Allocating Microdata from the National Longitudinal Survey of Youth Among Access Tiers: A Framework for Decision-Making and Initial Investigations

By Alison Aughinbaugh, Keenan Dworak-Fisher, Donna Rothstein, and Julie Yates

Proposal for Research to be Presented to NBER Conference on "Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences"

Previous cohorts of the National Longitudinal Survey of Youth (NLSY) have employed a tiered access approach to provision microdata in a way that balances the needs of its users with the imperative to safeguard the privacy of its respondents. Both the NLSY79 and the NLSY97 have maintained extensive public use data sets while providing access to additional, restricted data to qualified researchers who agree to act as temporary agents of the Bureau of Labor Statistics (BLS). As BLS is preparing to field a new NLSY cohort, it may reconsider its allocation of microdata between its public- and restricted-use datasets. This study will provide a framework for deciding how the allocations should be made and report on initial investigations as to how the framework could be applied.

This work is informed by the ongoing effort by Federal Statistical agencies to both: a) increase their effective use of survey and non-survey data for statistical purposes and b) redouble their commitment to the privacy of the data's subjects. The reports of the Commission on Evidence-Based Policymaking (CEP) and the Advisory Committee on Data for Evidence Building (ACDEB) have offered careful analyses of the inherent tradeoffs, and the Foundations of Evidence-Based Policymaking Act of 2018 (Evidence Act) has provided a foothold in the Federal Law to implement several of their recommendations, including the provision of separate tiers of access for data of varying degrees of sensitivity. As the Federal Statistical System has sought to implement the Evidence Act, it has sought to clarify the joint responsibilities of statistical agencies to balance their commitments to relevance, timeliness, accuracy, credibility and objectivity while protecting the trust of respondents by ensuring the confidentiality and exclusive statistical use of their data. Relatedly, the Federal Committee on Statistical Methodology (FCSM) has developed a data quality framework that includes these different components.

Coincident with these efforts, statistical agencies have advanced the research on the application of a differential privacy approach to disclosure limitation. This approach focuses specifically on the susceptibility of data programs to reidentification attacks relying only on the information provided by the data programs themselves. However, as Hotz *et al* (2022) has pointed out, these methods do not account well for al sources of reidentification risk, and they don't, by themselves, weigh the benefits of data provision against the potential costs.

In developing a framework to be applied to the allocation of data into public- and restricted-use components of NLSY data provision, we will focus on 3 broad considerations: 1) the legal requirements for protecting data privacy, especially as governed by the Privacy Act of 1974 and the Confidential Information Protection and Statistical Efficiency Act of 2018 (Title III of the Evidence Act); 2) the need to ensure the confidence of survey respondents and other providers of data to the Federal government; and 3) the research value of the microdata that is provided to users, which is itself determined by many dimensions of data quality such as accuracy, coherence, relevance, granularity, and accessibility. Although the application of these principles to a new set of microdata is likely to require subjective judgments, we will investigate how the framework can be made to work in practice by considering the data that has already been provided in previous NLSY cohorts. The analysis will include considerations of

the uniqueness of observed values, the closeness with which data may track with other readily available data sources that disclose identities, the sensitivity of the data to misuse, and the demonstration of research value implied by previous uses.

Bibliography

Abowd, John M. "Protecting the confidentiality of america's statistics: Adopting modern disclosure avoidance methods at the census bureau." *Census Blogs: Research Matters* (2018).

Advisory Committee on Data for Evidence Building: Year 2 Report, October 14, 2022. https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf

Federal Committee on Statistical Methodology. "A Framework for Data Quality." (September 2020). https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Hotz, V. Joseph, Christopher R. Bollinger, Tatiana Komarova, Charles F. Manski, Robert A. Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D. Spencer. "Balancing data privacy and usability in the federal statistical system." *Proceedings of the National Academy of Sciences* 119, no. 31 (2022): e2104906119.

Milner, Justin, Katharine Abraham, Ron Haskins, Karen Pittman, and Kathy Stack. "Realizing the Promise of Evidence-Based Policymaking." https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf