

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

What Are the Implications for SIPP Data Users?

In 2022–2023, the Committee on National Statistics within the National Academies of Sciences, Engineering, and Medicine convened a panel of experts to examine disclosure avoidance in the Survey of Income and Program Participation (SIPP). Following are some of the key implications of the report for SIPP data users.

DO THE SIPP DATA CURRENTLY PRESENT DISCLOSURE RISKS?

A recent re-identification study conducted by the Census Bureau found no disclosure risks that required immediate action. However, additional studies are needed, and there is a prospect that disclosure risks will be identified (Conclusions 3–5 and 3–6). In the case of SIPP, the initial study may underestimate risk since all of the variables and databases open to data intruders were not considered, risks were estimated only for primary respondents, and analysts did not consider the full dimensionality of the data, in which both data on other members of the household and data on change over time might be used to help identify respondents. For example, interracial marriages are relatively uncommon, as also are large age differences between spouses; thus, knowing that a black male is married to an Asian female who is 12 years younger helps to quickly narrow the number of possible matches.

WILL SIPP DATA USERS CONTINUE TO HAVE ACCESS TO A SIPP PUBLIC-USE DATAFILE?

The panel recommended continued release of a public-use datafile because data users need the capacity to

perform complex and highly customized analyses (Recommendation 9–3). However, there may be some changes to the public-use file.

- In the short term, there may be value in re-coding certain highly granular data to have fewer distinct values. Possible candidates for such recoding are the measures of age (while keeping key cutoff ages for determining eligibility for program participation) and of state (perhaps instead providing geographic region).
- Another short-term possibility is to require the acceptance of a user agreement in which the data user agrees not to identify individual respondents.
- In the longer term, highly disclosive and sensitive variables might be replaced with synthetic versions of those same variables.

WHAT CAN SIPP DATA USERS DO IF THEY NEED MORE DETAILED DATA THAN IS PROVIDED IN THE PUBLIC-USE FILE?

The panel recommended that the Census Bureau move towards allowing more tiers of analysis (Recommendation 9–4). In particular, offering secure online data access, but with fewer restrictions than those currently imposed for access to Federal Statistical Research Data Centers (FSRDCs), seems to offer the best prospects for allowing the complex and highly

customized analyses that are often performed by SIPP data users (Recommendation 5-1). Another prospect is to create a remote analysis platform in which researchers can obtain statistical estimates without having direct access to the raw data; however, a simplified table generator would not meet the needs of many SIP data users, and a full remote analysis platform is likely to require several years of development work (Recommendation 7-1).

HOW DOES SECURE ONLINE DATA ACCESS DIFFER FROM WORKING WITH A PUBLIC-USE FILE?

There are multiple ways of implementing secure online data access, but some common elements are the submission of a research proposal showing how the data would be used, the acceptance of a user agreement on how the data will be used, requirements for a secure data environment, and limitations on what data may be published or released.

WHAT HAS HAPPENED THAT MIGHT LEAD TO CHANGES IN CENSUS BUREAU PROCEDURES?

Once data were highly dispersed, sometimes requiring access to physical records in multiple locations. With data increasingly becoming digitized and becoming available online, data intruders have ready access to a wide variety of data, and the amount and types of such data continue to grow. Further, advances in computing technology, software tools, and internet search capabilities all have combined to enable the processing of large amounts of data from diverse sources in ways that might allow the identification of survey respondents.

The Census Bureau has long had a legal and ethical responsibility to protect the confidentiality of its survey respondents, but that task has become more difficult. At the same time, new approaches for providing data access while also protecting data are also being developed, and these can be incorporated into new strategies for protecting confidentiality.

WHAT DO I NEED TO KNOW TO WORK WITH PRIVACY-PROTECTED DATA?

A key decision at the start is deciding which tier of access you need to perform your research. A table generator or remote analysis platform may be the easiest way to access the data, and may be especially useful for preliminary work in developing a research proposal and data analysis strategy. Public-use files will meet the needs of many researchers, but you will need to check if they have the specific variables and level of detail that you need. Secure online data access will provide more complete access to the SIPP data, and access through FSRDCs may be needed if you wish to combine the data from SIPP with data from other sources.

Certain data protection tools also have implications for how analysis is conducted. Data analysis based on synthetic data may need to be verified through comparisons with the original data. Data with differential privacy protections may require special tools for estimating standard errors.

The panel recommended that the Census Bureau release documentation that will help data users to make these kinds of decisions and perform the analyses (Conclusions 4-4 and 10-1; Recommendation 3-2).

FOR MORE INFORMATION

This Consensus Study Report Issue Brief was prepared by the Committee on National Statistics and based on the Consensus Study Report *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (2023)*. The study was sponsored by the U.S. Census Bureau. 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. The Consensus Study Report is available from the National Academies Press, (800) 624-6242 or <https://www.nap.edu/catalog/27169>.

Division of Behavioral and Social Sciences and Education

**NATIONAL
ACADEMIES** Sciences
Engineering
Medicine

Copyright 2024 by the National Academy of Sciences. All rights reserved.