ATTACHMENT I – PROJECT TOPIC

Evaluation of Noise Infusion for the Survey of Doctorate Recipients

Key Objective

The objective of this research and development project is to evaluate the use of noise infusion for a demographic survey as a possible privacy-preserving method in the use of federally confidential data. Investigations will focus on use cases where noise infusion may be appropriate and use cases where noise infusion may introduce quality issues that reduce confidence in the use of estimates for decision-making. The Survey of Doctorate Recipients (SDR) is a sample survey that provides data on the characteristics of science, engineering, and health doctorate degree holders. The SDR provides data useful in assessing the supply and characteristics of U.S.-trained science, engineering, and health doctorates employed in educational institutions, private industry, professional organizations, and government in the U.S., as well as in other countries worldwide. To address disclosure concerns with the SDR while maximizing data utility, noise infusion is being explored as an alternative to augment other disclosure limitation methodologies currently in use with the SDR, both for the restricted-use data and the public-use microdata file. This project will inform the National Secure Data Service Demonstration Project (NSDS-D) by exploring noise infusion for a sample survey and assessing the disclosure protections and quality considerations for the resulting estimates.

Background

America's DataHub Consortium brings together capabilities and infrastructure to securely fill information gaps and to take on key analytic questions and evidence building challenges. As demand for access to confidential federal data assets increases alongside novel analytical approaches, privacy protections must be in place to ensure the protection of the privacy and confidentiality of the data. Noise infusion can be a powerful disclosure limitation methodology to protect the identity of respondents. Use of noise, however, presents concerns surrounding data quality. Balancing these two competing needs is critical in ensuring that high-quality data is available for decision-making while protecting the privacy and confidentiality of respondents.

The SDR is a sample survey of recipients of doctoral degrees in science, engineering, and health (SEH). It samples individuals who have earned an SEH research doctoral degree from a U.S. academic institution and are less than 76 years of age. The SDR is conducted biennially and collects demographic information of the respondent; educational history; employment status; field of degree; and occupation. The SDR biennial data collection allows for both cross-sectional and longitudinal analysis of the U.S.-trained doctorate population. To allow for this analysis, NCSES produces both cross-sectional microdata files and longitudinal microdata files. The cross-sectional data is a unique source of information about the educational and occupational achievements of U.S.-trained doctoral scientists and engineers in the United States and abroad. This project will focus on the cross-sectional restricted data and public-use file.

The restricted-use data are used internally by National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF) staff to produce aggregated statistics for the production of reports and information products for the public. In addition, the restricted-use data are available for use by researchers upon approval of a data request application, entering into a licensing agreement with NCSES, completion of annual data security training, and use of the data within a secure virtual enclave. Resulting products must meet disclosure standards and pass a disclosure review.

A public-use file (PUF) is also available for researchers who do not wish to request access to or do not need the restricted-use data. Currently, the SDR PUF and other publicly available SDR information products use multiple disclosure methodologies to protect the confidentiality of SDR respondents, including cell suppression, top-coding, and aggregation. While no known reidentification has occurred, the increasing risk of reidentification given the amount of publicly available data requires exploration of additional disclosure limitation methodologies. This project will explore the use noise infusion as an additional means to reduce the risk of re-identification in a sample survey.

This project will explore the following:

1. Multiple methodologies for noise infusion for both the restricted-use SDR data and the SDR Public-Use File (PUF), comparing the different methods with regards to data protection, resource needs, and ease of use, as well as resulting data quality and utility of estimates produced from the data. Impact on data quality and potential uses cases for noise infusion with the SDR, with a particular focus on the use of the estimates for evidence-based decision-making.
2. Assembling a Technical Expert Panel of 5-7 subject matter experts to evaluate the quality and fitness for use of the noise-infused data. This panel will include NCSES subject matter experts, as well as disclosure experts outside of NCSES. The offeror will organize and implement this review plan and produce a final report summarizing the panel feedback.
3. A strategy for messaging to data users the availability of noise-infused data if NCSES chooses to make the noise-infused dataset available within the restricted-data licensing program. This messaging will provide clarity on quality issues, use cases, and the utility of the noise-infused data for decision-making. In addition, this messaging should explain and emphasize the role of noise-infused data in the SDR's tiered access model by ensuring that public-use microdata can continue to be made available to researchers.

Information Gaps

This project will identify key components necessary to inform a future/potential NSDS including:

- Potential uses of noise infusion for a sample survey.
- Potential uses of noise infusion for cross-sectional microdata.
- Different methodologies for infusing noise and an evaluation of each.
- How to message noise infusion and the potential uses of estimates to researchers.

Key Evidence Building Considerations

- Key focus questions (address one or more) to assess innovation in the following areas: data security, privacy, and engagement:
    - Which novel techniques for data, privacy, and confidentiality protections can be used?
    - Are the resulting data and models fit to inform policy discussions and to make data available more equitability?

- What types of collaboration and stakeholder engagement are needed to help inform these questions, data, and analysis?
- What mechanisms are needed to access the resulting data that uphold privacy requirements?

Deliverables

At a minimum, offerors will provide the following if selected for an award. Additional deliverables may be required.

- Monthly report to document progress.
- Report detailing the results of the technical expert panel.
- All code, clearly documented; documentation of noise infusion methodology; documentation of data quality assessment; and any other data/documentation created under this award.