Exploring Improved Access to Individual Income Tax Data for Evidence-Building

Amy O'Hara, Barry Johnson, Paul Arnsberger,
Stephanie Straus, Ron Borzekowski

Submitted to the National Bureau of Economic Research (NBER) Data Privacy Protection and the Conduct of Applied Research Conference

05/16/2024

Table of Contents

Submitted Abstract

In recent years, important and headline-grabbing findings have emerged from research using individual income tax data for statistical purposes. Demand for these microdata, accessible under the tax administration authority of the Internal Revenue Code and through the IRS Statistics of Income (SOI) Division's Joint Statistical Research Program, continues to grow. However, such approaches constrain access to these data and impose substantial resource costs and risks on the public organizations providing the data.

This paper describes an alternative approach, using a trusted service provider and trusted PETs, to allow access to, and use of, tax information for critical applied research while recognizing the need to protect privacy. The project explores the feasibility of a privacy preserving secure query system (SQS) linking end-users of the data, a data intermediary, and a data provider. In the early stages of development, the end-users will be state or local governments, or nonprofit institutions; the intermediary will be at an academic institution; and all processing will be done within and by SOI staff. For example, a local government will use tools developed in this project to produce a dataset with personal identifiers and any tabulation groups necessary for their analysis. They will securely transfer those files to SOI, which the intermediary will help facilitate. Using approaches developed in this project, SOI will match these data to individual income tax information and produce privacy protected standardized output tables. The automated system should run efficiently enough to allow greater production of evidence at much lower cost to the data providing entities.

There are several key components of the system being developed in this project. First, the system imposes a strict schema and various quality controls on input data, along with the technology to enforce these. While the data can come from varied sources (and therefore address a broad array of questions), the automated system requires strict controls on the input file (e.g., format, completeness). Second, a common data model supports accurate and ultimately privacy protected record linkages. Third, the system produces limited output data using a combination of privacy preserving methods through automated disclosure review protocols. Enabling these steps, the intermediary manages the legal, security, and administrative aspects of the system through agreements with SOI and end-users.

The system's privacy protecting technologies and automated disclosure reviews ensure compliance with IRS statutory and regulatory requirements. SOI can use tax data under existing tax administration

authority under Internal Revenue Code (IRC) 6103(h) and special statistical studies (IRC 6108(b)). This project considers outputs from a secure query system as compilations of tax information under IRC 6108(b). The project focuses on individual tax data derived from information returns (Forms W-2, 1099) as well as individual income tax returns (1040). Future efforts could possibly utilize business tax returns, for example to measure outcomes for the Inflation Reduction Act. However, in all cases, the outputs of the system must meet the agency's disclosure guidelines and, notably, cannot reveal fact-of-filing or any specific federal taxpayer information. This may limit the outputs in terms of geography or other information that could be disclosive. Current laws penalize IRS employees if Federal Tax Information is published and this results in rigid and conservative approaches to output privacy protection. The project will explore potential changes to laws that would shift and share penalties for improperly using tax information to identify taxpayers.

Successful development and deployment of the query system should support both SOI and the user community. Most notably, an automated system would offer substantial reductions in the administrative burden at SOI. Currently, projects at SOI require tailored one-off data use agreements for each project; customized delivery of input files that then require data cleaning and preparation by SOI employees; and manual disclosure review. Such a process is burdensome and cannot scale, leaving only a limited number of researchers with access to this valuable resource.

The automated system would also limit the risks of disclosure while furthering the mission of the SOI. Fewer individuals accessing the data and standardized outputs based on automated review should also limit opportunities for any data leakage. Further, should the system develop toward a process where linkage is done with hashed identifiers, privacy protections will be even further enhanced. Considering this new tier of access should help SOI further comply with the Foundations for Evidence-Based Policymaking Act, which pushes toward a presumption of accessibility. This ambitious use of privacy enhancing technologies (PETs) could inform the National Secure Data Service and what services it may offer.

While this approach to automation and privacy protection limits the types and depth of available analyses, end-user outreach to date indicates strong support for the system as described in this paper. A number of states and non-profit institutions indicate that they are willing to agree to the terms, can provide data necessary to make this approach viable and would benefit from the, admittedly, limited standardized output from the system.

SOI has been exploring and implementing PETs, including synthetic data for non-filers and public-use files, and has implemented a form of differential privacy to protect data provided to the Department of Education in support of their College Scorecard webtool. The SOI secure query system adds a new tier of access, enabling statistical analysis of valuable, restricted federal administrative data to support evidence-building needs.

Introduction

The Internal Revenue Service (IRS) Statistics of Income (SOI) Division is the Federal Statistical Agency within the Treasury Department responsible for transforming administrative tax records into statistical products. SOI has long been a leader within the Federal Statistical System in responsibly releasing administrative data for research use. In 1960, when most government administrative data were available only to internal agency staff, SOI began producing a public use file (PUF) of tax returns for research. In general, tax records cannot be released and remain available only to internal agency staff. Technological advances in data capture, retention, and processing allow IRS to manage ever-increasing amounts of administrative data that have great potential for use in policymaking. However, access is curtailed due to legal barriers, logistical challenges, resource constraints, and privacy concerns.

This paper describes a new access mode that enables government agencies or nonprofit institutions to obtain statistics from federal tax information (FTI). With support from an intermediary, initially at an academic institution, this model relies on SOI staff for all FTI processing. By working with a trusted service provider and privacy enhancing technologies (PETs), this model will allow access to, and use of, tax information for critical applied research while recognizing the need to protect privacy. Below we describe how this privacy preserving secure query system (SQS) aligns with SOI's past and current deployments of tiered access models, and with the federal government's evidence building agenda.

Current Models for Providing Access

Individual income tax data cover the vast majority of Americans. Tax returns contain limited demographic information (i.e., marital status and number of claimed dependents), and geographic information from mailing addresses (Slemrod, 2016; Vartivarian et al., 2007; Larrimore et al., 2017). Tax data are collected with a great deal of specificity across different income types (available in instructions,

worksheets, and forms). All federal tax filers use these forms and worksheets, making this uniformly documented data source valuable for comparisons across geographies or over time.

SOI manages tiers of access to tax data, from the most restrictive, closed data to fully open data – and they have been doing so for over sixty years. This paper first describes SOI's current models for providing data access, weighing their benefits and challenges and assessing their reach across external researchers and nonprofits. The paper then describes the SQS, and how it provides approved requestors with greater access to administrative tax statistics, while retaining privacy protections.

Open Data SOI Products

SOI generates regular statistical series, data books, and special studies using FTI. SOI has changed these products and the data they make available over time, with some expansions in response to user requests and retractions due to privacy concerns. Key examples of SOI open data products are the annual ZIP Code level tax return data, US population migration data, and county income data products. These datasets are aggregated, containing no individual tax return information, and are therefore publishable outside the agency.

Use Case: Domestic Migration Research

Tax returns have large coverage of the population, are submitted annually, and include filers' mailing addresses, making them an ideal source of data for migration research. Compared to other common sources of migration data, such as the Decennial Census and the American Community Survey, tax return data has several advantages, including a more detailed time frame (annual vs. decennial for the Census or 5-year averages for ACS), and larger coverage of the population (Hauer & Byars, 2019). Hauer and Byars developed a single, flat, standardized file containing national county-to-county migration data from 1990 to 2010, estimating coverage at 95-98% of the tax-filing universe and their dependents, or approximately 87% of US households, allowing for research using detailed, annual data on domestic migration flows with wide coverage of the national population. They demonstrated the file's effectiveness in showing, for example, strong out-migration from New Orleans following Hurricane Katrina in 2005 (Hauer & Byars, 2019).

Joint Statistical Research Program (JSRP)

SOI established a Joint Statistical Research Program (JSRP) in 2012, permitting selected researchers access to federal tax microdata approved in a competitive proposal review process. The JSRP increases use of SOI's data by allowing limited researcher access specifically to support tax administration, advancing knowledge of how taxes affect people, businesses, and the economy (Slemrod, 2016). Given the sensitivity of tax return data and the potential damage from a breach, however, the program's data access controls are stringent. The number of research teams granted access is modest, with 25 projects selected in 2023. Researchers on approved projects must undergo a thorough background clearance and are bound by the same rules and penalties as IRS employees. Each approved project is assigned to SOI subject matter experts who monitor all aspects of the project to ensure compliance with data access and security standards as well as project scope. This adds burden to SOI employees who must provide monitoring, technical assistance, and disclosure avoidance review. These activities must be absorbed into the SOI employee workloads.

Table 1. Number of JSRP Projects Approved, by Application Year

Year	2012	2014	2016	2018	2023
Projects	17	12	18	29	25

Table 1 shows the number of projects approved by the JSRP increasing over the biannual program's history. The program is constrained by IRS staff availability and data access through an IRS approved site. According to IRS staff, for the 2023 round, 114 applications were received and reviewed by IRS staff to determine which would advance knowledge on tax administration.

Described in ACDEB report as "a small researcher access program which has yielded groundbreaking studies, as a partial solution toward expanding evidence-building capacity," the JSRP program proves useful for those whose projects are selected (ACDEB, 2022). But the program is resource intensive for SOI staff who must monitor the selected researchers who take on Intergovernmental Personnel Act (IPA) assignments during their four-year projects.

SOI Public Use File (PUF)

Early SOI Public Use Files (PUF) contained anonymized and altered individual income tax return records to protect against the risk of disclosure. A subsample of the SOI Individual Income Tax Annual Cross

Section file that the Office of Tax Analysis and Joint Committee on Taxation use in their microsimulation models, the PUF has been a critical resource for economic and tax policy researchers, enabling studies of the effects of tax policy changes, the distribution of tax burdens, and economic incentives (Bowen et al., 2020). Researchers using PUF files agree to terms of use (e.g. pay fee, do not attempt to re-identify any taxpayers, do not share their copy of the PUF). However, the growing availability of external data sources that could be linked to the PUF, as well as increasing technological capabilities to re-identify individuals in the sample, forced the IRS to distort the data in increasingly aggressive ways, making the data less useful for analysis (Bowen et al., 2020).

SOI Synthetic Public Use File (PUF)

In response to growing threats over use of actual tax records in the PUF, SOI, in partnership with the Urban Institute and Tax Policy Center, developed a synthetic PUF, using statistical methods to generate fake data that mirror the underlying IRS data (Burman et al., 2024). There are a number of concerns regarding the creation and release of a synthetic PUF. The structure of tax returns complicates the production of synthetic data, since virtually all variables are involved in accounting relationships and nonlinear tax computations across the forms and schedules. Synthesizing a specific variable, such as total income, necessarily involves synthesizing its components, such as wages and salaries, but the accounting relationships between all variables must be preserved for the synthetic PUF to be useful (Vartivarian et al., 2007). Despite these difficulties, a team at the Urban Institute and IRS persevered. They have produced fully synthetic supplemental PUFs for tax years 2012-2015, based on information returns such as forms W-2 for reporting wages and Forms 1099 that report pensions, interest, dividends, etc., for individuals whose income is likely below the tax filing threshold. They have also produced a fully synthetic beta Tax Year 2015 Form 1040 file that is being tested by current PUF purchasers.

College Scorecard

SOI supports projects that need custom tabulations of FTI for tax administration purposes and for special studies. The College Scorecard produces aggregate earnings data on a range of years after graduation, for students who received federal financial aid, by U.S. postsecondary school, credential level, and program of study. This web tool is "designed to help students make informed decisions about their education options after high school, bringing together information on college costs, graduation rates, student loan debt, post-college earnings, loan repayment rates, and more" (Kaouk et al., 2021). To release this fine-grained information, SOI developed a disclosure avoidance protocol using differential privacy that

reduces re-identification risks for students and programs. The protocol relies on SOI and ED assumptions about the acceptable tradeoff between privacy and accuracy.

As described above, SOI's efforts on open data resources, the JSRP, synthetic PUF, and College Scorecard meet many users needs for tax and income information. However, these products fail to meet the measurement needs of program administrators, evaluators, and policymakers who need information about specific groups, not the entire population.

Secure Query System (SQS)

Georgetown is assisting SOI with design options for the SQS, relying on outreach conducted with Yale University to federal agencies, state agencies, local governments, and nonprofit organizations to assess their needs for data regarding tax filing behavior and income metrics. These organizations cited many broad measurement goals including benchmarking, evidence-based budgeting, informing workforce development programs, increasing economic mobility, and understanding labor flows across state borders. State officials were eager to explain these varied needs, and to specifically explain gaps in their capacity for measurement. For example, state education and workforce officials need to know whether their learners are earners. Most are using their state Unemployment Insurance wage data but lack visibility for out-of-state earners and non-employer earnings. They are looking for aggregate statistics that show the margin of missing earnings, and they are looking for indicators about the extent of out-of-state earnings. State health and human services officials lack visibility into tax filing and credit claiming behavior; they cannot measure whether outreach efforts are inducing more state residents to file and claim refundable credits. State justice officials lack information on training programs prior to prisoner re-entry. State economic development officials lack insights about retention of in-state college graduates and career pathways. These officials also seek information on industries in which learners or beneficiaries work and the impact of program changes and interventions. Some are focused entirely on the individual, and others are interested in tax units.

These discussions informed Georgetown's SQS process design, especially the output statistics of interest and use to potential clients. Table 2 shows the initial planned set of SQS-1040 output statistics to be produced using individual income tax data linked to the client's input file.

Table 2. SQS-1040 Planned Output Statistics
1. Percent filed 1040 (1040 filers/total records)
2. Filing status frequency (for 1040 filers)
3. Percent claimed EITC (for 1040 filers)
4. Average EITC amount (for EITC>0)
5. Percent claimed CTC (for 1040 filers)
6.. Average CTC amount (for CTC>0)
7. Percent with Schedule C
8. Percent with only Wage income
9. Mean Total Wage Income on 1040
10. 25/50/75 Percentiles on Total Wage Income
11. Mean AGI (with standard deviation)
12. 25/50/75 Percentiles on AGI
13. Median AGI by Filestat

SQS Process Overview

To establish the SQS, SOI and Georgetown University, the temporary intermediary, are developing options for client data submissions, data linkages inside IRS, statistical analyses and tabulations, and disclosure avoidance methods. In parallel, Georgetown is working with Yale and potential clients to confirm that the pre-defined SQS outputs will meet their measurement needs. Georgetown is developing protocols to help potential clients validate their input data to verify the presence of sufficient identifying information for linkage and adequate cell sizes for output statistics. When potential clients pass these validation checks, they enter a standard agreement with Georgetown as the intermediary and prepare their data extract for submission. Georgetown facilitates transfer of encrypted client input files to SOI without seeing the file contents. SOI receives the client data and runs an automated matching, data transformation, and tabulation process. SOI populates the pre-defined output tables and applies privacy protections. The resulting aggregate statistics (no longer Federal Tax Information) are transferred back to the clients through Georgetown. Once Georgetown confirms receipt of the output tables, SOI destroys

the input files. Georgetown distributes the output tables back to the clients, and documents SQS inquiries, submissions, and completions for SOI.

Having Georgetown act as intermediary in this process removes SOI from the burden of negotiating one-off legal agreements, handling data transfers and receipts that do not conform to agreed standards, recoding input files to align with SOI schema, and designing disclosure review protocols on variable output statistics. SQS is an opportunity to standardize the external data access approach, from the first request of statistics by the outside agency or organization, down to the disclosure avoidance protocols before the statistics are released. By rethinking the access process and building it to scale, SOI will achieve greater equity, transparency, efficiency, and risk reduction.

How the SQS Differs from Existing Models

Within the SQS, constraining both inputs and outputs allows for automation, scaling, and efficient use of tax data for these analytic and policy goals. This approach reduces burden on SOI staff and decreases risks of inappropriate disclosures. Table 3 shows how SQS differs from current practice. Instead of SOI staff negotiating separate data sharing agreements with organizations wishing to have their data matched to produce tax and income statistics, SQS has a single template for clients to request SQS outputs.

Table 3. SQS Compared to Current SOI Practices for Special Studies

Current Practice	**SQS Changes**
Custom Data Sharing Agreement	Single agreement template
Open-ended inquiries and custom outputs	Pre-defined output statistics
Analysis conducted by IPAs managed by SOI staff	Periodic production by SOI staff
Reliance on Tax Administration authority	Reliance on Special Studies authority
One-off disclosure avoidance review of outputs	Standardized and automated disclosure review

SQS is designed to run with minimal IRS staff effort, with the bulk of data preparation and alignment work being carried out by the potential clients in state and local governments and nonprofits, and the design and administrative burden handled by Georgetown as the initial intermediary. Risk is minimized with SQS, as only employees use FTI and limited pre-defined outputs are produced.

Effective Privacy Protection

Georgetown conducted outreach to potential SQS clients—state and local government agencies, academic institutions, non-profit service providers, and evaluators—to learn how they view the trade-offs of various disclosure avoidance methods. We gave examples of the different levels of granularity and types of output statistics they could receive, and how to avoid receiving tables full of suppressed values. The potential SQS clients confirmed their eagerness to obtain privacy-protected output statistics, and their willingness to participate in conversations about privacy trade-offs. They asked about metrics to quantify benefits and risks when comparing disclosure avoidance methods.

SOI must abide by its laws and guidelines, including those in IRS Publication 1075, that, "Statistical reports may only be released in a form that cannot be associated with, or otherwise identify, directly or indirectly, a particular taxpayer" (IRS, 2021). Georgetown explained these constraints and found that the potential SQS clients understood the need for coarsening, suppression, or noise injection to obscure observations in output statistics. While they would like more information on how noise infusion would affect output utility, the potential clients supported a reduction in precision of estimates in order to gain strong privacy protections. As illustrated in Table 4, clients were supportive of receiving output in Column B, which represents the output statistics after they have undergone disclosure avoidance. They confirmed their willingness to receive (and the usefulness of) the aggregate tax and earnings information with values rounded to the nearest hundreds place, with the understanding that further coarsening may be needed (to 500s or 1000s, depending on the statistic). They also understood that percentages may be released as ranges, that quartiles may not include the bookends of value ranges, and that only certain mean income statistics would come with standard deviations. Finally, they understood that cell size limitations may trigger SOI suppression rules in cases where counts do not allow publication.

Table 4. SQS-1040 Planned Output Statistics, Before and After Disclosure Avoidance		
	Column A - before DA	Column B - after DA
1. % filed 1040 (in1040/total records)	0.7252643467	0.73
2. Filing status frequency (for in1040)	0.1962346623 Single 0.4905371305 Head of Household (HH) 0.3158271053 Married Filing Joint (MFJ) 0.0082493532 Other	0.20 Single 0.49 HH 0.32 MFJ 0.01 Other
3. % claimed EITC (for in1040)	0.3523504634	0.35
4. Average EITC amount (for EITC>0)	1097.8885357907	1100
5. % claimed CTC (for in1040)	0.4236475832	0.42
6. Average CTC amount (for CTC>0)	592.715799067543	600
7. % with Schedule C	0.0382405322	0.04
8. % with only Wage income	0.6480678329	0.65
9. Mean Total Wage Income on 1040	38246.2942712593 (467.3024839053 s.d.)	38000 (500 s.d.)
10. Quartiles on Total Wage Income	25th 18651.3029485322 50th 36005.4204837212 75th 43488.2039837294	25th 18700 50th 36000 75th 43500
11. Mean AGI (SD)	40098.2938273729 (901.1382784373 s.d.)	40100 (900 s.d.)
12. Quartiles on AGI	25th 19003.3627989032 50th 36989.0984039843 75th 44008.3534218398	25th 19000 50th 37000 75th 44000
13. Median AGI by Filestat	18994.1736200983 Single 34703.0378432671 HH 42499.9864210945 MFJ 24278.2938473274 Other	19000 Single 34700 HH 42500 MFJ 24300 Other

Table 4 illustrates the implications of using privacy protected data for estimation and inference when conducting applied research. Potential clients provided input on statistics of interest; Georgetown

confirmed feasibility with SOI, and shared possible rounding and suppression approaches with potential clients. When the need arises, differentially private statistics will be developed and socialized with potential clients.

In developing the disclosure avoidance protocols for SQS, Georgetown learned that the most vital aspect was trust of the end users. The potential SQS clients confirmed they are willing to provide their highly sensitive data for linkage to SOI data and accept imprecise aggregate statistics because a value proposition exists. Further, SQS success is measured not only by the number of clients who successfully input their data and receive useful statistics back, but also by minimizing the risk and burden SOI faces as they operate the SQS.

Legal Authority and Policy Environment

IRS Authority

SQS can be conducted under I.R.C. Section 6108(b), stating that: "The Secretary may, upon written request by any party or parties, make special statistical studies and compilations involving return information (as defined in section 6103(b)(2)) and furnish to such party or parties transcripts of any such special statistical study or compilation. A reasonable fee may be prescribed for the cost of the work or services performed for such party or parties." Special Statistical Studies must be conducted by SOI employees; they cannot rely on IPAs or contractors. Other projects, including the College Scorecard, have been conducted under this authority.

Policy Environment

The Commission for Evidence-Based Policymaking (CEP) noted in its Final Report that complex legal regulations and internal agency policies "limit the effective, efficient, and transparent use of existing data" (Abraham et al., 2017). CEP identified tiered access as a potential solution that balances greater access and robust privacy protections. The Commission recommended that Federal departments consider the sensitivity of the data, with input from stakeholders, including researchers and privacy advocates, to establish access restrictions based on law, context, and sensitivity.

In 2018, the passage of the Foundations for Evidence-Based Policymaking Act (Public Law 115–435) advanced many of the recommendations of the CEP report. For example, the Evidence Act, as PL 115-

435 is known, requires agencies to develop learning agendas, inventory their data assets, name statistical officials, chief data officers, and chief evaluation officers. The Evidence Act also required a committee to research whether and how the U.S. could establish a National Secure Data Service. The Advisory Committee on Data for Evidence Building (ACDEB) gathered evidence and proposed recommendations about improving secure and efficient access to government data. Recommendations relevant to SQS are listed in Table 5.

Table 5. SQS-Relevant ACDEB Recommendations
Rec. 1.6. OMB should adopt a risk-utility framework as the basis for standards on sensitivity levels, access tiers, and risk evaluations as part of the regulation on expanding secure access to CIPSEA data assets.
Rec. 1.7. OMB, in coordination with the ICSP, should promote the use of privacy-preserving technologies in the tiered access framework required under Title III of the Evidence Act by identifying an initial set of promising tools over the next 1 to 3 years.
Rec. 1.8. OMB, in coordination with the ICSP, should identify models for shared responsibility among data providers and users and provide guidance on applying such models through the regulation on expanding secure access to CIPSEA data assets.
Rec. 5.6. The NSDS should facilitate the development and application of statistical disclosure limitation methods. •Invest in open source tools and training •Encourage more researchers to contribute to this work •Use realistic risk models •Facilitate privacy risk assessments, and catalog info from past projects
Rec. 5.8. The NSDS should provide tools and support to users in conducting scalable, privacy-preserving record linkages, facilitating data preparation and review of matching metrics. As part of its data concierge services, the NSDS should coordinate with federal, state, territorial, local, and tribal government officials seeking linkage services.

SQS identifies what is important to clients and mitigates risks of producing those statistics. This aligns well to Recommendation 1.6, as SQS introduces new methods to handle input privacy (e.g., data from

clients) and output privacy (e.g., non-disclosive output tables), that could inform OMB efforts. Similarly, SQS can help inform both OMB (Recommendation 1.7) and NSDS (Recommendation 5.8) on privacy preserving record linkages (PPRL). Research is underway at Georgetown and through NSDS Demonstration Projects to understand whether clients have sufficient data quality and technical readiness to adopt PPRL. SQS can demonstrate the value of coordinating with federal, state, territorial, local, and tribal government officials seeking linkage services, and through the strategic use of an intermediary to offset burden on SOI staff.

Regarding models of shared responsibility called for in Recommendation 1.8, SQS will produce evidence to inform OMB: a shared responsibility model is embedded in SQS client DUAs. Clients must attest that they have authority to share data, demonstrate that their data conform to SQS schema, and agree not to attempt re-identification to learn about fact of filing or FTI details. SOI shares responsibility to prevent release of FTI, acknowledging that not all data are equally sensitive, to guide their application of disclosure avoidance tools to SQS output tables. This relates to Recommendation 5.6, which calls on the NSDS to use realistic risk models. Until the NSDS exists, SOI is asking important questions including: What could a client learn based on the output tables? How could a potential bad actor exploit the system? SOI and Georgetown have planned an expert review of SQS disclosure methods prior to production.

Even as the NSDS and other initiatives work to improve access, the SQS team has identified a real and tangible need for evidence that can be met with a standardized and automated query service. This need will likely exist even if, and when, greater access to underlying data comes online. Each of these will have an important role in the future toolkit for evidence-based policymaking.

Conclusion

There is a perception that expanded access to FTI requires legislative change and radical increases in IRS funding and infrastructure. SQS demonstrates that academic-government partnerships can accelerate research and development, and that agencies have authority to establish new tiers of access without legislative change. SQS shifts burden from SOI in two ways: SQS relies on clients to review and recode their own data, prepping it for submission to SOI, and it relies on Georgetown (as intermediary) to manage the administrative aspects of the project. Sharing the burden and costs of innovation makes it possible. Still, SQS is not costless for SOI. Depending on the frequency and scale of SQS, SOI will need to plan for cost recovery.

To inform cost models, Georgetown will capture useful metrics from pilot queries, including successes and failures of the input data validation system, legal and agreement issues, and data quality challenges that may hamper or hasten client interest in SQS output tables. Outreach with potential clients indicates high interest in additional individual income SQS queries, including more sophisticated analyses that handle lagged matches (computing outcomes on older cohorts), interventions including data from randomized controlled trials, and specific populations (e.g., learners on credential pathways).

These innovations require funding for SOI staff to run the queries; funding is also needed for research and development, outreach, capacity building in state and local agencies, whether these activities take place within SOI or with assistance from an intermediary. The SQS model can expand within SOI to corporate tax data and can inform efforts at other federal agencies.

References

Abraham, K., Hasking, R., Glied, S., Groves, R., Hahn, R., Hoynes, H., Liebman, J., Meyer, B., Ohm, P., Potok, N., Mosier, K., Shea, R., Sweeney, L., Troske, K., & Wallin, K. (2017). The Promise of Evidence-Based Policymaking Report of the Commission on Evidence-Based Policymaking. https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Advisory Committee on Data for Evidence Building (2022). Year 2 Report Supplemental Information. https://www.bea.gov/system/files/2022-10/supplemental-acdeb-year-2-report.pdf

Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., & Williams, A. R. (2020). A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings* (pp. 257-270). Springer International Publishing.

Burman, L., Johnson, B., MacDonald, G., (others TBD). (2024). Protecting Privacy and Expanding Access in a Modern Administrative Tax Data System, National Tax Journal (forthcoming).

Hauer, M., & Byars, J. (2019). IRS county-to-county migration data, 1990–2010. *Demographic Research*, *40*, 1153-1166.

Internal Revenue Service. (2021, November). Publication 1075: Tax Information Security Guidelines For Federal, State and Local Agencies, Safeguards for Protecting Federal Tax Returns and Return Information https://www.irs.gov/pub/irs-pdf/p1075.pdf

Kaouk, T., Fortelny, G., & Allen, R. (2021). Chief Data Officers Presentation for the Advisory Committee on Data for Evidence Building. https://www.bea.gov/system/files/2021-02/OCDO-ACDEBpresentation-FEB2021.pdf

Larrimore, J., Mortenson, J., & Splinter, D. (2021). Household incomes in tax data: using addresses to move from tax-unit to household income distributions. *Journal of Human Resources*, *56*(2), 600-631.

Slemrod, J. (2016). Caveats to the research use of tax-return administrative data. *National Tax Journal*, *69*(4), 1003-1020.

Vartivarian, S., Czajka, J. L., & Weber, M. (2007, July). Measuring Disclosure Risk and an Examination of the Possibilities of Using Synthetic Data in the Individual Income Tax Return Public Use File. In *American Statistical Association Meetings*.