\}$ and these are the high-risk records in the public-use file that have the potential to be population uniques. The DIS measure (Skinner & Elliot, 2002) is defined as:

$$DIS = \frac{\sum_k I(f_k = 1)}{\sum_k I(f_k = 1)F_k}$$

where *I* is the identify function. The *DIS* measure is the number of sample uniques divided by the sum of the population sizes in those cells defined by SU. The advantage of the *DIS* measure is that it can easily be estimated without the need for any probabilistic modelling. The estimate for the *DIS* is:

$$\widehat{DIS} = \pi n_1/[\pi n_1 + 2(1-\pi)n_2]$$

where π is the sample fraction, $n_1 = \sum_k I(f_k = 1)$ the number of sample uniques and $n_2 = \sum_k I(f_k = 2)$ the number of sample doubles.

As mentioned in Chapter 3, generally the 'fishing' scenario is not sufficient for statistical agencies as they need to assess the consequences of releasing a public-use file. Therefore, a more relevant disclosure risk measure is based on the attack scenario, that is, when an intruder has access to the confidential microdata and wants to make an identification based on an external data source containing information about the population. Two main disclosure risk measures are:

(1) Expected number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$, and

(2) Expected number of correct matches for sample uniques assuming a random assignment within cell *k*.

For example, if a sample unique matches to three individuals in the population, the match probability for that sample unique would be 1/3. Aggregating all match probabilities over the sample uniques leads us to: $\tau_2 = \sum_k I(f_k = 1)\ 1/\ F_k$

Shlomo and Skinner (2022) and Shlomo (2010) (and references therein) provide an overview of the literature on measuring the risks of re-identification (1) and (2) using probabilistic modelling to estimate the population sizes in cell *k*.

Copyright National Academy of Sciences. All rights reserved.

Using a probabilistic model, the risk measures are estimated by:

$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(F_k = 1|f_k = 1)$ and $\hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(1/F_k|f_k = 1)$

Skinner and Holmes (1998) and Elamir and Skinner (2006) propose a Poisson distribution and a log-linear model to estimate the disclosure risk measures. In this model, they assume that $F_k \sim Pois(\lambda_k)$ for each cell *k*. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell *k*: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k\lambda_k) \text{ and } F_k|f_k \sim Pois(\lambda_k(1-\pi_k))$$

where the population cell counts F_k are assumed independent given the sample cell counts f_k.

The parameters λ_k are estimated using log-linear modelling. The sample frequencies f_k are independent Poisson-distributed with a mean of $\mu_k = \pi_k\lambda_k$. A log-linear model for the μ_k is expressed as: $log\ (\mu_k) = x_k'\beta$, where x_k is a design vector that denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimators $\hat{\beta}$ are obtained by solving the score equations: $\sum_k (f_k - \pi_k exp(x_k'\beta))x_k = 0$.

The fitted values are then calculated by: $\hat{\mu}_k = exp(x_k'\hat{\beta})$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Individual disclosure risk measures for cell *k* are:

$$P(f_k = 1) = exp\ (\lambda_k(1-\pi_k))$$
$$E(1/F_k|f_k = 1) = (1 - exp\ (\lambda_k(1-\pi_k)))/\ (\lambda_k(1-\pi_k))$$

Plugging $\hat{\lambda}_k$ for λ_k in the formula above leads to the estimates $\hat{P}(F_k = 1|f_k = 1)$ and $\hat{E}(1/F_k|f_k = 1)$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$. Skinner and Shlomo (2008) develop a method for selecting the main effects and interactions of the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. There have been many extensions and similar approaches to estimate the risk of re-identification based on probabilistic modelling and these are mentioned in Shlomo and Skinner (2022). In addition, the R package SDCNway on the CRAN network has been developed that carries out the probabilistic modelling for estimating the risk of re-identification.

Copyright National Academy of Sciences. All rights reserved.

Appendix B Inferences Based on Multiple Synthetic Data

Denote the number of synthetic datasets by m, the estimated of the parameter of interest θ is $\hat{\theta}^{(j)}$ and the within-set variability of $\hat{\theta}_j$ by $w^{(j)}$ (Liu, 2022). Let $W = m^{-1}\sum_{j=1}^{m} w^{(j)}$ (average within-set variability) and $B = \sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \hat{\theta}\right)^2/(m-1)$ (between-set variability).

Fully synthetic data: The final estimate of θ over m synthetic sets is $\hat{\theta} = m^{-1}\sum_{j=1}^{m} \hat{\theta}^{(j)}$ and its estimated variability is given by $T = (1 + m^{-1})B - W$. Hypothesis testing and confidence interval construction are based on the asymptotic assumption of $T^{\frac{1}{2}}\left(\hat{\theta} - \theta\right) \sim N(0,1)$.

Partial synthetic data with or without DP: The final estimate of θ over m synthetic sets is $\hat{\theta} = m^{-1}\sum_{j=1}^{m} \hat{\theta}^{(j)}$ and the variance estimator is $T = W + m^{-1}B$. Hypothesis testing and confidence interval construction are based on the asymptotic assumption of $T^{\frac{1}{2}}\left(\hat{\theta} - \theta\right) \sim t_{\nu}(0,1)$, where the degrees of freedom are $\nu = (m-1)\left(1 + \frac{W}{m^{-1}B}\right)^{-1}$. Though the inferential approaches based on multiple synthetic datasets are the same with or without DP, what is captured in the between-set variance component B is different between the two. For DP data synthesis, B has the extra variability in the synthetic data due to the employment of randomized mechanisms for achieving DP guarantees, compared to the case without DP.

Copyright National Academy of Sciences. All rights reserved.

Appendix C
Technical Details for Differential Privacy Table Builder

The technical details on the use of differential privacy in flexible table generators is described in this appendix using the exponential mechanism with a Discrete Laplace Distribution (Rinott et al., 2018). The mechanism $M(.)$ is defined as follows: given a count $a \in A$, select a count $b \in B$ (where B is the range of the output b) with probability proportional to $exp\left(\frac{\frac{\varepsilon}{2}u}{\Delta u}\right)$ where Δu is the sensitivity defined as:

$$\Delta u = \max_{b \in B} \max_{a \sim a' \in A} | u(a, b) - u(a', b) |$$

Given the need to cap perturbations: $P(M(a) = b) < \delta$, then for all neighboring $a \sim a' \in A$, if $P(M(a') = b) = 0$ implies $|a_k - b_k| \leq m$ for all k, then $M(.)$ satisfies (ε, δ).

Table C1 displays examples of Discrete Laplace perturbation probability vectors when the sensitivity Δu is 1 for the internal cells of the table. For each level of ε and δ we show the amount of cell perturbation depending on the draw of a random uniform number, the probability of perturbation, and the cumulative vector of probabilities of perturbation. For example, for $\varepsilon = 0.5$, $\delta = 0.008$, if we draw a random uniform number of 0.8, which falls between the range of 0.77 and 0.87, the cell value will be perturbed by adding 2 to the cell total. Note that if the mechanism leads to perturbations of negative counts, they can be set to zero without invalidating the property of differential privacy.

TABLE C-1 Examples of Two Discrete Laplace Perturbation Vectors for $\varepsilon = 1.5$, $\delta = 0.00002$ and $\varepsilon = 0.5$, $\delta = 0.008$

Amount of cell perturbation	$\varepsilon = 1.5, \delta = 0.00002$		$\varepsilon = 0.5, \delta = 0.008$	
	Probability of perturbation	Cumulative probability	Probability of perturbation	Cumulative probability
-7	0.00002	0.00002	0.0076	0.00760
-6	0.00008	0.00010	0.0125	0.02010
-5	0.00035	0.00045	0.0206	0.04070
-4	0.00157	0.00202	0.0339	0.07460
-3	0.00706	0.00908	0.0559	0.13050
-2	0.03162	0.04070	0.0922	0.22270
-1	0.14172	0.18242	0.1520	0.37470
0	0.63516	0.81758	0.2506	0.62530
1	0.14172	0.95930	0.1520	0.77730
2	0.03162	0.99092	0.0922	0.86950
3	0.00706	0.99798	0.0559	0.92540

Copyright National Academy of Sciences. All rights reserved.

4	0.00157	0.99955	0.0339	0.95930
5	0.00035	0.99990	0.0206	0.97990
6	0.00008	0.99998	0.0125	0.99240
7	0.00002	1.00000	0.0076	1.00000

The way to ensure a single privacy budget is to ensure the property that any time a cell is aggregated in any table, the 'seed' determining the perturbation amount is fixed, that is, the 'same cell-same perturbation' rule is applied. This is carried out by assigning to each individual in the microdata a random number, the microdata key. When aggregating individuals into a cell, the microdata keys are also aggregated, and this forms the seed (the 'Cell-Key') of the perturbation. Thus, the same cell will always have the same perturbation. Define:

- *Cell consistency* – Across multiple users, if the same set of records contributes to a table cell, the same results are attained. This is attained by a simple sum of the microdata key among cell members.
- *Query consistency* – Across multiple users using the same query path (e.g., same specification for universe definition and requested table), the same results are attained. This is attained by a function of the simple sum of the microdata key for the cell, marginals associated with the cell, and the table universe.

Attaining cell consistency has less protection than attaining query consistency. For instance, in the extreme scenario of table differencing for explicit tables that differ by one case, attaining cell consistency has the consequence that the potential identification of the true attribute can be found because all cells but one have zero sum of weights in the implicit tables from differencing. This is why one needs to ensure the building blocks (hypercubes) as the input of the table generator, which will not allow for this extreme scenario.

When applying differential privacy to a flexible table builder, there needs to be reflection on what lower-level margins would be available. In general, if one allows 4 dimensions in the tables, this leads to 1 four-way table, 4 three-way margins, 6 two-way margins, and 4 one-way margins, meaning that an individual can appear multiple times in the table. This means that the sensitivity is now $d = 2^4 - 1 = 15$, thus changing the sensitivity by d. In general, this means that one needs to define an overall privacy budget ε as $\frac{\varepsilon}{d}$ (or at least ensure that the different $\varepsilon's$ across margins add up to the overall privacy budget according to the Composition theorem). This can lead to a rather large overall privacy budget. Therefore, preliminary work needs to be undertaken as to what margins will be released, thus lowering the sensitivity of the privacy budget. Alternatively, one can direct more research on using correlated noise to ensure marginal distributions (in expectation) as evidenced in the early disclosure avoidance literature by changing from the Laplace distribution to the Normal distribution (Shlomo & De Waal, 2008) or placing the property of invariance on the perturbation vectors (Shlomo & Young, 2008).

Since the SIPP is a probability sample and has survey weights, one can adjust for the weighted counts in the tables, as shown in Shlomo et al., 2019. The perturbation p is applied to the sample counts. Then, one adds (or subtract) $p \times \underline{w}$ from the weighted sample count where $\underline{w}$ is the average weight.

For continuous variables, such as sums, averages and quantiles, correlations, one can use the same concept of the microdata keys to obtain the same perturbations, but more research is required on how this is actually implemented in a differentially private setting.

Copyright National Academy of Sciences. All rights reserved.

In a non-differentially private setting, one can add multiplicative noise to the statistic by multiplying the statistic by (1 + p) and p is determined by the microdata keys where the perturbation vector is in a pre-set range, for example [-0.2, …, +0.2]. As an example, for a weighted sum $\hat{Y}$, one perturbs as follows: $\hat{Y} + \hat{Y} \times p \times \underline{w}$ (note that p can take a positive or negative value). Similarly, one can use this approach for averages, where the denominator is now: $\hat{w} + p \times \underline{w}$ (Shlomo, Krenzke, & Li, 2019).

For more advanced modelling in the remote analysis server, one can add noise p to the estimating equations, that is, instead of setting the score functions to 0 one solves them to s = p*max (residual). For a simple regression model, the solution for a perturbed regression coefficient is: $\beta^{Pert} = \beta^{Orig} + (X'X)^{-1}s$ (Shlomo, 2020).

Copyright National Academy of Sciences. All rights reserved.

Appendix D
Technical Details for Geography Variables

Hierarchical modelling is often conveniently thought of using three stages: data model (e.g., likelihood), process model, and parameter model. Using the notation $[Y|X]$ to denote the conditional distribution of Y given X, one may consider the following conditional distributions [data|process, parameters], [process|parameters], and [parameters]. Then, using Bayes' rule, the posterior distribution of the process and parameters given the data can be expressed as

[process, parameters|data]$\propto$ [data|process,parameters] $\times$ [process|parameters] $\times$ [parameters].

The well-known Fay-Herriot (FH) model (Fay III & Herriot, 1979) can be expressed as

$$Z = Y + \epsilon$$
$$Y = X\beta + \xi,$$

where $\epsilon \sim N(0, D)$, $D = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)$, $\xi \sim N(0, \sigma_\xi^2)$, Z are the direct estimators, and X are known covariates with associated parameters β_i $(i = 0,1, \dots, p)$. In this context, σ_i^2 $(i = 1, \dots, N)$ is the sampling error variance for area i $(i = 1, \dots, N)$ and p is the number of covariates (β_0 is the intercept). Thus, this can be expressed hierarchically with $[Z|Y, \beta, \sigma_\xi^2]$ equal to the distribution of data model given the process and parameters, $[Y|\beta, \sigma_\xi^2]$ the distribution of the process given the parameters, and $[\beta, \sigma_\xi^2]$ the distribution of the parameters.

Here is a description of the multivariate spatio-temporal mixed effects model (MSTM) (Bradley et al., 2015). For ease of exposition, this discussion presents the multivariate spatial case. Details surrounding the spatio-temporal case can be found in Bradley et al. (2015). Similar to the FH model, one has

$$Z \sim MVN(Y, D)$$
$$Y = X\beta + S\eta + \xi.$$

Here, MVN denotes the multivariate normal distribution, D contains the known sampling error variance, S is spatial basis functions with associated coefficients given by the elements of η (see Bradley et al., 2015, for a comprehensive discussion), and $\xi \sim MVN(0, \sigma_\xi^2 I)$ represents an additional error term to capture fine-scale variation. Importantly, the FH model can be viewed as a special case of the MSTM.

For the unit-level case, one first considers a model for an ignorable design (Battese et al., 1998). Specifically, consider the linear mixed-effects model

$$y_{ij} = x_{ij}\beta + v_i + \epsilon_{ij},$$

where y_{ij} is the response for unit j in area i $(i = 1, \dots, m)$, x_{ij} are fixed covariates associated with unit j in area i. β is associated regression coefficients, v_i is area-level random effects for area i with an iid mean zero normal

Copyright National Academy of Sciences. All rights reserved.

distribution having variance σ_v^2. Finally, ϵ_{ij} is iid normally distributed sampling error random effects with mean zero and variance σ_ϵ^2. Importantly, this model can be rewritten in the form of a hierarchical model.

In the case of an informative sample design, one path forward proceeds through the Bayesian pseudo-likelihood (PL) (Savitsky & Toth, 2016). The PL is given by

$$PL(\theta) = \prod_i f(y_i|\theta)^{w_i},$$

where unit i ranges over the sample and w_i denotes the sample weight for unit i, scaled to sum to the sample size. Combined with a suitable prior distribution on the model parameters θ, this leads to a pseudo-posterior distribution.

Copyright National Academy of Sciences. All rights reserved.

Appendix E
Data Collection Report

The panel felt that it needed to understand how SIPP data users access SIPP data and how they use the data. Panel members pursued multiple approaches. First, they studied publications based on SIPP data. There were three separate studies: one based on a bibliography of published studies that used SIPP data compiled by the U.S. Census Bureau, a second based on articles in the last 5 years across a mixture of methodological approaches, and the last based on the 50 most cited studies based on SIPP data as defined by the Scopus citation index. The results provided in Chapter 5 reflect the papers that are included in each category, though the details in the assessments based on these studies may reflect the judgment of the analyst who summarized the papers.

The panel also decided to collect information directly from SIPP data users. However, there is no list of SIPP users to provide a valid sampling frame for a scientific sample survey. The approach that was used and its limitations are described in this appendix.

DATA COLLECTION AND RESPONSE RATES

On September 13, 2022, CNSTAT published notice of an invitation to SIPP users and potential users to complete a questionnaire concerning their experiences in using SIPP data. The notice was published online, through an e-blast to CNSTAT subscribers, to the American Association for Public Opinion and Survey Research Methods Section of the American Statistical Association listservs, to users of the SIPP synthetic beta, and by using #EconTwitter. A follow-up notice was published on September 2, 2022. The last response was received on October 5, 2022.

A total of 65 responses were received, with 41 completing the full questionnaire and 24 partially completing the questionnaire. Because the notification was published as a general announcement, and neither the number who received the announcement nor the total number of SIPP users are known, true response rates cannot be calculated. However, there are 275 known users of the synthetic data, and 18 respondents reported using the synthetic data; this is approximately a 6 percent response rate for the synthetic data users.

The call for information was not designed to produce nationally representative results. Thus, these data provide a picture of the experiences of the SIPP users that responded to the call for information, but a truly representative sample might show different proportions in the responses. Typically, those who are most involved and have the strongest opinions are the most likely to respond to surveys, so the results are likely to overrepresent those who are heavy users of SIPP, who care deeply if access might be changed, and who had either highly positive

Copyright National Academy of Sciences. All rights reserved.

or highly negative experiences. They are likely to underrepresent those who might want only peripheral involvement with SIPP data, such as those who only wish to download tables.

The respondents to the questionnaire were promised that only summary statistics and possibly verbatim quotations would be made publicly available. For this reason, the database containing their responses will not be released.

CHARACTERISTICS OF THE RESPONDENTS

A total of 42 respondents provided information on their personal situation, with 10 being government employees, 9 being faculty members, 9 being students, and 9 belonging to research organizations (Figure E-1).

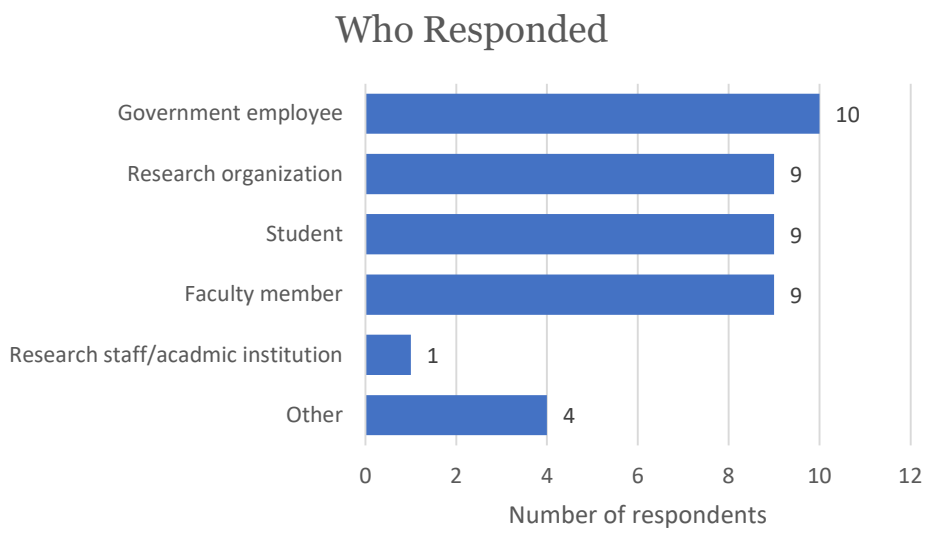


Figure E-1 Characteristics of respondents to call for information.

TYPES OF DATA USED

Respondents were asked which ways they used the SIPP data, with multiple responses allowed. Overall, 42 reported downloading the public-use file, 18 accessed the synthetic data, 17 downloaded tables or reports, and 7 accessed the data through a Federal Statistical Research Data Center (FSRDC). Additionally, 8 of the respondents had not yet used the SIPP data.

Looking at what combinations of data sources were used, 2 reported using all 4 sources, 25 reported using only one source, and 23 reported using some other combination of sources (Figure E-2).

Copyright National Academy of Sciences. All rights reserved.

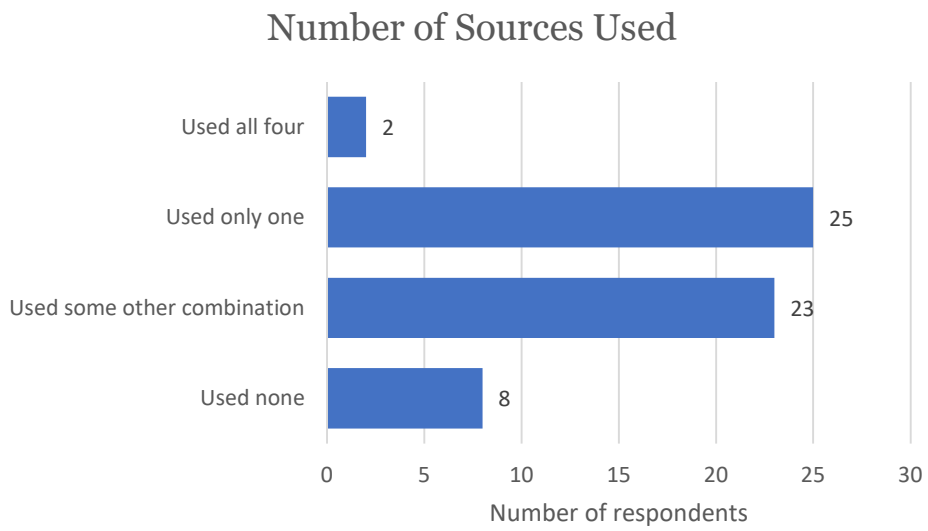


Figure E-2 Number of different SIPP data sources used by respondents to the call for information.

Faculty members and government employees were the most likely to use downloaded public-use data, while students and government employees were the most likely to use the synthetic data. Only government employees and faculty members used FSRDCs (Figure E-3).

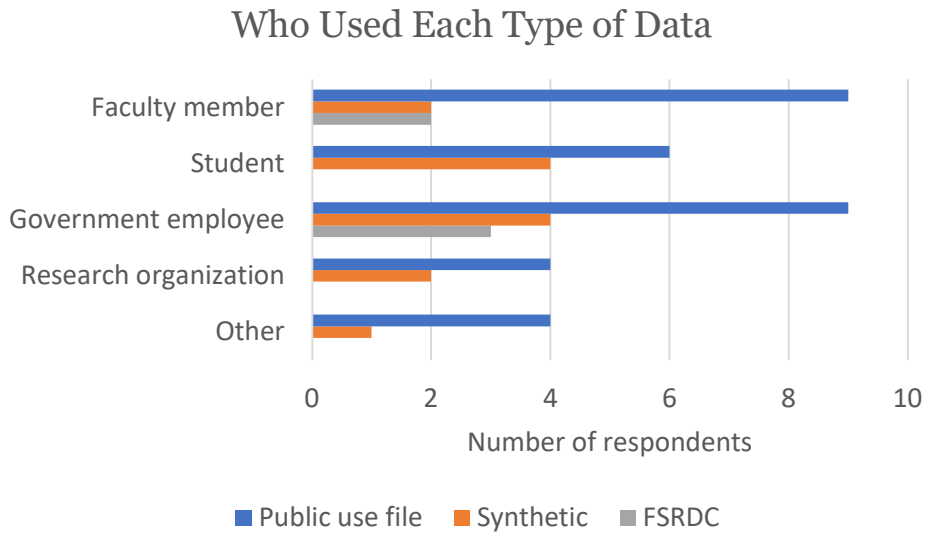


Figure E-3 Types of respondents to the call for information that used each format of SIPP data file.

WHICH DATA WERE USED

Every listed SIPP module, or topic area, was used by at least 3 respondents, and 6 were each used by at least half of those responding (Figure E-4). The most commonly used modules were income, employment and earnings, demographics, family and household, education, and poverty.

Copyright National Academy of Sciences. All rights reserved.

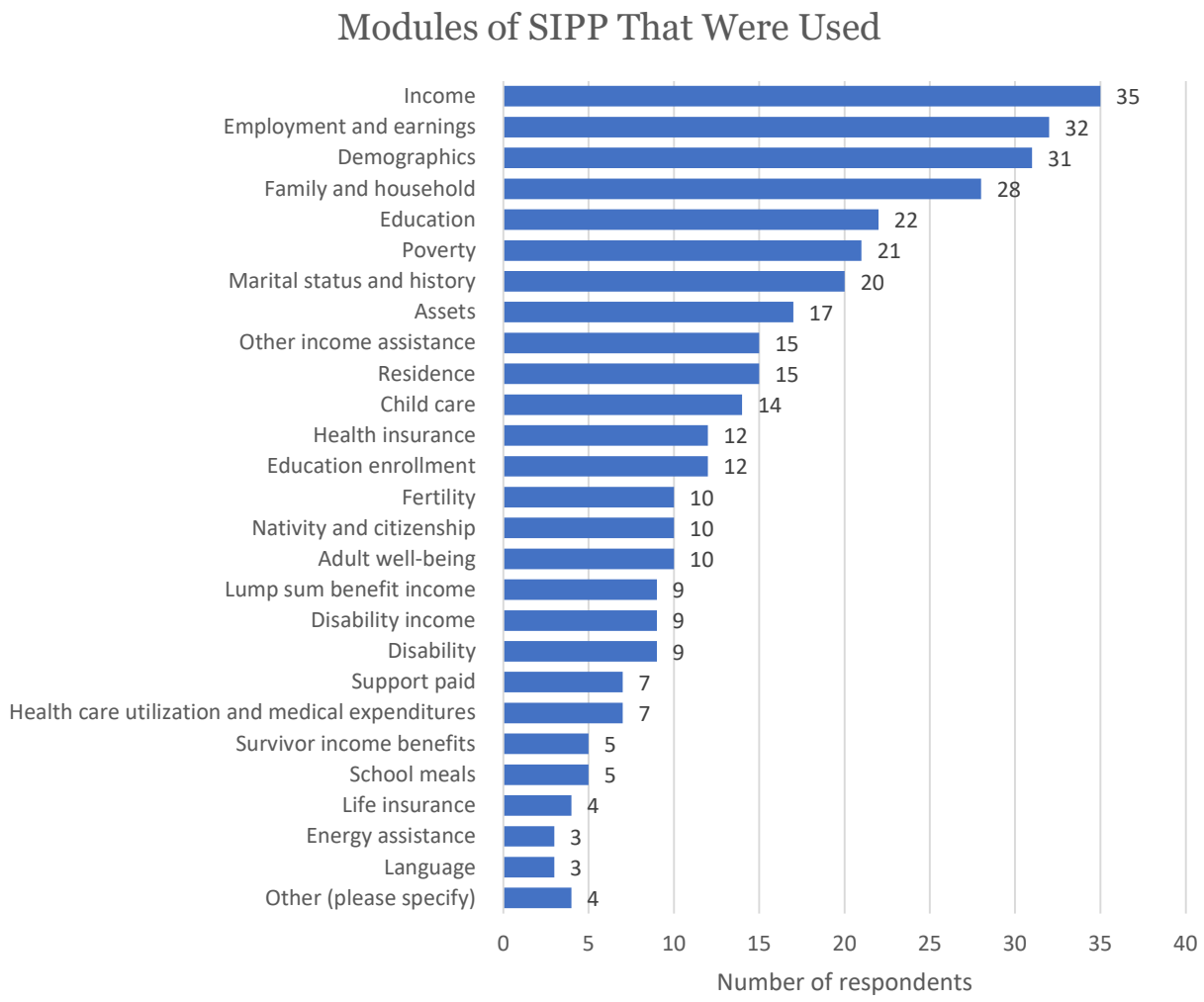


Figure E-4 Number of respondents to the call for information who used each module or topic area within SIPP.

Under "Other," respondents wrote in:

- "Commuting and work schedule"
- "Informal care module"
- "Work schedules"
- "Tax."

As might be inferred from the above statistics, SIPP data users typically used multiple modules, with every respondent using at least 2 of the listed modules, and one using all 26 (Figure E-5).[75] More than half used between 1 and 10 modules, but if a package of the top 10 modules were created, that would have met the needs of only 6 of the 41 users. No two respondents used exactly the same set of modules.

[75]Of the four respondents who wrote in additional modules, only the person adding "tax" failed to mark any other modules as having been used.

Copyright National Academy of Sciences. All rights reserved.

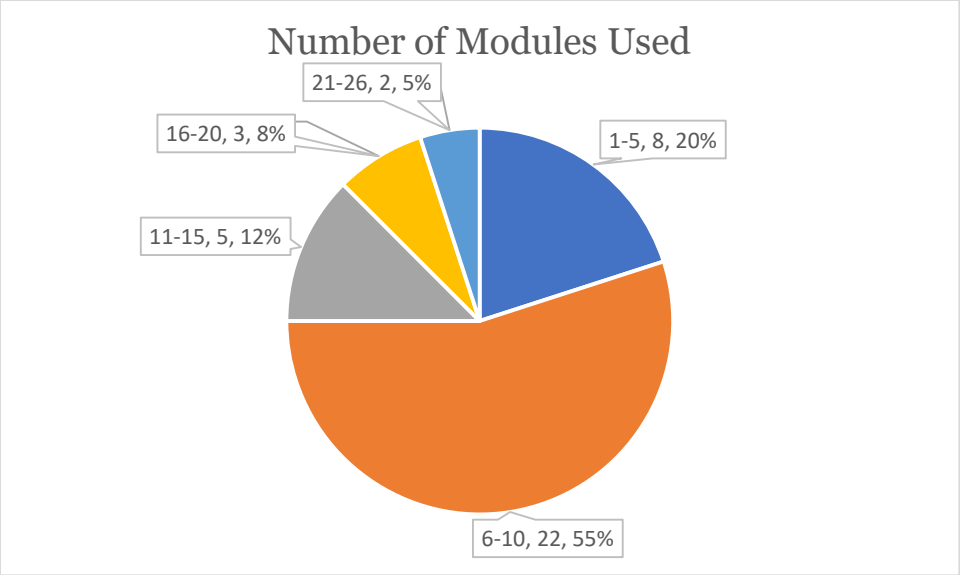


Figure E-5 Number of modules or topic areas used by respondents to the call for information.

Most typically, the analysis each respondent had performed was both cross-sectional and longitudinal, and next most often it was longitudinal only (Figure E-6).

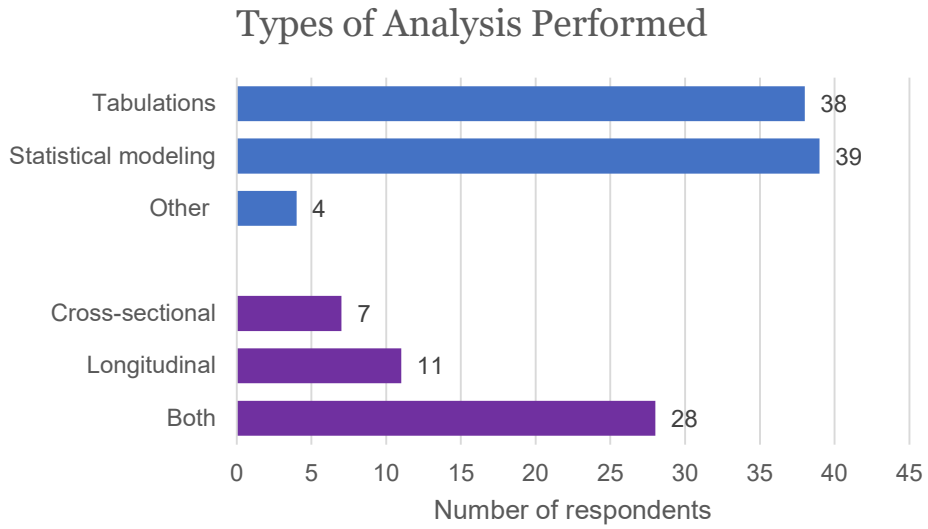


Figure E-6 Types of analysis performed by respondents to the call for information.

Almost all respondents performed both tabulations (83 percent) and statistical modelling (85 percent). Next to "Other," respondents offered the following comments:

- "Small area estimation"
- "No results"
- "None."

Copyright National Academy of Sciences. All rights reserved.

DIFFICULTIES ENCOUNTERED WHEN USING THE PUBLIC-USE FILES

Of the 42 respondents who had downloaded public-use files, 10 experienced some kind of difficulty, including both problems with the actual data (not having needed measures or losing important details) and more general issues reflecting weaknesses in documentation and in clarity on which files to download (Figure E-7). Next to "other," respondents wrote in:

- "Difficulties merging across waves due to sample redesign, questionnaire changes, especially when topical modules were used and then when they were not."
- "The older panels were difficult to understand. Changes to the structure and variables of the panels made it difficult to compare data over time."
- "Found the oldest panels confusing to use. Changes to data structure and individual variables made it difficult to compare results across panels."

Eight respondents said they either could not complete a research project or had to significantly change it because of a lack of access to detailed categories in the public data.

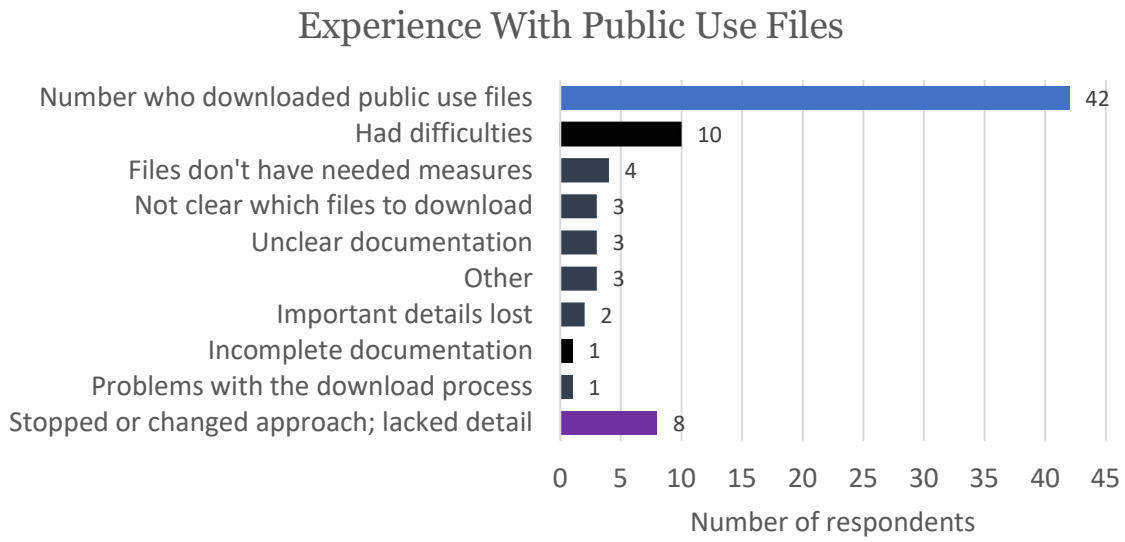


Figure E-7 Difficulties experienced by respondents to the call for information when using the public-use files.

DIFFICULTIES ENCOUNTERED WHEN USING SYNTHETIC DATA

Fewer respondents reported using synthetic data than public-use data (18 vs. 42), and fewer reported encountering difficulties (5 vs. 10), but the percentage encountering difficulties was roughly the same (28 percent vs. 24 percent) (Figures E-7 and E-8). Equal numbers reported having difficulties getting clearance for access, incomplete documentation, and unclear documentation. One person, responding to "Other," commented:

Copyright National Academy of Sciences. All rights reserved.

> The Cornell server and the validation server had different programming software versions. My code would run on the Cornell server but not the Census server, and I couldn't help fix the problem. This delayed my validation request for a few weeks, but Census staff were super nice and got it figured out.

Only one respondent reported having to stop or change the research approach because of inadequate data, but this may reflect how access to the synthetic data is obtained: to gain access to synthetic data, researchers must submit a proposal that Census determines to be feasible.

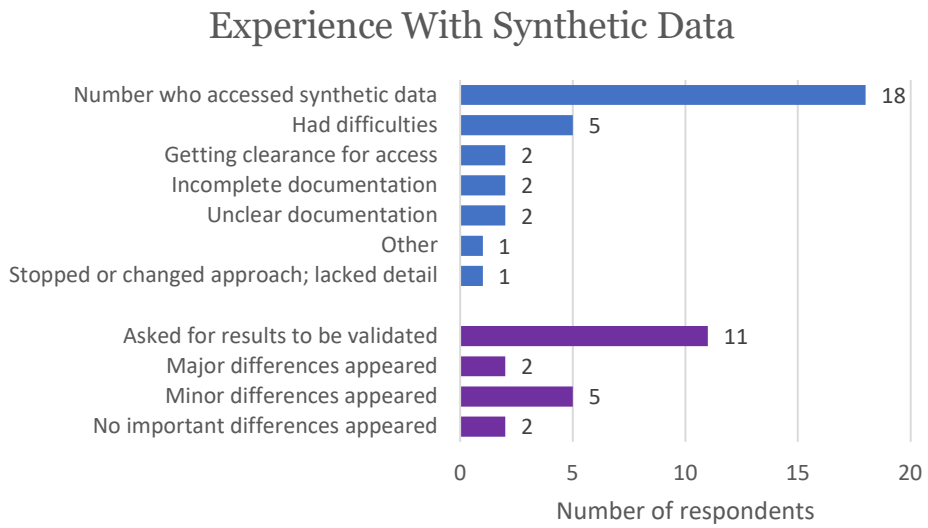


Figure E-8 Difficulties experienced by respondents to the call for information when using the synthetic data.

A difference between the public-use data and the synthetic data is that synthetic data users are encouraged to have their data validated on the Gold Standard File. Eleven of the 18 users reported asking for their results to be validated; 2 said major differences appeared, and 5 said minor differences appeared. Following are the comments provided by the respondents on this experience:

- "It was many years ago and I don't remember. I used the results from the internal file, not the synthetic results."
- "It was years ago, I don't remember. I used the results from the confidential file (after having them reviewed)."
- "Below are some of my experiences using the SSB/GSF and disclosing results. I say 'we' because all of this work is coauthored. The big caveat here is that we haven't done anything simple with the SSB/GSF, meaning, we haven't done anything that uses the SSB in a way that matches how they synthesized the data (at least as I understand it). I can't think of anything we've done where I would be comfortable using the point estimates from the SSB as a substitute for the GSF. An example of where things 'worked' is our work on mobility. In this paper, we estimate a bunch of decile transitions matrices, and some other measures of mobility that are closely related to the transition matrix. If I recall correctly, we wrote an entire draft of the paper with the results from

Copyright National Academy of Sciences. All rights reserved.

the SSB. When we got the GSF results, we had to change point estimates, but none of the substantive conclusions changed. In general, mobility was lower in the GSF than the SSB, but relative mobility among groups and at different points in the initial earnings distribution remained the same and changes through time were largely the same (trends, not absolute levels). To me, this is a tough test for the SSB, and I recall being really impressed with how qualitatively close the SSB was to the GSF. An example of where things sort of work, but it is much harder to draw qualitative conclusions from the SSB, is the work onfine volatility. Volatility is a measure [of] short-run earnings [in-]stability. On the one hand, though the level of volatility is a lot higher in the SSB than the GSF, overall trends are remarkably similar. Subgroup trends by gender and race are also similar between the SSB and the GSF. Trends by education, if I recall correctly, are different enough that qualitative conclusions change between the SSB and the GSF. It's also difficult to draw conclusions from decomposition models that try to separate short-run instability from long-run trends because transitory inequality is much higher in the SSB, and the measure of earnings used in much of the related literature relies on knowing an individual's level of education (I'll explain this below). Also, it's impossible to compare estimates on the SSB to other datasets because of the substantial difference in levels between the SSB and the GSF. The upshot is that results on the SSB work fine for specification and sample testing, and for drawing most qualitative conclusions that rely only on SSB/GSF estimates, but are not useful at all for comparing those results to other results and point estimates [which] are substantially different. An example of where things have not worked is our work on the effect of access to TDI on labor market outcomes. This work relies on a set of diff-in-diff regressions comparing people with/without a work limiting disability in/out of states with a TDI program. We've had two problems in this context. The main issue here is that point estimates have changed signs (e.g., if I recall correctly the SSB has positive effects for women and negative for men, while the GSF shows the opposite). With this project, we really haven't been able to draw any conclusions at all from the SSB. Like volatility by education and the decomposition models, this work also relies on SIPP survey information that cannot be treated as fixed prior to the individual's SIPP panel, in addition to the admin data. In summary, I think there are two basic issues. First is whether a sample uses SIPP panel characteristics that can be treated as time invariant both before and after an individual's SIPP panel. The GSF links all SIPP participants, including kids. The SSB synthesizes the pooled SIPP panels, including those kids. Volatility by race largely works because race can be treated as fixed prospectively and retrospectively, so those kids age into the sample. Volatility by education doesn't work because it can only be treated as fixed after a SIPP panel, so kids can't age into the sample. Our mobility estimates overall and by gender worked better than those by education for the same reason. The second is that, unless you're comparing means of variables using the full GSF, comparing point estimates in the SSB to any other dataset will likely be problematic. So, in our experience, it's usually possible to do all sorts of specification/sample/method testing on the

Copyright National Academy of Sciences. All rights reserved.

SSB and expect those changes to effect the GSF in the same way, but it may take a number of disclosures before you can know whether this is true. But, there hasn't been any situation in which we can rely on point estimates in the SSB given the type of work we're doing. And, again, it would take a number of disclosures before anyone could learn where their specific work can rely on SSB point estimates."

- "The coefficients were mostly of the same sign and significance, but were more precise in validated results."
- "There were differences in estimating how the variance of log earnings changes had evolved over time—in particular around the 2007 recession."
- "Age was a key variable in constructing microdata results; birthdate variable was very difficult to use."
- "Some barely significant differences (in the synthetic data) in earnings across groups became more significant in the GSF results. This was particularly true when looking at more narrowly-defined groups (e.g., by sex and education) where sample sizes became smaller."
- "Certainly, differences did appear. But keeping in mind that the SSB was not meant to be scientifically valid for all use cases on its own, the SSB results were similar enough to the GSF results that I was able to still build my samples and analyses using the synthetic data."
- "I am still waiting for validation."
- "Our results were qualitatively unchanged."

DIFFICULTIES ENCOUNTERED WHEN USING FSRDCS

Only one of the seven FSRDC users reported having difficulties, expressing the following issues: a time delay in getting clearance for access, unclear documentation, and the file(s) did not have the measures that were needed.

IMPACT OF DIFFICULTIES

Describing the impacts resulting from their difficulties in accessing the data, five gave up on obtaining access, including three who had difficulty accessing the public-use files (Figure E-9). Four said their research was never completed, and three said there was a time delay in obtaining access. Notably, all three who said there was a time delay were users of the public-use file, which can be downloaded, so they were not referring to clearance procedures.

Copyright National Academy of Sciences. All rights reserved.

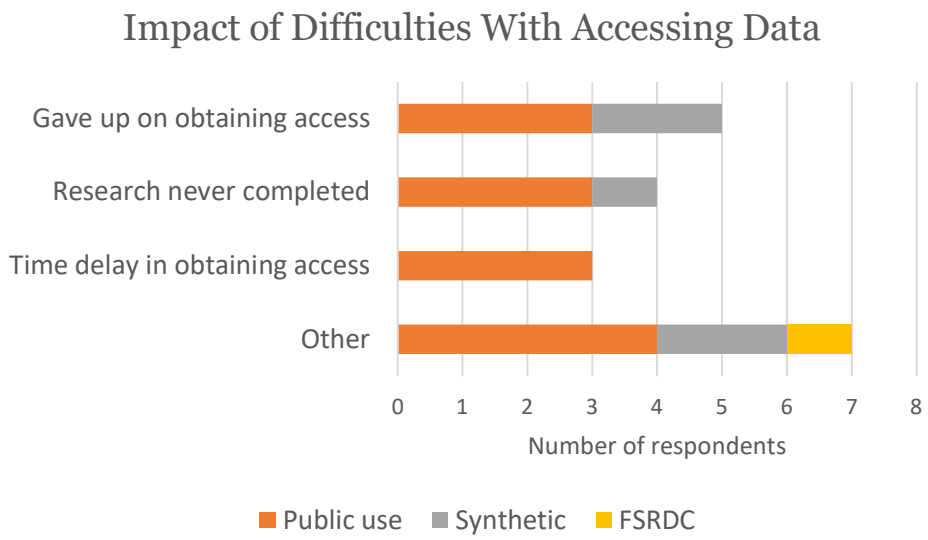


Figure E-9 Impact of encountering difficulties with accessing SIPP data.

Following are the text comments for those who responded "Other."

- SSB: "Data not available now."
- SSB: "Delayed our analysis of results."
- Public-use: "Scaled down our research project to only recent panels."
- FSRDC: "Adjusted measures used in research."
- Public-use: "Delay research progress."
- Public-use: "Incomplete analysis."
- Public-use: "Time delay in using the data."

HOW THE DATA WERE USED

Eighteen respondents reported having published a journal article or report as a result of using the SIPP data, 6 reported the results supported the public policy process, and 5 reported that the data were used to make projections of program or policy impacts (Figure E-10). Some respondents reported multiple uses: one reported both published articles/reports and supporting the policy decision process, one reported both publishing and making projections, and 2 reported all three. Nineteen respondents reported no results yet from their use of the SIPP data.

Copyright National Academy of Sciences. All rights reserved.

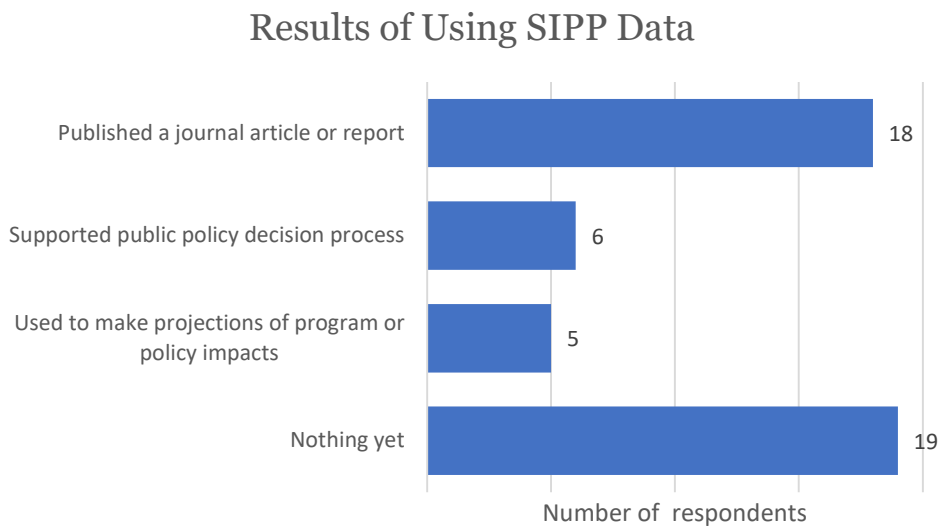


Figure E-10 How the results from SIPP data were used.

Those 18 respondents who reported publishing articles or reports based on SIPP also reported publishing across a wide span of areas. Most often they reported publishing on earnings, employment, survey methodology, and sociology (Figure E-11). Of the 17 fields listed on the questionnaire, respondents reported publishing in 14 of them, with each publishing in one to 6 fields.[76] Among the 18 who reported publishing, 16 supplied the number of articles they published, for a total of 57 articles or reports, a mean of 3.2 per respondent.

[76]The respondent checking "Other" wrote in "NA."

Copyright National Academy of Sciences. All rights reserved.

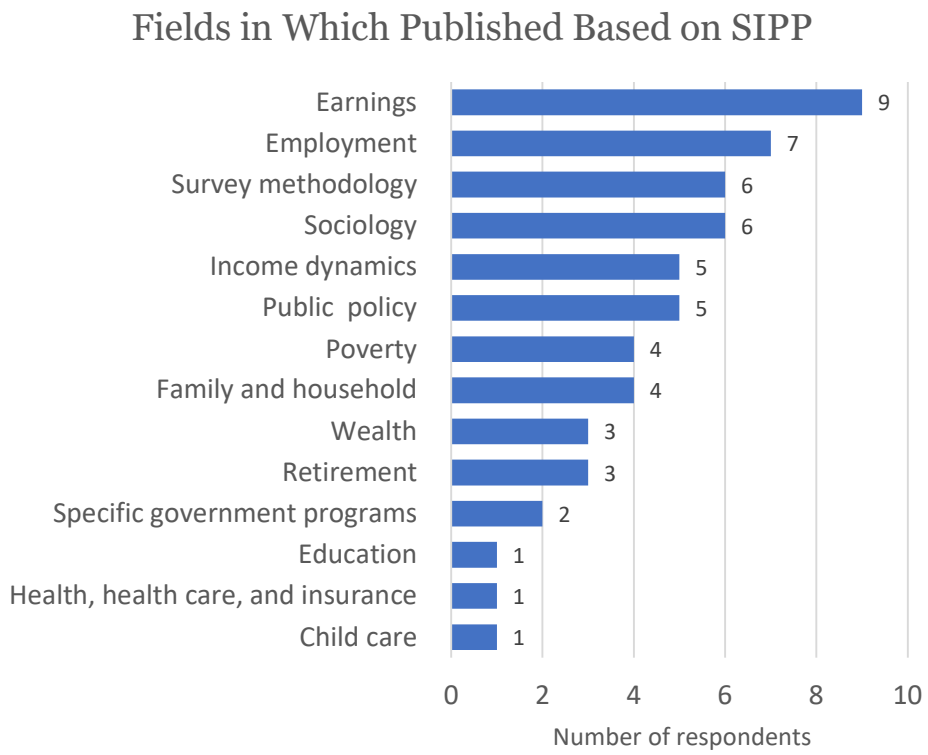


Figure E-11 Fields in which SIPP data findings were published.

WHETHER RESEARCH NEEDS COULD BE MET THROUGH STANDARDIZED TABLES

Only 7 percent of the respondents said their research needs could largely be met through the availability of a comprehensive set of standardized tables (Figure E-12). By contrast, roughly three-fourths said their research needs would be met only to a little extent (31 percent) or no extent (43 percent). These results are consistent with the results above that the respondents often engaged in statistical modelling rather than depending only on tabulations. However, these results may be biased in the sense that those most heavily involved in using SIPP data, and having the greatest vested interest in access to the data, may have been the most likely to respond. There may be a pool of less committed SIPP users, or of potential future users, if the data were more readily accessible, people who are not well represented in these results.

Copyright National Academy of Sciences. All rights reserved.

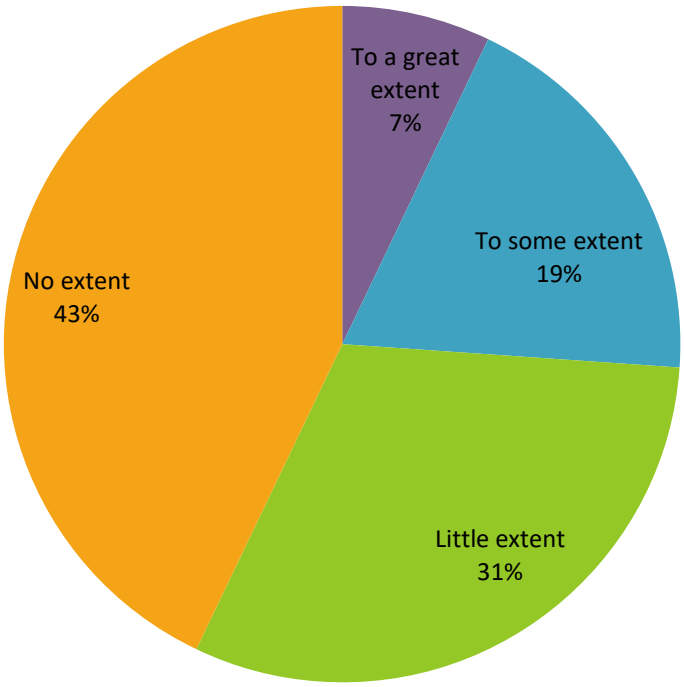


Figure E-12 Degree to which SIPP findings could be met by standardized tables.

OTHER COMMENTS BY SIPP USERS

The respondents were also invited to provide open-ended text comments that they felt would be helpful to the panel. Eighteen respondents provided comments. These are presented verbatim below.

Ease of Use

- "I watched the Census Bureau's webinar series about using SIPP data, and those webinars helped me a lot. But it was a steep learning curve to gain even basic skills accessing and interpreting SIPP data. I would like to be able to create state-level tabulations using SIPP data."
- "The variables showing receipt of housing subsidies are not internally consistent across waves of the same panel. The editing and imputation procedures that the SIPP uses produce inconsistent four-months spells of subsidized housing receipts. This makes the data very hard to use for any research on housing subsidies. This problem does not exist for the 1996 Panel."
- "Please provide example code/syntax for various statistical software programs using the synthetic data. Please include limitations that should be considered, especially if subpopulation analyses could be problematic."

Copyright National Academy of Sciences. All rights reserved.

- "The biggest issues I've had with the SIPP are important changes across panels that are poorly (or even incorrectly) documented, and general documentation, website, and data presentation not being user friendly (and especially not beginner friendly)."

Value of the Data

- "I had hoped to use the tax module, but it was almost empty. This was many years ago, I don't know if the situation has improved."
- "The information was sound; the income in 2020 compares to now."
- "Learning tool for me. And it's been amazing and enlightening."
- "I am relying on the SIPP Synthetic Beta (SSB) for the second paper in my dissertation and currently access the file through the VRDC (CED2AR) at Cornell."
- "Both the public-use SIPP and Synthetic SIPP have become invaluable tools in my research. I am very sad to hear we may be without the SSB for some time. The W-2 earnings records in the SSB are priceless for my work on earnings dynamics. I hope the SSB comes back and perhaps with even more merged/generated variables from the public-use SIPP."
- "I was interested in correlations in income across time with different lag periods."
- "Data coding of categories across panels. No newer data beyond 2008 panel for Informal Care (adult caregiving)."

Accuracy of the Data with Respect to Disclosure Avoidance

- "When considering the probability of reidentification, also consider how complex the SIPP data are (and how high the learning curve for using it), and all of the uncertainty introduced from editing the data. In general, it seems that everyone assumes that respondents respond honestly and correctly to every question, and the public data reflect that. While I do think the data are of high quality, I suspect there is an additional layer of protection introduced just through human error and data processing that is typically ignored by those concerned about disclosure."
- "The risks of disclosure avoidance have to be balanced with the benefits of research insights. Over-weighting disclosure avoidance means that only a few researchers can have access to these important datasets. This would undermine everyone's understanding of important socioeconomic questions and increases inequality in research (only the very well-established, well-connected researchers will have data). A lot of our staff already have Census Sworn Status among other certifications. So there are ways to virtually eliminate disclosure issues without limiting access. Access to microdata is crucial to our work as published tabulations are virtually useless to the work that we do."
- "Many data users fail to appreciate the amount of error that is already in the data; both error due to traditional survey error components and error for disclosure avoidance. It is important that data

Copyright National Academy of Sciences. All rights reserved.

users understand that changing/strengthening disclosure avoidance does not mean the survey is changing from 'the truth' to 'not the truth.' "

Providing Access to the Data

- "I think the GSF is the single most important 'non-standard' data product that Census has produced because it provides access to the types of administrative data that are increasingly being considered 'the best' thing for researchers. It is of course debatable whether this is true, but as long as people continue to believe that it's true, it is critical for as many researchers as possible to have access to these types of data. The current way of providing access to the GSF is not sustainable. It seems like some amount of automation is critical. My experience working with the SSB/GSF suggests that it is unrealistic to think that a synthetic SIPP would be a suitable substitute to the real thing."
- "I would be happy to comply with a more detailed onboarding process + disclosure process if I got to see the microdata in better detail. The current tradeoff is difficult to use and often requires repeated interfacing with Census staff to get things right; it would be easier to access better data remotely, with a more comprehensive data disclosure process at the end."
- "I think individual data is necessary for statistical analysis, which is why I am not sure that comprehensive summary tables are a great solution. I would love for the data to be harmonized so it is easier to use several panels of the SIPP together. Especially because the codebooks prior to 2014 are not in such a user-friendly form and with all the variable name changes, it is quite an undertaking to merge datasets."
- "I had a great experience using the SSB. Its synthetic nature didn't cause any problems for either of the projects I used it for. If researchers are no longer going to be allowed access to the full public-use SIPP, I'd rather have to use something like the SSB with a validation step, than only be given aggregated data."

Copyright National Academy of Sciences. All rights reserved.

Appendix F
Biographical Sketches of the Committee

TRIVELLORE RAGHUNATHAN is a professor of biostatistics in the School of Public Health and a research professor at the Institute for Social Research, both at the University of Michigan. He is also a member of the research faculty at the Joint Program in Survey Methodology at the University of Maryland. Raghunathan is a fellow of the American Statistical Association, and he received the Remington award from the American Heart Association, SPAIG and Monroe Sirken Awards, both from American Statistical Association. He has served on the National Academies' panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods and the panel to Review Programs of the Bureau of Justice Statistics. Raghunathan served as chair of the department of biostatistics and director of the survey research center. His research interests are in the analysis of incomplete data, multiple imputation, Bayesian methods, design and analysis of sample surveys, small area estimation, confidentiality and disclosure limitation, longitudinal data analysis and statistical methods for epidemiology. He also developed a free software (IVEware) for implementing sequential regression approach for multiple imputations which can be downloaded from the website www.iveware.org. He received MS degrees from both Miami University and Nagpur University, India, and a PhD degree from Harvard University.

SCOTT H. HOLAN is professor of statistics at the University of Missouri, and is co- editor-in-chief of the International Statistical Review. He currently serves on an Intergovernmental Personnel Act Agreement as a senior research fellow at U.S. Census in the Office of the Associate Director for Research and Methodology (Research and Methodology Directorate), where he mainly conducts research on various uses of multivariate spatio-temporal modelling, such as applications to the American Community Survey and the 2020 Census. Holan is an elected fellow of the American Statistical Association, an elected member of the International Statistical Institute, and an elected fellow of the Institute of Mathematical Statistics. His research expertise includes dependent data modelling (spatial, spatio-temporal, functional, and multivariate, among others), Bayesian methods, official statistics, and survey methodology. Holan previously was co-awarded the Statistical Partnerships Among Academe, Industry, and Government Award . He has previously served on a National Academy of Science Panel for Addressing Priority Technical Issues for the Next Decade of the American Community Survey and a Steering Committee for Review of Confidentiality Criteria for Survey of Earned Doctorate Data. Holan has a PhD in statistics from Texas A&M University.

V. JOSEPH HOTZ is the Arts and Sciences Distinguished Professor of economics and public policy at Duke University. Prior to his current appointment, he held faculty positions at the University of California, Los Angeles, the University of Chicago, and Carnegie Mellon University. Hotz is an elected fellow of the Econometric Society, the Society of Labor Economists, and the International Association of Applied Econometrics. His research areas include the economics of the family, economic demography, labor economics, population health, and applied econometrics. Hotz previously served as the chair of the Population Association of America's Committee on Population Statistics, as a member of the NORC New National Longitudinal Survey Assessments Oversight Committee., as a member of the Committee on National Statistics, and he has served on many National Academies consensus panels and workshop planning committees; including the planning committee for the Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations, which he co-chaired. Hotz received his PhD in economics from the University of Wisconsin–Madison.

Copyright National Academy of Sciences. All rights reserved.

THOMAS KRENZKE is vice president and senior statistical fellow in Westat's Statistics and Data Science. He currently leads Westat's Confidentiality Work Group and he serves on the Westat Institutional Review Board. Krenzke is a fellow of the American Statistical Association (ASA), and he is the Council of Chapters Governing Board District 2 vice chair. He was chair of the ASA Committee on Privacy and Confidentiality, and the Washington Statistical Society president. Krenzke leads research in statistical confidentiality, sample design, non-probability sampling, small area estimation, variance estimation, imputation, and nonresponse bias. He adds new statistical capabilities by developing software for disclosure avoidance; nonresponse bias analysis; area sampling unit formation, stratification, and selection; classification trees; generalized regression estimation; and imputation. Krenzke holds an MS in applied statistics from Bowling Green State University.

FANG LIU is professor of applied and computational mathematics and statistics (ACMS) and associate chair of the ACMS department at the University of Notre Dame. At the University of Notre Dame, she is also affiliated with the Harper Cancer Research Institute, the Eck Institute of Global Health, the Technology Ethics Center, and the Lucy Family Institute for Data and Society. Liu is the co-editor of the 'O Privacy, Where Art Thou' column for CHANCE, and is associate editor of Transactions on Data Privacy, founding co-editor-in-chief of ACM Transactions on Probabilistic Machine Learning. Liu is an elected fellow of the American Statistical Association. She serves on the Committee of Privacy and Confidentiality of the American Statistical Association. Liu's research focuses on the development and application of modern approaches in statistical machine learning and data privacy protection. She also develops and applies statistical methodologies, especially Bayesian statistics, to analyze data originating from the medical, biological, and social sciences. This research is mainly funded by the National Science Foundation, National Institutes of Health, Gates Foundation, and UNITAID. Liu holds a BS from Peking University, an MS from Iowa State University, and a PhD from the University of Michigan, Ann Arbor.

ROBERT A. MOFFITT is Krieger-Eisenhower Professor of Economics at Johns Hopkins University and holds a joint appointment in the Johns Hopkins Bloomberg School of Public Health. He is a member of the National Academy of Sciences, a fellow of the American Academy of Arts and Sciences, a fellow of the Econometric Society, a fellow of the Society of Labor Economists, and a recipient of a Guggenheim Fellowship and a National Institute of Health MERIT award. Moffitt serves on the Advisory Board for a new synthetic data file of tax information for the Urban Institute and the U.S. Department of Treasury. His research interests include social policy, welfare programs, poverty, family structure, labor markets, income volatility, and applied statistical methods. He is a past president of the Society Labor Economics and of the Population Association of America, and was chief editor of the *American Economic Review*, coeditor of the *Review of Economics and Statistics*, and chief editor of the *Journal of Human Resources*. Moffitt taught at Rutgers University and Brown University before his current teaching position at Johns Hopkins University. He holds a BA in economics from Rice University and a PhD in economics from Brown University.

AMY PIENTA (she/her) is a research professor and director of business and collection development at the Inter-university Consortium for Political and Social Research (ICPSR). She is also a research/faculty affiliate of the Michigan Center on the Demography of Aging, the Population Studies Center, the Michigan Institute for Data Science, and the Center for Global Health Equity at the University of Michigan. At ICPSR, Pienta directs several federally-funded data infrastructure projects that provide broad dissemination of confidential and/or restricted-use data using various modes of secure access. She chairs the research advisory board for the Qualitative Data Repository at Syracuse University, and co-chaired the social sciences subsection of the COVID-19 Working Group within the Research Data Alliance which includes establishing disclosure review practices for COVID-19 data. Pienta earned her Ph.D. in sociology from SUNY Buffalo and completed a National Institute on Aging postdoctoral fellowship in the Population Research Institute at the Pennsylvania State University.

NATALIE SHLOMO is professor of social statistics at the University of Manchester. She is currently the president of the International Association of Survey Statisticians 2023-2025. She also serves on editorial boards and is a member of scientific methodology advisory committees for National Statistics Institutes, currently for the United Kingdom, Italy, and Canada. She is a fellow of the Academy of Social Sciences in the United Kingdom, an elected member of the International Statistical Institute and a Fellow of the Royal Statistical Society. Before that, Shlomo was senior lecturer in the School of Social Sciences at the University of Southampton and senior methodologist at

Copyright National Academy of Sciences. All rights reserved.

the Israel Central Bureau of Statistics. She publishes widely in the area of survey statistics, including small area estimation, adaptive survey designs, non-probability sampling, confidentiality and privacy, data linkage and integration. Shlomo has a BS in mathematics and statistics and a MA and PhD in statistics from the Hebrew University in Israel.

ALEKSANDRA (SEŠA) SLAVKOVIĆ is a professor of statistics & public health sciences, and associate dean for graduate education in Eberly College of Science at Penn State. She is associate editor of the *Annals of Applied Statistics* and *Journal of Privacy and Confidentiality*, and editor for *Statistics and Public Policy*. Slavković is a fellow of the American Statistical Association (ASA), Institute of Mathematical Statistics, and the International Statistical Institute. Her research focuses on methodological developments in the area of data privacy and confidentiality in the context of small- and large-scale surveys, health, genomic, and network data, including differential privacy and broad data access that offers guarantees of accurate statistical inference needed to support reliable science and policy. She served as a chair of the ASA's Privacy and Confidentiality committee and ASA's Social Statistics Section. Slavković served on the National Academies of Sciences, Engineering, and Medicine's Committee on National Statistics' Panel on Measuring and Collecting Pay Information from U.S. Employers by Gender, Race, and National Origin, and the National Research Council's Transportation Research Board. She held visiting scholar positions at Cornell University, University of California Berkeley, University of Minnesota and Utrecht University. Slavković received her BA in psychology from Duquesne University, her MS and PhD in statistics from Carnegie Mellon University, and a MS in human-computer interaction from Carnegie Mellon University.

HEEJU SOHN (she/they) is currently an assistant professor of sociology at Emory University. Her research examines how unequal social safety-nets and kin structures exacerbates health and social inequities in the United States. An ongoing project quantified key outcomes among undocumented immigrants living in the United States by leveraging data from the Survey of Income and Program Participation, Los Angeles Family and Neighborhood Survey, and the Survey of Migration at the North Border of Mexico (EMIF, [Encuesta sobre Migración en la Frontera Norte de México]). Sohn has also worked as an expert consultant to the United Nations on estimating mortality patterns in data-sparse contexts such as armed conflicts and severe disasters. She received her MEng in information engineering from Cornell University and PhD in demography and sociology from University of Pennsylvania. Sohn completed her National Institutes of Health-funded postdoctoral training at UCLA's Health Policy and Management and California Center for Population Research.

SALIL VADHAN (he/they) is the Vicky Joseph Professor of computer science and applied mathematics in the Harvard School of Engineering and Applied Sciences. For the past decade, he has led the Harvard Privacy Tools Project, a broad effort to advance a multidisciplinary understanding of data privacy issues and build computational, statistical, legal, and policy tools to help address these issues in a variety of contexts. Vadhan is a Simons investigator and an Association for Computing Machinery fellow. He has received many awards including a Godel Prize, a Guggenheim Fellowship, and a Bowden Award for Outstanding Research in Privacy-Enhancing Technologies. Vadhan serves without compensation on two advisory boards: the Validation Server Advisory Board at the Urban Institute and for the Abdul Latif Jameel Poverty Action Lab Innovations in Data and Experiments for Action Initiative. He co-founded and co-directs OpenDP, a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data with the strong protections of differential privacy. He received his PhD in applied mathematics from MIT.

JENNIFER VAN HOOK (she/her) is Roy C. Buck Professor of sociology and demography at the Pennsylvania State University, and non-resident fellow at the Migration Policy Institute. Her research focuses on the demographics of immigrant populations, estimates of the size and composition of the unauthorized foreign-born population, and the socioeconomic integration of immigrants and their children, including several articles that relied on data collected by the Survey of Income and Program Participation. Van Hook received the Clifford C. Clogg Award for Mid-Career Achievement and was elected to the Sociological Research Association. She has served as the vice president of the Population Association of America. Van Hook served on the National Academy of

Copyright National Academy of Sciences. All rights reserved.

Sciences, Engineering, and Medicine's panel Reengineering the SIPP. She received her PhD in sociology from the University of Texas at Austin.

Copyright National Academy of Sciences. All rights reserved.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics was established in 1972 at the National Academies of Sciences, Engineering, and Medicine to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant, a National Agricultural Statistics Service cooperative agreement, and several individual contracts.

Copyright National Academy of Sciences. All rights reserved.