American Economic Association

Office of the Data Editor, Email: dataeditor@aeapubs.org

SI-FTAC Response to "*Request for Information To Improve Federal Scientific Integrity Policies*"

Dear Members of the Scientific Integrity Task Force,

My name is Lars Vilhuber, Executive Director of the Labor Dynamics Institute at Cornell University. I am an academic researcher, and I hold positions in various professional societies that relate to the topic you are tasked with. I am a member of the public and a taxpayer.

I am the inaugural and current Data Editor at the American Economic Association, and it is this role that I am responding to your Request for Information To Improve Federal Scientific Integrity Policies, issues on 2021-06-28.

In my role as Data Editor, I am tasked with the job of assessing and ensuring the integrity of the empirical and numerical results published in the AEA's eight journals. Most of the time, this entails ensuring that data is as broadly accessible as possible, because that is the simplest way in ensuring that many researchers can trust the results published in scientific articles. However, we also conduct active checks on the computational reproducibility of the code provided by authors, which requires that we, or a trusted third party, has active, short-term access to data, including data that is subject to legal or procedural access restrictions. In the past two years, my team and I have assessed over 1,000 articles that were conditionally accepted in the AEA's journals. I thus have an acute understanding of the many challenges that accessing data and computing resources for the purpose of integrity checks pose.

I am also the current Chair of the American Statistical Association's Committee on Privacy and Confidentiality, which is tasked with informing the ASA's membership of developments in the field of protecting the privacy and confidentiality of respondents in surveys and administrative data. One could argue that our job there is to make sure that our membership is aware of the tradeoff between privacy and access. In my response, I speak for neither the ASA nor for the committee, but draw on my experiences in that role.

As one of the editors and member of the governing board of the Journal of Privacy and Confidentiality, I have encouraged an informed but scientific discussion of the issues surrounding privacy protection, and of the latest technological and legal developments in that field. As such, I am quite aware of many of the tricky issues surrounding privacy and access. In my response, I do not speak for the journal's editorial or governing board, but draw on my experiences in that role.

I am also chair of the scientific advisory committee of the French research data access system, and on the board of the Canadian research data center network, both organizations tasked with the difficult job of providing secure but broad access to confidential data. My opinions here do not engage or constitute a position undertaken by these foreign institutions, but I draw on my experiences observing how other countries handle issues of data access, privacy, and integrity.

Finally, in the early 2000s, I was a leading member of the team that implemented the statistical data production and publication system underlying even today the Census Bureau's Quarterly Workforce Indicators. As such, I was acutely aware of the many challenges, but also opportunities, when attempting to make detailed and confidential data on the US workers available to the broadest possible audience, while ensuring transparency, traceability, and integrity of the statistical production process. I am not currently a member of that team, and I most definitely do not speak for the U.S. Census Bureau in any capacity.

> *(1) The effectiveness of Federal scientific integrity policies and needed areas of improvement; (2) good practices Federal agencies could adopt to improve scientific integrity (3) other topics or concerns that Federal scientific integrity policies should address.*

In the following, and in response to your request for information, I will highlight a few issues that should merit your attention.

1. Several federal statistical agencies have policies on scientific integrity, which are well laid out. They are not, however, universally adopted. I would encourage your taskforce to ensure that every **federal statistical agency adopt and publish explicit policies on scientific integrity**.
2. Policies define possible actions and activities, which need to be performed by federal staff. **These activities need to be funded**. Adherence to policies by federal staff is improved by making compliance with policies a part of job evaluations, and by providing staff with the resources (time, funds) and training to understand and support such policies.

3. **Predictability of access times** when data are not freely downloadable. When there are unavoidable application procedures, the process should be as transparent and predictable as possible. Public statements of processing times, public statistics on compliance with those processing times should be available. Importantly, such application procedures also must be suitably funded, so that compliance with stated deadlines is actionable, not wishful thinking.
4. **Simplification and standardization of access procedures, nomenclature, and legal basis across the federal government**, for instance by implementing streamlined and simplified application procedures when those are necessary (trusted researchers, enhanced legal foundations and mandates of access). The federal statistical system has a bewildering array of access procedures, ranging from click-through licenses, to legal agreements that need to be signed by requestor's organizations, to security clearances necessary to access highly secure facilities. At the AEA, we regularly investigate and test the procedures, to ensure that others, not just the original authors of a manuscript, can reasonably access the data. The terminology and legal framework differs across even similar agencies - "special sworn status" at the Census Bureau is broadly similar to "designated agent" at the Bureau of Labor Statistics, but not identical. This creates a steep learning curve for anybody wishing to navigate the system. Even simple contractual transactions for data access involve many individuals, many rounds of correspondence. Scaled up to 100s of researchers, these varied and incompatible processes cost researchers and the U.S. government a lot of money, and create friction in the efficient use of data, even when guaranteeing the security of the data. While work on a single access portal for the National Secure Data Service is underway, simplification of terminology for other data files that are available outside of that framework should also be considered.
5. **Streamlined researcher certification**: In certain other countries, individuals can be pre-vetted for access. "Accredited researchers" are uniformly vetted in the UK, and listed on public pages, regardless of access environment or project. In Canada, expedited access for experienced researchers - those with certain affiliations and prior experience on secure data access - is being considered. Pilot projects in the United States on university-managed "researcher passports" have not found the traction they deserve, maybe because this needs to be a task centralized with the federal government.
6. **Streamlined approval for reproducibility-related access**. At the AEA, when assessing the computational reproducibility of research conditionally accepted for publication, we often run into the problem of timely access to confidential data. In general, we have two options: we can request access to the data ourselves, or we can ask others to conduct such reproducibility assessments. In order to request access to the data ourselves, we almost always have to initiate access

requests that are completely disconnected from the authors' original request. Yet we attempt to do no more - by design - than the original authors aimed to do, and we do not intend to publish any new results, only verify existing results. It would be extremely helpful if a **streamlined access for a reproducibility team** were feasible - and it would save the federal government time and money. Every reproducibility attempt accesses data under the exact same justification that was previously authorized, and generates no new publications. A simple reference to the previously approved access request should suffice, in combination with any personal assurances that are legally required. At the AEA, we are trialling such streamlined access procedures for German secure data access and with certain commercial providers of proprietary data. We have had no success with the federal statistical system.

7. The second path to obtain assurances that researcher results are computationally reproducible is to request support from the federal agencies that control access to the data, i.e., ask them to run code, or to otherwise verify that the code provided by authors has successfully been used to generate the result. The former option - a staff member runs the code - requires that staff dedicate some time to such a task - akin to providing a referee report when a journal editor asks for input. However, in almost all cases where we have asked federal agency's leadership for such support, we have heard that there is funding to support use of staff time, since it "is not in the mandate of the agency." **Providing both funding, and a mandate, for reproducibility checks by agency staff should be encouraged**. I note that such activities don't just generate costs for agencies - they also provide benefits to agencies. Agencies can become aware of the latest type of analyses being conducted with their data, may learn about new econometric or programming techniques, and can serve as a skill enhancing activity.
8. **Publication of permanent digital identifiers**, using industry standards, that move with the data throughout the archival lifecycle. For instance, data should not change identifiers when moving from an agency to the National Archives. Examples include DOI and Handles, but the federal government is large enough that the application of a US government identifier system would be sufficient if universally applied. Such identifiers should be assigned at the earliest possible opportunity, not just upon publication of the data. Identifiers should be assigned to all relevant objects, including especially those accessible only through application procedures of varying degrees. This is already being done by German and French systems of secure data access, see for instance the confidential "Linked Personnel Panel" in Germany (assigned the DOI 10.5164/IAB.LPP1617.de.en.v1), or the "Panel tous salariés", the French equivalent of the Census Bureau's Longitudinal Employer Household Panel

(assigned the DOI 10.34724/CASD.85.1177.V1 for the 2017 cross-section, and similar DOIs for all prior years).

9. Scientific integrity is supported by verifiability, and traceability. For the publications of the AEA, we assess the provenance of data back to their source, and then request that authors provide computer code and instructions documenting all subsequent modifications. This should be standard practice for all government publications, whenever they use or convey data - backed up by publicly available code and traceable source data. While any such publications go through an extensive review process, that review process does not leave public artifacts. **Replication packages for government publications** are a good way to start, and are not hard to implement when following best-practices in coding and development. The AEA, for authors publishing in our journals as well as in collaboration with other journal editors, has compiled guidance that would also be applicable to many other publication types.
10. **Public-use code and specification documents** should be envisioned for all new and revised data products. By following industry-standard secure coding standards, code and specification documents can be written from the outset in a way that does not reveal any unavoidable secret parameters. While this is hard to implement for existing code - millions of lines of code would need to be reviewed - it is quite a bit easier to never let secret parameters enter code as it is created and continuously reviewed. (These are practices we followed when implementing the original QWI codebase, and when defining policies for maintaining the codebase)

I appreciate this opportunity to provide you with these recommendations. I am available and would be happy to brief OSTP, the White House Scientific Integrity Task Force, and OMB on the recommendations, their origins, effectiveness, and their importance to the economics community.

Sincerely,

Lars Vilhuber
Data Editor
American Economic Association