Statistical Use of U.S. Federal Tax Data

by Nicholas H. Greenia, Internal Revenue Service

"*The makers of the [U.S.] Constitution conferred the most comprehensive of rights and the right most valued by all civilized right to be let alone.*" Louis D. Brandeis, U.S. Supreme Court Justice

"…the rush to ensure complete levels of privacy in the research context paradoxically results in less social benefit, rather than in more… .[T]hrough the additional concept of utility, people will recognize that, while they surely have the right to privacy, they may also come to the realization that they have a duty to share information, if the common good is to be furthered." Peter Madsen, The Privacy Paradox

The importance of tax data to the Federal statistical system, in identifiable, anonymized, and publicly available form, is due to the rich financial information on both individuals and organizations. First and foremost, these data underpin the administration of the Federal tax system, which collected $2.5 trillion in tax revenue in Fiscal Year 2006, funding most Federal Government operations and public services. Obviously, this is the core purpose of tax data, but it is not the only one. A second purpose served by tax data is almost as important–namely, their role as inputs to statistical systems that inform analysts and policymakers both inside and outside of government. Since a system of administrative records already exists, there would be no additional burden to respondents, and minimal additional burden to taxpayers, if the tax record system were used for analytical purposes. In a time in which the cost of collecting data through surveys is skyrocketing, survey response rates are plummeting, and fiscal constraints are very evident, the importance of examining avenues for utilizing tax data is clear.

However, the use of tax data is predicated on the protection of taxpayer confidentiality. Because voluntary compliance is a cornerstone of the tax system, and is in turn based on the protection of taxpayer confidentiality, any usage of tax data–including statistical usage–that is even perceived to threaten confidentiality may be viewed as problematic. As a result, a longstanding issue for both the tax administration system and the statistical system is the determination of when a compelling need exists to use microlevel or identifiable Federal tax information (FTI) rather than aggregate and anonymized data. This paper describes the potential for FTI to describe economic activity. It then sketches the U.S. legal framework for permitting access to FTI, and the confidentiality challenges that have been addressed. It concludes with an outline of potential future directions for research.

▶ Data Description

The U.S. Internal Revenue Service (IRS) collects data[1] for a variety of entities: over 130 million individuals; over 50 million businesses[2], over 800,000 tax-exempt organizations; and approximately a million employee benefit arrangements, mostly retirement plans. Record level data for the population of filers are posted to one of several master files. The Business Master File contains the Form 1120 series, representing corporations; the Form 1120S and Form 1065, representing passthrough entities; the Form 941 series, representing employment tax returns; and the publicly available Form 990 series,[3] representing nonprofit and charitable organizations. The Individual Master File contains the Form 1040 series of individual tax returns, and the Information Returns Processing File contains the Form W-2 returns completed by employers but also filed by

This paper was presented at the International Seminar on the Use of Administrative Data for Economic Statistics and Register-Based Population Census on May 19-20, 2008, in Daejeon, Korea.

[1] The numbers presented here are for Tax Year 2005.

[2] Includes over 40 million sole proprietorships, represented by Schedule C and Schedule F filings.

[3] Not all of the Form 990 series data are publicly available; e.g., Form 990-T.

employees, mostly to record their taxes withheld, as well as the Form 1099 series to record investment income payments. Finally, the Employee Plans Master File contains records for employee benefit plans, especially retirement plans, such as the publicly available Form 5500 series.[4]

Tax return data, including complete balance sheets and financial statements, contain information on everything from net business profits to charitable contributions made by individuals. In addition, nonmagnitude information affords a variety of limited demographic data such as name and address, but also marital status–including previously divorced and widowed; number and names of dependents–including those living with a divorced spouse; unemployment status; status of retirement distributions; disability status–including blind; education expenditures; physical relocation; major medical expenses; and even military combat status. In short, it is apparent that tax return information could be useful for answering a whole host of important socioeconomic questions in addition to those directly related to tax.

Business data cover returns filed for corporations, partnerships, and sole proprietorships, and for both employers and nonemployers. Data are collected not only for these typical businesses but also for entities not typically thought of as businesses, such as governmental entities and nonprofit organizations, increasingly important economically due not only to the activities they perform but also to their role as employers. In addition, for corporations filing on a consolidated basis, it is possible to link subsidiary corporations with the parent, identifying corporate families.

The richness of the data stems in part from their universality; for example, because of the tax benefits accruing to businesses, it is in their interest to be captured by the tax system. This enables important measures, such as employment totals, to be compiled at the employer level through the employment tax returns filed by businesses (for example, the Form 941 series long used by the U.S. Census Bureau). Employment totals can also be compiled at the employee level and associated with related employers through SSN/EIN crosswalks (for example, using the Social Security number (SSN) and Employer Identification Number (EIN) captured from the Form W-2 used to report annual wage and salary payments and taxes withheld). Just as useful are other measures, such as firm entries and exits–critical data for any economic analysis of business formation as well as job creation and destruction. These are captured by the tax system, including whether a firm is an employer or nonemployer, or indeed, whether the economic unit is a commercial enterprise, a nonprofit organization, or even governmental in nature.

The richness stems also from the ability to link returns across economic entities. For example, the link between employers and employees captured from a crosswalk provided by the Form W-2 also enables linkages of detailed data on individual tax returns (including sole proprietorship returns filed with them) with employer tax returns. Similarly, because of dependent SSNs available on individual tax returns, it is possible to identify demographic families. In conjunction with the corporate families described above, it is possible to link the demographic population of workers, including a large segment of their family dependents, with the employer population.

Finally, the richness stems from data quality. The tax compliance program makes it possible to ensure a certain degree of data quality through legal requirements for not only the timely filing of returns, but also their accurate completion. Thus, the tax system is able to capture these business and demographic populations regularly–annually, quarterly, and even monthly for some returns. In sum, the variety of tax return filers, the financial and entity detail provided on their returns, the regularity of filings, the universe of coverage, and their data linkability make the tax system a potent resource for research and analysis. Nevertheless, even with a compliance program, the IRS database, like other databases, is imperfect in response quality. This inconvenient fact is supported by the most recent estimate of the tax gap–the difference between what is owed and what is paid—at $345 billion.[5]

[4] Not all of the return data are publicly available; e.g., Schedule SSA.

[5] For Fiscal Year 2001, see "IRS Updates Tax Gap Estimates," IR-2006-028, available at www.irs.gov/newsroom.

▶ Legal Framework for Statistical Uses of Tax Data

Access to administrative records is determined by societal needs at the time. For U.S. tax records, the rule of law has been inconsistent; for example, tax records were once made publicly available. However, in 2008, it is unmistakably clear that all access to tax data begins with statute. Even access by employees at the Internal Revenue Service (IRS) is governed by statute. For example, staff at the Statistics of Income Division (SOI) are authorized by statute[6] to access tax data to produce statistics of income both authorized and required by another statute.[7] This legalistic focus has long been recognized as the basis for tax data's confidentiality protection, but it has proven a challenge for other uses that tax data serve, including their important role in economic analysis.

Presently, any entity, including Federal statistical agencies, may access FTI only if a statute provides such authorization. This requirement is formidable, as it means that legislation containing such authorization has been proposed and passed by both bodies of Congress and signed into law by the President. For some recipients, e.g., the U.S. Census Bureau, Treasury regulations are also necessary, and may stipulate not only the specific purpose for which the FTI may be used, but also the specific tax items the recipient may receive. Although the process is less arduous than that needed to change the statute, a regulations amendment must still undergo scrutiny by both IRS and the Treasury Department, and requires approval by the Assistant Secretary of Tax Policy, often considered to be the nation's highest tax official after the Treasury Secretary himself. Both statute and regulation policy require that only the minimal amount of tax data be provided to accomplish an authorized task.

For even authorized recipients, there are also official protocols for provision of tax data that include official request letters at the departmental level, although delegation orders provide for some routine correspondence to be done at lower executive levels. Annual reimbursable interagency agreements may then be developed, allowing IRS to recover the costs of providing the data to recipients.

In sum, access to tax data–statistical or otherwise–must be authorized by statute. The core elements are summarized as follows for statistical analysis:

Statistics of Income

Under section 6108(a), the Secretary of the Treasury is required to direct the production of regular publications of statistics of income, based on tax return filings. These are produced by the IRS's statistical office, the IRS Statistics of Income Division (SOI), from annual stratified random samples of tax returns for the two major programs: the Form 1040 series for individual tax return data and the Form 1120 series for corporate tax return data. The resulting publications provided to the public take the form of annual complete hardcopy and electronically available reports, but tabular data from smaller studies, e.g., the Partnership program based on the Form 1065 series, may be published with articles in the quarterly *SOI Bulletin.* These and still other data are also posted directly to the Federal tax statistics Web site at http://www.irs.gov/taxstats. A listing of SOI projects and contacts can be found at http://www.irs.gov/taxstats/article/0,,id=169439,00.html.

Outside Requests for Special Tax Data Tabulations, Studies

Many user needs for tax statistics can be satisfied by the publicly available aggregate statistics described above. However, requests for different variations, e.g., classification categories, of these tabulations or even special studies involving the processing and production of new datasets and statistical analysis, can be made under section 6108(b). If accepted, these requests become con-

[6] Section 6103(h)(1) authorizes tax data access for tax administration purposes, which include statistical and research components.

[7] Section 6108(a) mandates the Secretary of Treasury to prepare and publish annual statistics with respect to the operation of the internal revenue laws, including various variables and taxpayer classifications.

tractual agreements and may be reimbursable in nature. Factors figuring in approval include disclosure considerations and the resource needs of SOIusually, but not always, the IRS function fulfilling such requests. The customers in such arrangements are permitted to receive only disclosure-proofed or anonymous data products.

Tax Administration, Statistical and Research Components

Section 6103(h)(1) authorizes access to confidential tax data for tax administration, which includes specifically recognized statistical and research components. Although official, discrete functions of the IRS are named for these responsibilities, namely, the Statistics of Income Division (SOI) and the Office of Research, other areas within IRS have come to maintain their own research functions in recent years. Although the proliferation of these statistical/research functions has also increased the amount of statistical analysis done with tax data, much of the non-SOI work is internally applicable to the main IRS mission of tax collection, including how to improve performance for compliance-related objectives.

Tax Analysis–Treasury Department and Joint Committee on Taxation

The analysis and estimation of tax consequences for legislation (both proposed and previously enacted) often require access to confidential tax data from either the SOI sample files or from population data on the IRS master files.

The two offices responsible for the lion's share of such analysis are essentially mirror images of each other. The Office of Tax Analysis (OTA) in the Treasury Department performs such analysis for the Executive Branch of Government, usually obtaining tax data under section 6103(h)(1), while staff for a tax-writing committee, the Joint Committee on Taxation (JCT)—so-called because the committee's 10 members are from both houses of Congress—performs this work for the Legislative Branch of Government and obtains tax data for such purposes under section 6103(f).

In addition, agents of tax writing committees, often for, or in conjunction with, JCT, may be designated to conduct analysis that requires access to FTI. Such agents include the Governmental Accountability Office (GAO) and the Congressional Budget Office (CBO).

Statistical Agencies

Under section 6103(j) only three of the fourteen major Federal statistical agencies are authorized to access tax data for purposes not related to tax administration *per se*: the U.S. Census Bureau and the Bureau of Economic Analysis (BEA) in the Department of Commerce, and the National Agriculture Statistics Service (NASS) in the Department of Agriculture.[8]

CBO—Although not a Federal agency, the Congressional Budget Office (CBO) is authorized to access FTI for purposes of long-term modeling to analyze the Social Security and Medicare programs. Most of the FTI used by CBO is obtained from the Social Security Administration (SSA)[9] along with other non-FTI data related to the SSA benefits under study.

Census—The U.S. Census Bureau receives the most FTI–in terms of both the volume of records and the number of variables–for satistical purposes, and by statute is authorized to receive both business and demographic data. FTI has a variety of uses, including the tracking of firm entries and exits, and in general

[8] In 2008, the thirteen major statistical agencies that acquire confidential microdata are: Statistics of Income (IRS); National Agriculture Statistics Service and Economic Research Service (Agriculture); Energy Information Agency; Office of Research Evaluation and Statistics (SSA); Bureau of Census and Bu reau of Economic Analysis (Commerce); Environmental Protection Agency; National Center for Health Statistics; National Center for Education Statistics; Bureau of Transportation Statistics; Bureau of Justice Statistics; and Bureau of Labor Statistics. The National Science Foundation is the fourteenth major statistical agency.

[9] Such sharing of FTI is enabled by section 6103(p)(2)(B) and the associated regulation.

is viewed by Census as the lifeblood of their business register. Because FTI is inextricably commingled with nontax data, the Census business register is subject to IRS safeguards review.

For both its demographic and economic surveys, Census uses FTI for sampling frame purposes, as well as for imputation and even some limited programmatic purposes, e.g., number of employees for its County Business Patterns program.

BEA—BEA is authorized to receive FTI only for corporations, and mostly uses microdata to support its sampling frame for foreign company surveys. Aggregate tax data are instrumental for tracking the nation's economic performance, including its balance of payments.

NASS—A limited amount of FTI for companies engaged in agricultural activities is received by NASS in order to conduct the quinquennial Census of Agriculture. Primarily, the data are used for validating its frame for mailing purposes, but, unlike as at Census, the data do not remain on its register.

Contractor Access, Tax Administration

Under section 6103(n), Treasury and IRS have the flexibility of engaging contractors, with access to FTI, for purposes of tax administration, including its statistical and research components. For example, the Federal Reserve Board uses a limited amount of tax data, for sampling frame purposes, in conducting the invaluable and longstanding Survey of Consumer Finances. For such access, FRB is officially a Treasury contractor under section 6103(n). Such purposes might also include helping to fulfill SOI's mandate under section 6108(a).

▶ Challenges in Protecting Confidentiality[10]

Protecting the confidentiality of tax data is challenging and expensive, for two reasons: there is no statute of limitations and the Tax Code treats all FTI the same with respect to confidentiality protection. That is, to IRS, a business name or address is as deserving of confidentiality protection as are income items for a large corporation or individual tax return, and all must be protected in perpetuity.

The challenge is to identify acceptable risk, and the approach utilized to date is taking steps that prevent reidentification of tax data through "reasonable means." The interpretation of reasonable means includes the use of reasonably available computer technology, mathematical/statistical techniques, and a working knowledge of the related subject matter. The reasonable means standard is a technology-relative concept and, thus, may be a moving target. Nevertheless, it represents an attempt at due diligence in balancing the two goals for tax data: their protection and their effective usage. The protection approach taken by IRS is two-pronged.

Part of the protection is physical in nature: statistical agency recipients of FTI must undergo regular onsite safeguards reviews that include examinations of not only physical and computer security systems, but also scrutiny of past uses.[11] These reviews confirm the recipients' understanding and implementation of the many requirements covering physical and computer security, data need and use, and appropriate documentation. Related requirements include separate systems for processing or accessing FTI and background checks on individuals accessing FTI within facilities certified for such purposes. All these requirements are intended to

[10] See Greenia, Nick, *The Release of IRS Data: Challenges and New Approaches*, IRS Reseearch Conference, 2002.

[11] Safeguards standards are described in *Publication 1075, Tax Information Security Guidelines for Federal, State and Local Agencies and Entities*, IRS.

preserve the confidentiality of FTI, whether maintained in its original form or commingled with data from other sources.

Part of the protection is legal. The access of FTI must be only for purposes authorized by statute, possibly supplemented with regulations and, infrequently, policy agreements. Proposed new uses of tax data may be scrutinized not only as part of the official interagency request correspondence process, but under a more formal review process established by IRS and the recipient agency.[12] This "need and use" review is another tool used by the agencies to ensure that due diligence is exerted, including documentation, for such accesses of tax data.

Given these constraints, the resource consequences of safeguarding taxpayer confidentiality over time are nontrivial. These constraints are exacerbated by the potential for complementary disclosure, or the reidentification of taxpayer data using indirect means, e.g., using data in other publicly released data to identify FTI related to a particular taxpayer. Given the ever-increasing public releases of tax and other data, the task of protecting FTI is daunting, especially over time.

Another part of the protection strategy is to minimize access: statutory policy on tax data authorizes provision of *the minimal* amount of tax data for an authorized purpose. This leads to an historical tension with statistical agencies, such as Census, since their mandate on administrative records is to *maximize* such usage. The tension may lead to friction unless a mutual agreement on process, protocols, and access parameters addresses the needs of each agency in the provision and usage of tax data. Sometimes, this agreement may result only after a catalytic crisis, followed by some period of "turbulence" and bargaining towards an equilibrium position.[13]

In sum, the tax system seeks to control or regulate the use of tax data by conceptually limiting, physically confining, and tracking such access in order to provide a documented audit trail that will withstand outside or third party scrutiny.

All of these constraints are driven by concerns for both actually protecting and being perceived as protecting taxpayer confidentiality, which is essential to preserving voluntary compliance, a critical cornerstone of the tax system.

Implicitly, both IRS and Congress recognize that this approach does not guarantee complete confidentiality, as the only means for such an absolute assurance would be not to release any data at all. However, since padlocking the treasure of tax data is viewed as neither a desirable nor a viable outcome, some disclosure risk is accepted as part of the necessary balance of protection and access. But there is clearly an overall limit on tax data access, even if that limit is not precisely known. The need for this limit can be attributable to both resource costs of protection and what might be termed the perception of a *plausible access quantity* limit. To see why such a limit makes sense, consider that even large amounts of safeguarding resources cannot enable unlimited access to FTI. The reason is credibility. It is simply not credible that unlimited access would ever pass a perceptions test on confidentiality protection, especially for third party scrutiny. That is, such an outcome would not seem plausible, as it would seem to turn the very concept of confidentiality on its head.

▶ Analytical Research, Future Directions

An increasingly sensitive subject in recent years has involved the role of tax data in analytical research, not

[12] For the U.S. Census Bureau, this process is described in *Criteria for the Review and Approval of Census Projects That Use Federal Tax Information*, effective September 15, 2000, at www.ces.census.gov.

[13] See Greenia, Nicholas H., "Developing Adoptable Disclosure Protection Techniques: Lessons Learned from a U.S. Experience," *Privacy in Statistical Databases*, CASC Project Final Conference, PSD 2004, Barcelona, Catalonia, Spain, June 9-11, 2004 Proceedings.

merely for producing the descriptive aggregate tables historically produced by statistical agencies. Through increased efficiencies, technological improvements have enabled the pursuit of different directions such as new access modes that include the Research Data Centers (RDCs) and remote access sites–in order to more fully inform decisionmakers in both the private and public sectors.

The value of such an access mode for analytical research has been to avail the researcher community of more microdata access, as well as to improve existing data collections through more rigorous use of the data, especially at the microdata level, that produce more precise understandings of the data, including their limitations and future needs for improvement. As beneficial as this new direction may sound initially, one problem is that it is not always clear that existing statutory language and supporting regulations and policy statements support it. In addition, it is perceived to expand access to tax data, which raises at least the question of whether taxpayer confidentiality is still being protected.

A case in point is the Census Bureau's establishment of RDCs as part of its Center for Economic Studies program, which have access to business data at the microlevel. Given the ubiquity of tax data in Census files, this access produced some unease, given that the statutory language of section 6103(j) in Title 26, U.S. Code, had not changed in decades. That is, tax data were still being provided to Census for

> "…the structuring of censuses…and conducting related statistical activities authorized by law."

Fortunately, Title 13, the Census Bureau's statute, was more flexible, enabling authorizations to change according to the needs of the Census Director. In this instance, resolution was achieved through interagency cooperation, namely, SOI proposed, and both agencies crafted, a policy statement by the Census Director that, in essence, codified past practice at the RDCs through its explicit–albeit, slightly delayed–recognition of the vital role played by analytical research in order to accomplish its Title 13 mandate.[14] This clarification helped modernize and synchronize consistency with the Treasury Regulations on the authorized purpose of Census access to tax data, namely, for purposes of Title 13. This policy statement is insufficient alone, but has helped strengthen the interagency relationship in conjunction with continued compliance by Census on all IRS safeguards reviews, and the continued application of the interagency review process for all new RDC research proposals. As a result, the program has succeeded in being an example of what is possible through an interagency cooperative model that emphasizes, especially through its documentation, the *demonstrable credibility* of the process.

Another reason for concern about analytical research is an exclusive mandate on tax policy analysis–involving access to FTI–for proposed legislation that seems to be afforded only certain groups, e.g., the Joint Committee on Taxation (JCT) in Congress and the Office of Tax Analysis (OTA) in Treasury. One concern by both groups is that they may be "blindsided" on some controversial issue by the conclusions of outside researchers with access to tax data, particularly if they have access to too little tax data to be completely informed about the tax policy issue under consideration. The legislative process is complicated and stressful enough without such outside factors intervening or even appearing to disrupt the process.

There is some justification for a different view on outside researcher access, however. Namely, given that statistical tax data are produced with publicly provided funds and given that publicly available data are often too imprecise–due to their anonymous form after disclosure processing–to answer questions so important for all citizens, why should there not be an outlet for such access? In effect, this might be seen as an additional systemic check for purposes of further "democratizing" the decision process. Although some might argue that the electorate has already spoken with the process currently in place, including statute, several factors might argue for at least some additional access by outside economists and analysts.

[14] See Memorandum, Analytical Research Policy, by Charles Louis Kincannon, January 4, 2007, available at www.ces.census.gov.

First, the analytical questions being answered by JCT and OTA are largely driven by the political discussion in Congress and the Administration. It is possible that other questions–not being actively pursued–might also be relevant to the public debate, and this role might be played by outside researchers, on a carefully controlled basis. Second, analytical resources are so heavily burdened at JCT and OTA that there is sometimes less than optimal opportunity to ensure that analytical results are comprehensively accurate. Third, both JCT and OTA may call upon outside researchers to assist them with their analysis, but such requests are strictly at their discretion. Perhaps a different type of "third party" scrutiny would be provided by more outside researcher analysis, especially if it might be viewed as helpful, not divisive or destructive, to the ultimate decision-making process.

Finally, the increasing capacity of technology itself, including more powerful computers and techniques for both research analysis and confidentiality protection, may argue for some expansion in tax data access. The case for expanded access could be compelling–if the likely outcome is more informed decisions in both the private and public sectors, with increased utility for society as a whole, and if that can be accomplished without sacrificing privacy.

▶ Summary

This paper outlined the potential for FTI to describe economic activity, but noted that the use of tax data is predicated on the protection of taxpayer confidentiality. It summarized the legal framework within which microlevel federal tax information (FTI) could be used rather than aggregate and anonymized data. It concluded with an outline of potential future directions for research.