The Secure Query System: Unlocking Linked Employment and Earnings Outcome Measurements

Joshua Snoke[1], Amy O'Hara[1], Vincent Dorie[1], Jacob Pasner[1], Kangheng Liu[1], and Ron Borzekowski[2]

[1] *Georgetown University*

[2] *Yale University*

[*] *Corresponding author: joshua.snoke@georgetown.edu*

Keywords— Employment, Earnings, Administrative Data, Internal Revenue Service, Privacy, Confidentiality

Abstract

The Secure Query System (SQS), a partnership between the Georgetown Massive Data Institute and the Internal Revenue Service Statistics of Income Division provides a novel resource for measuring earnings and labor force participation in areas not captured by existing data sources. Official statistics deliver aggregated statistics or estimates gathered through surveys, while databases built from public data scraping or data on unemployment insurance wages are susceptible to lack of representation. Among products that target outcomes by specific institutions, many only include individuals who attend 4-year academic institutions, missing those who complete other degree or credentialing programs. The SQS was built to deliver statistics to help measure the outcomes for targeted cohorts of individuals who are missing from these traditional sources. By enabling clients to obtain measurements for specific cohorts, SQS unlocks the potential to measure earnings and labor outcomes for any subset of the market. Combining data linkage and data privacy methodology with best practices in data governance, the SQS delivers accurate and safe outputs while minimizing the administrative burden and efficiently allocating the cost of running the system. This proposed paper describes the methodology underlying the SQS and the governance and administrative structure. The SQS went live in early 2026, and we will present three case studies with real clients comparing the SQS outputs with their prior data sources.

1 Existing systems for accessing earnings data

A growing need exists to measure the employment and earning outcomes for specific cohorts that is not currently addressed by official statistics or other available data sources (Berger et al., 2024). These data are necessary in order to evaluate the effectiveness of programs and initiatives fielded by

many non-profit organizations, state and local governments. For example, non-degree credentials are growing much more rapidly than post-secondary degrees (HolonIQ, 2023), but organizations are struggling to identify accurate data sources to evaluate the value of different credentials. Additionally, state and local governments may have an interest in providing statistics on outcomes for individuals in their areas, or they may be interested in assessing the effectiveness of policy initiatives. In either case, they need high quality employment and earnings data which can be linked to specific cohorts.

Existing data sources for measuring employment outcomes and earnings in the U.S. labor market fail to provide the granularity needed to assess programs or initiatives conducted by many regional non-profit organizations or local governments. Official statistics, such as those produced by the Census Bureau[1,2], or the Internal Revenue Service[3] contain measurements which are useful for studying certain groups, but they cannot be tailored to specific areas. Surveys, e.g., the Panel Study of Income Dynamics[4], may be conducted to study novel groups, but they are expensive and inefficient for smaller organizations to field.

Numerous data resources exist for studying post-secondary outcomes, such as Post Secondary Education Outcomes[5] of the U.S. Census Longitudinal Employer-Household Dynamics program, the College Scorecard[6], or the National Survey of College Graduates[7], but these are limited to traditional academic institutions and generally do not include non-degree credentialing programs. Expanding these data to include individuals in other programs could be an option, but given the rapid rise of credential programs and ongoing non-response issues in surveys, this would likely be quite burdensome with questionable representativeness.

Two additional options exist for gathering information on employment and earnings. Researchers wishing to target individuals in specific geographic areas may use state-level unemployment insurance (UI) wages data. This is most commonly used by state or local government agencies who have data sharing agreements and can link individuals to the state UI systems. The main drawback to this approach is that these data exclude informal or contracting work which can bias estimates, particularly when studying new or emerging elements of the labor market.

Finally, some organizations use data which is harvested through other "big data" methods. For example, data can be accessed through the credit bureau Equifax (Equifax, n.d.), which receives W2 information from numerous payroll providers. This data includes employment and earnings

[1] https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html
[2] https://lehd.ces.census.gov/
[3] https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi
[4] https://psidonline.isr.umich.edu/default.aspx
[5] https://lehd.ces.census.gov/data/pseo_experimental.html
[6] https://collegescorecard.ed.gov/
[7] https://www.census.gov/programs-surveys/nscg.html

based on reported W2s, but it is only representative of individuals who work in certain sections of the labor market. Other sources include compiling or scraping self-reported employment and inferred wage data, such as the Burning Glass Institute Credential Value Index database[8]. While potentially large, these data have numerous measurement issues including being a convenience sample and using indirect wage estimation methods. Both these data sources rely on limited PII, so linking specific individuals relies on fuzzy matching techniques which may have limited success and generate biased results.

2 The SQS: a solution for linked earning statistics

The Secure Query Service (SQS) is a collaborative project between the Georgetown Massive Data Institute (MDI) and the Internal Revenue Service Statistics of Income (SOI) Division that creates a new tier of access for individual income tax data. The overarching goal of the SQS is to provide clients with aggregate statistics by securely linking client data with income tax data and applying principled disclosure avoidance to ensure confidentiality is systematically preserved in the output statistics. Linked statistics enable targeted analyses where current data sources fall short. For example, linked cohorts can be used to evaluate training programs or randomized controlled trials for new policy initiatives.

2.1 Background on the SQS

The SQS started with two members of the Advisory Committee on Evidence-Based Policymaking, created under the Foundations of Evidence Based Policymaking Act of 2018 (United States Congress, 2019), who recognized the need for aggregate statistics measuring labor force participation and earnings and the need for efficient, secure use of administrative data. The SQS was designed by Georgetown's Massive Data Institute in collaboration with the statistical unit within the IRS to address unmet measurement needs in state and local government agencies and nonprofit organizations. The goal of the SQS was to provide predefined aggregate statistics rather than open-ended data exploration or microdata access. This approach limits the burden on the IRS and simplifies the threat surface to which privacy protections must be applied. By producing metrics for previously unreached users, IRS complies with the Foundations of Evidence Based Policymaking Act of 2018.

The original SQS design planned to provide a uniform set of output statistics to all clients, as described in O'Hara et al. (2024). Substantial market research was conducted by Georgetown and partners at Yale University's Data Intensive Social Science Center and Tobin Center in order to develop the suite of desired statistics. However, ongoing environmental scans and submission development with pilot project clients revealed that allowing SQS clients to select only the statistics of interest (rather than receiving a standard output) could provide higher quality statistics that meet clients' project goals without expending the additional privacy risks associated with releasing

[8] https://www.credentialvalueindex.org/

more statistics.

In spring 2026, the first clients went through the process and obtained aggregate statistical information about their program participants. This was a major breakthrough for the IRS. In the past, projects at SOI required tailored one-off data use agreements for each project; customized delivery of input files that then required data cleaning and preparation by SOI employees; and manual disclosure review. Such a process was burdensome and could not scale, and it would not be viable as a long-term solution. The SQS was created to provide efficient delivery of employment and earnings statistics for linked cohorts.

2.2 Overview of the SQS

MDI developed methodology and code to conduct the matching, tabulation, and disclosure avoidance protocols, and MDI serves as an intermediary between IRS and the client by managing client data governance and data sharing agreements. IRS maintains complete control over the execution and tax data, ensuring compliance with all regulations. Utilizing the SQS follows these steps:

1. Agreement Execution: The Client and Georgetown enter into a formal Memorandum of Agreement (MOA) that establishes MDI as the facilitator for the SQS. This process requires the Client to affirm they possess the legal authority to conduct research on the data and transmit it for SQS purposes. The agreement is finalized through authorized signatures from the Client, the MDI Director, and Georgetown's Joint Office of Research Administration (JORA).
2. Preparation & Verification: The Client uses Georgetown-provided documentation and software tools to ensure their data meets SQS technical requirements.
3. Secure Transmission: The Client uploads end-to-end encrypted data to the Kiteworks platform, a secure storage environment controlled by Georgetown.
4. IRS Processing: Pursuant to a Memorandum of Understanding between Georgetown and the IRS detailing data flows, expected processing, and affirming security requirements, IRS retrieves the encrypted data from Kiteworks and processes it within internal, secure IRS computing environments.
5. Linkage, Tabulation, & Disclosure Control: The IRS runs jointly developed code to match Client identifiers with tax data and performs tabulation while applying strict disclosure avoidance protocols.
6. Results Delivery: The IRS generates summary statistical reports containing no Federal Tax Information (FTI) or Personally Identifiable Information (PII). These are sent to Georgetown, which then transmits them to the Client.

Figure 1 provides a schematic diagram of the SQS process. First clients go through a vetting process and enter into an agreement with Georgetown. Only non-profit organizations are eligible for SQS. This provides an additional layer of security ensuring that the clients agree to follow

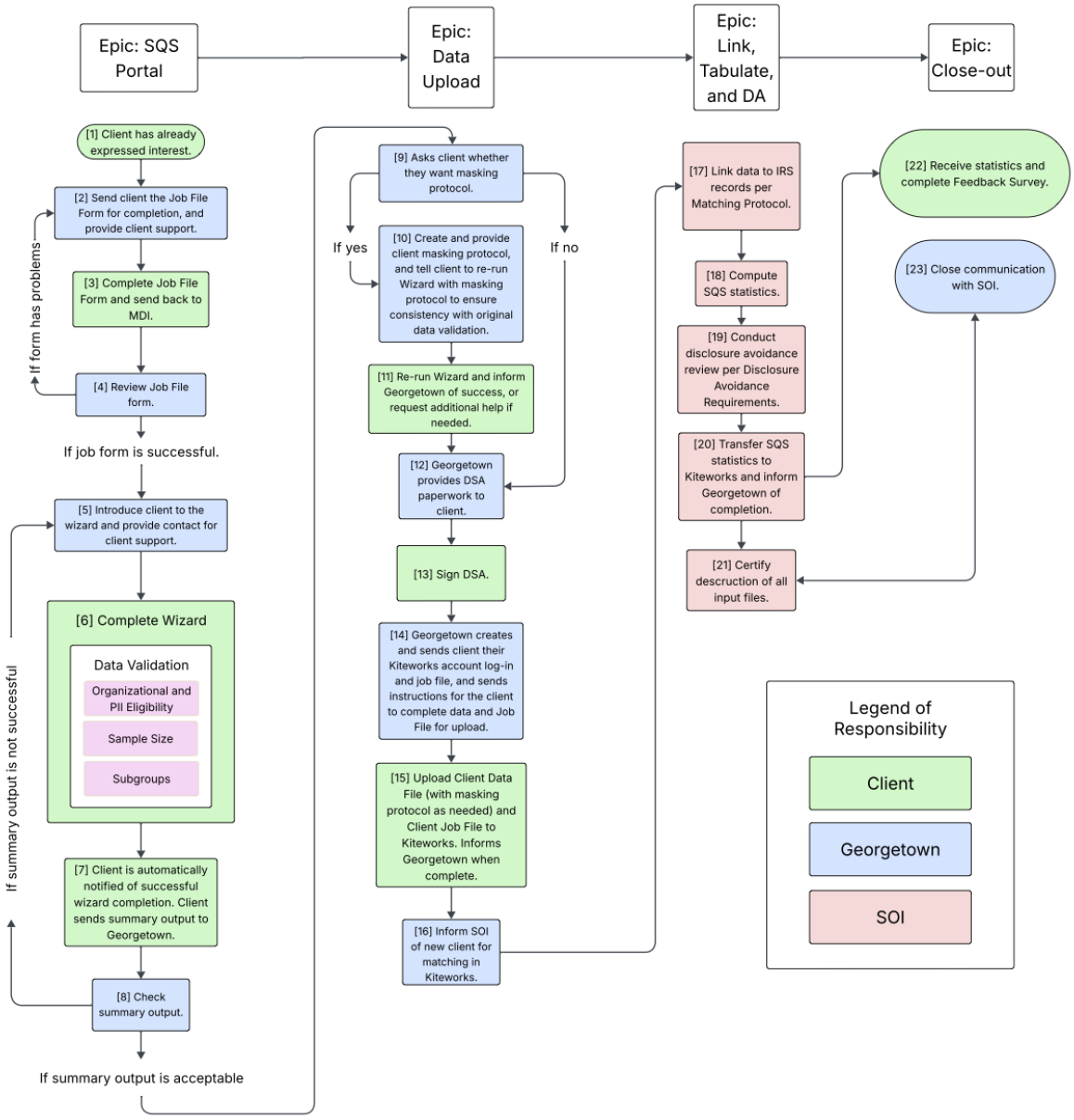


Figure 1: SQS Process from Beginning to End

the submission procedures and responsibly use the output. After entering into an agreement, the clients prepare their data, with technical assistance from MDI. This offloads data preparation efforts onto clients, reducing the administrative burden of running the system and ensures that data preprocessing is done by clients, who have subject matter expertise and familiarity with the input data.

Georgetown developed a data validation interface for clients to prepare and verify their data, ensuring the input file has sufficient data to compute statistics while enforcing the schema naming conventions to expedite automatic runs at IRS. The metadata produced through the data validation interface about input fields and their desired output statistics are important innovations, as are the onboarding steps that educate clients about data readiness, security, and SQS output interpretability. SQS deploys a more automated robust and repeatable service than had previously existed, which saves SOI processing time, makes disclosure avoidance checks easier, ultimately increasing access to SOI aggregated statistics for clients.

The SQS supplies statistics from individual tax data derived from information returns (Forms W-2, 1099) as well as individual income tax returns (1040). Table 1 in the appendix provides a complete list of the statistics currently available through the SQS. Clients select the statistics of interest for their study as well as the desired tax year(s). The statistics which can be requested were selected based on extensive client outreach, the policy decision regarding the trade-off between accuracy and privacy was informed by additional outreach to understand users' needs, and the system is automated to substantially reduce the administrative burden at SOI.

For a viable match to tax records, eligible clients must submit certain combinations of personal identifiers that may include: SSN, SSN/name, name/zip code/DOB, or name/DOB . Individual income tax forms (Form 1040) collect SSN, first name (FN) and middle initial (MI) and last name (LN) in one field. Forms also collect addresses (house number and street name are separate from apartment number), city, state, and zip code. Wage and Tax Statements (Form W-2) include the employee's SSN, FN, MI, and LN in one field, and address in one field (not separating house number/street address, city, state, and zip code). Non-employee compensation forms (Form 1099-NEC) include the recipient's Taxpayer Identification Number (TIN), name (in one field), street address including apartment number, and a single field for city, state, and zip code.

We conducted outreach to potential SQS clients—state and local government agencies, academic institutions, non-profit service providers, and evaluators—to learn about the readiness of their input files. Specifically, we sought to understand the quality and existence of their personally identifiable information fields which match the IRS collections. Across these cohorts, potential clients offer a range of personal identifiers. These include SSN, FN, MI, middle name (MN), LN, and address (typically residential address at time of service/participation/enrollment), which guided our requirements.

In addition to PII for linkage purposes, clients may submit categorical covariate information which is used to provide disaggregated statistics by groups. For example, individuals interested in comparing wages for men and women in their cohort may submit a grouping variable with these levels and receive separate statistics across groups, enabling multivariate analyses. The amount of disaggregation allowed is subject to required minimum cell sizes.

Once data are linked, statistics are tabulated by grouping variables and disclosure avoidance is applied to ensure outputs can be safely released. Privacy is ensured in the outputs through a combination of suppression and noise addition using the framework of Gaussian Differential Privacy (Dong et al., 2022). Publication 1075 of the IRS (Internal Revenue Service, 2021) regulates the disclosure rules, which specifically requires that outputs minimize the risk of revealing fact-of-filing, specific federal taxpayer return information, and restricts the publication of any statistic based on fewer than 20 records.

Our suppression rules enforce minimum cell sizes of 20 for all statistics, including for each level of submitted grouping variables. When a grouping variable does not meet this threshold, all levels are aggregated. In addition to this, random gaussian noise is added to all requested counts, means, and variances. The magnitude of the noise scales proportionally to the number of requested statistics, which reflects the increased risk from providing more information on a set of individuals. The amount of noise added was calibrated to ensure adequate privacy protections while incurring only small additional errors in the statistics. This approach also encourages clients to only request statistics which they require for their study, incentivizing data minimization. The SOI disclosure review board conducts final review of all outputs, and if approved, results are returned to MDI and the client.

3 Case Studies

The SQS completed the first official client run on April 2nd, 2026. Pending final approvals from our partners, full results from three clients will be included as case studies in the final paper. Below we provide outlines of the case studies only in general terms.

3.1 Case Study 1: Comparison with State UI data

The first study concerns a local government client interested in evaluating a transit subsidy experiment designed to increase mobility and earnings for disadvantaged individuals. Earnings are compared after a trial period to determine whether the subsidy program translated to increased outcomes for individuals in the treatment group, compared with individuals who did not receive assistance.

Prior to the SQS, this study could evaluate earnings only using state-level unemployment insurance (UI) wage data. This data has some advantages, such as being available at the quarterly level and covering all UI-eligible employment in the state. It also has limitations, specifically that it does not cover informal or independent contracting work, which may bias earnings estimates

downwards. Accessing UI data also generally requires researchers forming a partnership with a state agency, which may not always be possible or may be time consuming. Lastly, UI systems typically require researchers to link individuals based on social security numbers, which limits the applicability for use cases that submit other forms of PII for linkage.

By utilizing the SQS, this client obtained earnings statistics which covered all reported 1040 wages, W2 wages, and 1099 NEC compensation for the period following the transit study. These statistics are compared to the UI wage estimates to evaluate potential bias in those estimates and to better understand the impact of the transit assistance on earnings. Results will include a direct comparison between UI and SQS estimates and an assessment of the impact on the study results depending on the data source.

3.2 Case Study 2: Comparison with Payroll Data from Equifax

The second case study compares earnings for individuals who receive different types of non-degree credentials. These programs are growing rapidly in comparison with traditional degrees, and they are providing a basis for a growing portion of the labor market. Yet, it remains unclear which credentials provide real value in terms of employment and earning outcomes. The client, a national non-profit, is interested in evaluating the employment and earnings outcomes in order to assess the value of different credential programs.

Because the study encompasses a large number of programs across many geographic areas, payroll data compilers currently provide the most effective data source. The credit bureau (Equifax, nd) is one source which provides data based on reported W2 wages. This data has known limitations, specifically that it only covers certain types of employment, and it exhibits poor match quality when linking individuals to earnings. This results in biased estimates from missing earnings and smaller sample sizes from low match rates.

The SQS provides expanded sources of earnings for this application by allowing researchers to query 1099 NEC compensation and all combined 1040 wages, in addition to W2 income alone. The SQS also provides broader and more accurate linkage methods to increase the match rates and boost sample sizes. Results presented in this paper will compare the impact of using SQS statistics to evaluate non-degree credentialing programs in place of the credit bureau data.

3.3 Case Study 3: Comparison with a Convenience Survey

The final case study considers earnings outcomes for individuals who participate in a program with a training provider. The client, a workforce training provider, conducts their own RCTs to evaluate their programs by randomly assigning some participants to receive a newly proposed training program, with the remaining participants receiving the typical training. They evaluate earnings outcomes between the groups to determine the effectiveness of the proposed training.

This client currently evaluates their program by surveying participants and asking them to

report their own employment and earning outcomes. This has the advantage that the survey can gather additional information that may be used to analyze the outcomes. Conversely, this approach is susceptible to non-response bias or misreporting, which may be particularly problematic when it comes to asking participants to self-report their earnings. Additionally, fielding surveys is time consuming and expensive, so it may be challenging for the client to repeatedly utilize this approach for evaluating many proposed interventions.

The client can obtain earnings estimates from the SQS for both the treatment and control groups. These statistics should be based on larger samples, without suffering from non-response, and they should be more accurate (assuming individuals are more likely to accurately report their earnings to the IRS than to a survey). Results from this case study will directly compare estimates between the survey and the SQS, and we will consider how the data source affects the outcome of the RCT.

4 Conclusion

This paper presents the SQS and its ability to fill a growing need for more targeted employment and earnings statistics. The SQS is currently running live at IRS, and it has been designed with long-term sustainability in mind. We will evaluate the value of the this system through the use of three case studies, which will compare the SQS output to currently available alternatives. These represent the beginning of a larger effort, which we expect to grow to serve many non-profit organizations and state and local governments across the United States.

References

Berger, G., S. Seibel, S. Spitze, and S. Towns (2024). The importance of understanding non-degree credential quality. Report, Burning Glass Institute.

Dong, J., A. Roth, and W. J. Su (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 3–37.

Equifax (n.d.). Higher Education Data: Academic Research Data Set. https://assets.equifax.com/marketing/US/assets/academic-research-data-HED.PDF. Accessed April 10, 2026.

HolonIQ (2023, May 18). The Future of Post-Secondary Education in the US. https://www.holoniq.com/notes/the-future-of-post-secondary-education-in-the-us.

Internal Revenue Service (2021). Tax Information Security Guidelines for Federal, State and Local Agencies: Safeguards for Protecting Federal Tax Returns and Return Information. Technical Report Publication 1075 (Rev. 11-2021), Department of the Treasury, Washington, DC.

O'Hara, A., S. Straus, R. Borzekowski, P. Arnsberger, and B. Johnson (2024). A Secure Query System to Improve Access to Individual Income Tax Data. Working Paper w32969, National Bureau of Economic Research.

United States Congress (2019). Foundations for evidence-based policymaking act of 2018. https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf. Public Law 115-435, 132 Stat. 5529–5557.

Table 1: Available SQS Statistics

Description	**Population**	**Available Statistics**
Matched	Individuals with PII matched to IRS	Count
Form 1040	All individuals with filed 1040 in selected tax year	Count
Child Tax Credit (CTC)	Individuals with non-zero CTC amount on 1040	Count, Mean, Variance, Median
Earned Income Tax Credit (EITC)	Individuals with non-zero EITC amount on 1040	Count, Mean, Variance, Median
Wages	Individuals with non-zero Wages amount on 1040	Count, Mean, Variance, Median
Adjusted Gross Income (AGI)	Individuals with non-zero AGI amount on 1040	Count, Mean, Variance, Median
Schedule C	Individuals with a Schedule C on 1040	Count
Combined Info Return Income	Individuals with non-zero income on all W2s OR non-zero compensation on all 1099-NECs in selected tax year	Count, Mean, Variance, Median
W2 Only Income	Individuals with non-zero income on all W2s	Count, Mean, Variance, Median
1099NEC Only Income	Individuals with non-zero compensation on all 1099-NECs	Count, Mean, Variance, Median
Number of W2s Returned	Individuals with non-zero income on all W2s	Count, Mean, Variance, Median
Number of 1099NECs Returned	Individuals with non-zero compensation on all 1099-NECs	Count, Mean, Variance, Median
Info Return Location for Selected State	Individuals with non-zero income on all W2s OR non-zero compensation on all 1099-NECs in selected tax year	Counts (All Forms In-State, Some Forms In-State, All Forms out of State)