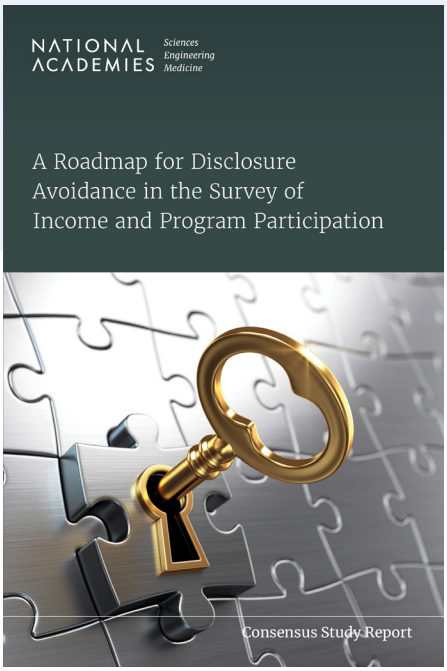
A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation



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 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; however, they also present challenges in maintaining the confidentiality of its survey respondents.

Protecting the privacy of survey respondents is a core responsibility of the Census Bureau. It is a legal requirement, as specified in Title 13; a moral requirement based on the promises given to the survey respondents; and possibly a key requirement in getting people to respond to the survey and give honest answers. This may especially be true of SIPP, which collects highly personal information such as data on incomes, assets, liabilities, and assistance received through governmental programs.

At the same time, protecting privacy is becoming increasingly difficult because numerous databases exist with personal identifying information that collectively contain data on household finances, home values, purchasing behavior, and other SIPP-relevant characteristics. The digitization of records has made data available that once would have required searches through paper records scattered across multiple locations. Software and internet tools have become available to allow data intruders to identify, collect, and combine data to create highly comprehensive databases. These sources could be used by potential intruders to identify a respondent in SIPP. In response to this growing threat, new tools are also being developed to protect privacy.

The Census Bureau asked the National Academies of Sciences, Engineering, and Medicine to convene a panel of experts to formulate a roadmap for making data from SIPP available to researchers and policymakers while protecting the confidentiality of survey respondents. It asked the panel to consider factors such as evolving privacy risks, the development of new methods for protecting privacy, the nature of the data collected through SIPP, the practice of linking SIPP data with administrative data, the types of data products produced, and the desire to provide timely access to SIPP data.

MEASURING THE CURRENT LEVEL OF RISK

The best current measure of disclosure risk for SIPP is a re-identification study recently conducted by the Census Bureau, in which it sought to match the SIPP 2014 data against IRS data from 2013 to 2016 based on 36 indirect identifying variables. The study found that age was the most disclosive variable among those that were used, and concluded that no changes are currently needed to the public-use file.

The panel was impressed by the work performed by the Census Bureau, but determined that more analyses are required to properly assess the disclosure risk. In particular, there should be analyses looking at the capacity provided by SIPP to examine changes in the data over time, the capacity to measure how combining information across multiple members of a household might be disclosive, the inclusion of potential matching variables that are in other databases but not in the tax data, the use of a more comprehensive database than only taxpayers, and the use of lower thresholds when identifying putative risks.

Based on data contained in the 2020 SIPP, the panel found that using a combination of between five and eight personal features (sex, race, ethnicity, state, birth year, highest level of education, number of children, and change in the number of children during 2019) was sufficient to uniquely characterize most panel respondents, particularly if data on multiple household members or on changes over time were included. Because uniqueness within the SIPP panel is quite different from uniqueness within the U.S. population, additional re-identification studies are needed to measure the actual level of risk. However, based on the panel's findings, and considering that still additional identifying characteristics could be added, the SIPP public-use file may put some respondents at risk of disclosure.

METHODS OF REDUCING DISCLOSURE RISKS

The panel sought to balance minimizing the risk of disclosure against allowing researchers and policymakers to have timely access to data that support valid inferences. After examining the most recent and the most often cited research in SIPP, the panel concluded that researchers often perform complex analyses that depend on the high granularity of the SIPP data, and that collectively use a wide range of variables. The two simplest solutions—coarsening the data or reducing the number of variables—would frustrate much of the research performed with SIPP. The panel concluded that it is important to continue providing a public-use file. At the same time, recognizing that additional limits are likely to be required (e.g., by dropping or coarsening measures such as age and state), there needs to be another way of providing the data that is readily accessible for analysis. Currently, the only alternative to using the public-use file is to access the data through Federal Statistical Research Data Centers (FSRDCs), but obtaining such access is both difficult and expensive, creating inequities in access to data and likely discouraging many researchers from using SIPP.

The panel recommends three broad approaches for limiting disclosure without hampering access.

- First is the creation of secure online data access (SODA). FSRDCs provide a type of SODA, and FSRDC access may still be needed for researchers wanting to access certain SIPP data or to merge the SIPP data with other databases. But as envisioned by the panel, SODA would have a less difficult application and approval process and would be less costly. The panel believes this is possible because the current process used for FSRDCs is more restrictive than is actually required by Title 13, and Title 13 can be reinterpreted based on more recent legislation that places a priority on evidence building and supporting multiple tiers of access.

- The second approach is to provide access to data through a remote analysis platform (an expanded version of a table generator). Creating such a platform that could support the complex analyses often performed on SIPP is likely to be difficult and time consuming.

- The third approach is to develop public-use files that are synthetic datasets, with or without differential privacy, along with automated verification and validation systems. The current state of technology for both creating synthetic data and for applying differential privacy is not adequate for handling all the variables in a file with the size and complexity of the SIPP data.

It will take time to develop SODA, and even more time to develop a remote analysis platform and for the creation of synthetic datasets with automated verification and validation systems. In the meantime, the Census Bureau may want to consider interim measures to lessen the risk of disclosure. These might include traditional statistical disclosure limitation methods such as coarsening the data, requiring users to register to download the public-use file, and requiring data users to accept a User Agreement that limits their use of the data. One might consider an incremental approach in which a few variables are synthesized for inclusion in the public-use file with or without differential privacy. Another possibility is to create a remote analysis platform or table generator with built-in differential privacy for the variables that are too granular in the public-use file. However, these approaches currently would provide too limited capabilities to constitute a general solution. These short-term solutions do not provide the full extent of protection needed, but will increase awareness of the issues and provide some protection. It is important to minimize any loss of data utility when taking these steps.

PANEL ON 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

COMMITTEE ON NATIONAL STATISTICS **Brad Chaney,** Study Director; **David Johnson,** Senior Program Officer; **Nancy Kirkendall,** Senior Program Officer; **Madeleine Goedicke,** Senior Program Assistant; **Joshua Lang,** Senior Program Assistant

FOR MORE INFORMATION

This Consensus Study Report Highlights was prepared by the Committee on National Statistics 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*