Chapter 9: Comparing Administrative & Survey Data

Consider the Source: Differences in Estimates of Income and Wealth From Survey and Tax Data

Barry Johnson, Internal Revenue Service, and
Kevin Moore, Board of Governors of the Federal Reserve System

One implication of the decentralized nature of the statistical system in the United States of America, composed of over 70 Federal Government organizations, is that the data used by lawmakers and researchers to develop and evaluate Government policies come from a variety of sources. Survey and administrative data sources are frequently blended to create information systems capable of supporting a variety of research purposes. Because these two types of data are primarily designed for different purposes, one inherently created for research and the other for administration of Government programs, blending them generally poses serious challenges. This paper examines the comparability of administrative and survey data, focusing specifically on data from Federal income and estate tax returns collected by the Statistics of Income (SOI) Division of the U.S. Internal Revenue Service (IRS), and the Survey of Consumer Finances (SCF) sponsored by the Board of Governors of the Federal Reserve System. Through the use of two case studies, we detail key similarities and differences between these two data sources and demonstrate methods for reconciling estimates produced from them.

We then briefly discuss the Statistics of Income program and the Survey of Consumer Finances. We also discuss in detail differences between administrative and survey data, using administrative data from tax returns and SCF data to illustrate key points. We then present detailed comparisons of wealth estimates derived from U.S. estate tax returns and from the SCF, followed by a section comparing estimates from U.S. income tax returns and the SCF. The final section summarizes key points.

▶ The Statistics of Income Program

The Statistics of Income Division of the Internal Revenue Service was established almost immediately after the adoption of a Federal income tax in 1916 and was charged with the annual preparation of statistics with respect to the operation of the tax law. The first SOI report, based on income tax returns filed by individuals and corporations for Calendar Year 1916, was released in 1918. From the very beginning, SOI reports were almost entirely used for tax research and for estimating revenue, especially by officials in the Office of Tax Analysis of the Department of the Treasury and in the Congressional Joint Committee on Taxation. In the 1930's, a third major user of SOI data was added, the Bureau of Economic Analysis in the Department of Commerce, which uses SOI data extensively in constructing the National Income and Product Accounts. As the SOI program and products have expanded, users in other Government agencies, such as the Census Bureau, as well as many private and academic researchers, have come to rely on tax data produced by SOI for evaluating tax policy initiatives (see Wilson, 1988 for a complete history of the SOI program).

In order to fulfill its charge, SOI created a structured mechanism for transforming administrative data into statistical files, using its own data collection systems, completely autonomous of main IRS tax return processing. SOI currently conducts approximately 110 different projects involving data collection from returns and information documents; this paper will highlight two of these projects, the individual income tax file (ITF) and the estate tax data file (ETD). Data content is developed working closely with data users so as to ensure both continuity and usefulness. For most studies, data are extracted from stratified random samples of returns as they are filed to ensure timeliness. Specially trained employees located in IRS submissions processing centers collect the data under the supervision of subject matter experts from SOI headquarters. These specialists supply data editing instructions, conduct training classes, and review difficult cases. Data are entered into computer databases and checked using embedded tests that verify coded values and key mathematical relationships. In addition, subsamples of edited returns are subjected to field-by-field quality review. Finally, subject matter experts carefully review all files for accuracy before they releasing them to customers.

▶ Survey of Consumer Finances

The SCF is a survey of household balance sheets conducted by the Federal Reserve Board in cooperation with the SOI division of the IRS. Beginning with 1983, the survey has been conducted triennially, with data collected by the Survey Research Center at the University of Michigan in 1983, 1986, and 1989, and by NORC, a national organization for social science and survey research at the University of Chicago, from 1992 forward. Besides collecting information on assets and liabilities, the SCF collects information on household demographics, income, relationships with financial institutions, attitudes toward risk and credit, current and past employment, and pensions (for more details on the SCF, see Aizcorbe, Kennickell, and Moore, 2003).

The SCF uses a dual frame sample design to provide adequate representation of the financial behavior of all households in the United States. One part of the sample is a standard multistage national area probability sample (Tourangeau et al., 1993), while the list sample uses the IRS-SOI Individual Taxpayer File (ITF) to oversample wealthy households (Kennickell, 2001). This dual frame design provides the SCF with efficient representation of both assets widely held in the population, such as cars or houses, and assets more narrowly held by wealthy families, such as private businesses and bonds. Wealth data from the SCF are widely regarded as the most comprehensive data available for the United States.

Sample weights constructed for the SCF allow aggregation of estimates to the U.S. household population level in a given survey year (Kennickell and Woodburn, 1999; Kennickell, 1999). Missing values in the 1983 and 1986 SCF were imputed using a single imputation technique, while missing values in the subsequent 1989-2001 SCF were imputed using a multiple imputation technique (Kennickell, 1991, 1998b).

▶ Administrative Records and Survey Data

The American Statistical Association (1977) defines an administrative record as "[data] collected and maintained for the purpose of taking action on or controlling actions of an individual person or other entity." In the U.S. Government, administrative records have a long history of use in the production of Government statistics. In recent years, technological advances have made it easier for statistical agencies to process large datasets, encouraging even greater use of administrative records for research purposes. As a research tool, administrative records have many potential uses, including direct tabulation and indirect estimation of models or other statistics, as well as construction of survey frames and evaluation of survey results (Brackstone, 1987). In the best situations, administrative data may have several advantages over traditional survey data, including more complete coverage of a population (sufficient for regional statistics), low data collection costs, reduced respondent burdens, and better data quality. The potential problems with using administrative data for statistical purposes include the stability of a program over time, privacy concerns about nonadministrative use of data, conceptual issues relative to the population and items collected, and costs of transforming the data into a form useful for research purposes.

Surveys differ from administrative data in terms of their purposes, and such differences often have implications for their statistical structure, conceptual framework, and content. Almost all surveys are conducted to answer specific classes of research or public policy questions versus fulfilling an administrative function. This difference in purpose is reflected in the population frame, the unit of observation, the sample size, and the scope of the data. Some advantages of survey data over administrative data include the targeting of a specific population and variables of interest, the interaction with the respondent, and the ability to pledge that the data will be used solely for statistical (that is nonadministrative) purposes. Potential problems with survey data include difficulties in constructing a suitable frame, lack of legally mandated participation, high costs of increasing sample size, unit and item nonresponse, and measurement error. The following sections will examine all these issues in more detail.

Frame Issues

The population covered by a system of administrative records is defined through legislation, based on the scope of the program the records are intended to sup-

port. Often this population is truncated in some way, restricted based on specific demographic or economic characteristics. In some cases, individuals may have to take some action to become part of the administrative system (e.g., filing a tax return); so, it is important to consider what incentive there is for individual units to be registered. There may be perceived advantages for some individuals to evade registration, particularly if their circumstances place them at or near a threshold requiring mandatory participation. The populations of both Federal income and estate tax filers, for example, include only those U.S. citizens and resident aliens whose gross incomes, or gross estates, concepts defined by statute, were above specified thresholds. For each tax system, nonresident aliens are subject to different filing requirements, based on income earned or assets owned in the U.S. Income tax filers represent roughly 61 percent of the U.S. individual population, while estate tax filers have generally represented fewer than 5 percent of total annual U.S. deaths (see Sailer and Weber, 1999; Johnson and Mikow, 2002). Recent income tax filing gap estimates for Tax Year 2000 suggest that as many as 11 million taxpayers, or about 9 percent of the potential income tax filing population, either file returns late or not at all (see Brown and Mazur, 2003).

The Federal Committee on Statistical Methodology's (FCSM) *Statistical Policy Working Paper 6--Report on Statistical Uses of Administrative Records* points out that the unit of observation useful for statistical purposes often focuses on the attributes of groups of individual entities, while administrative records are often focused on identifying specific entities in order to take some sort of action based on their individual characteristics. Thus, the unit of observation available from administrative records may make certain research difficult or impossible. Records may contain information about individuals rather than families or households, or may be a mix of both individuals and households. In the case of Federal income taxes, married couples may file returns jointly, but they are also allowed to file separately in cases where marginal tax rates favor treating the two incomes separately. Dependent children and others living in a home may also be required to file separate returns to report both earned and unearned income. Differences in the economic unit reported on income tax returns limit the data's usefulness for some types of research. Similarly, Federal estate tax returns represent only the decedent's wealth, including one-half the value of all community property [1] and property held as joint tenants [2]; assets owned independently by a surviving spouse are not reported.

The population targeted by a survey is determined by the purpose of the survey, the availability of a sampling frame, and the cost of the sample. The sampling frame for most surveys is derived from existing sources, such as geographically based population data, address listings, telephone directories, or administrative sources. Often, one of the most difficult issues with designing a survey is finding an appropriate frame (Lessler and Kalsbeek, 1992). Selecting the wrong sampling frame may lead to issues of undercoverage and may bias any results obtained from the survey data. A related problem arises if a survey targets a population that is difficult to locate or measure.

Directly related to the availability of a sampling frame is the potential cost of obtaining the frame information and the cost of interviewing a sample of the desired size. For target populations that are difficult to locate or appear infrequently in the frame, the cost of simply increasing the sample size to obtain better coverage can be prohibitive, although, sometimes, a frame contains information that may be used to target rare groups more efficiently. For example, one of the main goals of the Survey of Consumer Finances (SCF) is to measure the wealth of U.S. households. However, because wealth is highly concentrated in the population, sufficient coverage would require a very large area-probability sample. To this end, the SCF uses a dual-frame sample design in which an oversample of "wealthy" households is targeted using statistical records derived from tax returns provided by SOI [3]. Use of this sampling frame allows the SCF to collect data from wealthy households in a cost-effective and statistically efficient manner.

For survey data, the unit of observation is usually determined by the type of data required to answer certain research or policy questions. However, the choice of the unit of observation is also influenced by the type of sampling frame available to survey designers. In the SCF, the area-probability sample uses a sampling

frame in which the household is the unit of observation, but, for the list sample, the unit of observation is the tax-filing unit. Often, the tax-filing unit is analogous to the household, but, for certain households, such as households where a married couple files separately and households with multiple subhouseholds located within a household, there are differences. While there is the possibility of frame errors in the list sample, adjustments are made during the construction of the frame and during the sampling stage to limit the distortions (see Kennickell and McManus, 1993; Frankel and Kennickell, 1995; Kennickell, 1998a; and Kennickell, 2001).

Content Issues

The purpose for which administrative records were collected can have a profound effect on their usefulness for statistical purposes in terms of the amount of data available, data definitions, year-to-year consistency, and quality of the data. Many times, the usefulness of administrative record systems is limited because only those variables needed to administer the program are collected. These variables may be only a small fraction of the data reported on an administrative form.

In addition, because program requirements are established by legislation, data concepts and definitions used to meet program needs may not necessarily coincide with those required for social or economic analysis (Brackstone, 1987). For example, income for married couples is combined for joint filers of U.S. income tax returns; however, for some research purposes, it would be useful to know the amounts earned by each individual. When research and administrative needs differ, it can be very difficult to affect changes or improvements in content since statistical uses are often seen as secondary to an agency's primary purpose (FCSM Working Paper 6). This can pose serious limits on the overall usefulness of administrative data systems or require that the administrative agency undertake additional data collection and/or editing, incurring costs and delaying data availability.

Another consideration is that, while administrative records have much potential as a source of information on small geographic areas, to be useful, a precise geographic location code is needed. However, mailing addresses, frequently present on administrative records, may not always be the appropriate location, as when a post office box number is supplied rather than a street address. For Federal tax returns, addresses might be those of paid preparers rather than filers. In some instances, a filer may even own several residences.

An important aspect of data content is continuity over time, both in the items included and in the data definitions. Coverage and content in administrative records systems can be subject to discontinuities resulting from changes to laws, regulations, administrative practices, or program scope (Brackstone, 1987). For example, income tax law revisions in 1981, 1986, 1990, and 1993 all made significant changes to both the components of income subject to taxation and the allowable deductions from income that had significant impact on the statistical uses of tax return data (see Petska and Strudler, 1999). More recent changes in tax law will incrementally increase the filing threshold for estate tax return filers, from $675,000 in 2001 to $4,000,000 by 2009, and then abolish the tax entirely in 2010.

Data quality may also be a concern in administrative records systems. FCSM Working Paper 6 cautions that there can be considerable variation in quality across variables in an administrative records system. Information that may be statistically important, but only marginally relevant to administrative purposes, is often imperfectly reported, checked, and processed. Data items used primarily as background information may be of particularly low quality or even incomplete. This can also be the case for data collected specifically for statistical purposes using existing administrative channels. These items may be of lower quality if their priority is not very high to the administering authority or to the subject supplying the information (Jensen, 1987). Finally, data reliability may also be affected if the information respondents provide may be used to cause gains or losses to individuals or businesses. Underreporting on tax returns, for example, may have resulted in underpayment of as much as $120 billion in income taxes and $3.5 billion in estate taxes for Tax Year 1998 (Brown and Mazur, 2003).

FCSM Working Paper 6 suggested that administrative records sources are often a reliable source of timely data produced with predictable frequency. However,

since data collected and processed for administrative purposes are generally given priority over those required for statistical purposes, the amount of postprocessing required to render administrative data suitable for statistical purposes may affect data timeliness. In addition, the time and difficulty required to create desired statistics can vary considerably depending on a variety of factors. For example, for some research purposes, income data for households, rather than individuals, are required. To reconstruct households requires linking information documents with income tax returns filed by dependent filers and married couples who filed separately, using unique taxpayer identification numbers, all at the cost of significant resources (see Sailer and Weber, 1996).

Because surveys are freer than administrative systems to specify a conceptual framework, many issues related directly to the definition and scope of the data are less pressing. However, content and valuation issues of a different sort are present in survey data. One key issue is the voluntary nature of response to surveys versus the legally mandated participation in most administrative data programs. In most surveys, interviewers (either in person or via telephone) attempt to convince respondents to voluntarily donate time and information when there may be no direct benefit or punishment if a respondent refuses. Even if a respondent agrees to participate in the survey, it is still possible that the respondent will refuse to answer the questions truthfully and completely. Unit and item nonresponse are two important sources of nonsampling error in surveys; however, there are methods to help deal with both these issues, such as weighting and imputation.

For respondents who agree to participate and answer all the survey questions, measurement error is still a concern in survey data. Respondents may "guestimate" answers to questions; even if respondents' guesses overall are unbiased, such approximation reduces the estimation efficiency of the data. Respondents may also have difficulty recalling past events. Other typical measurement errors include rounding of dollar amounts, misunderstanding questions, and altering responses due to stigma or prestige attached to certain behaviors or a desire to protect privacy. A large volume of research exists on measurement error and its effects on survey data (see Lessler and Kalsbeek, 1992 and the references within).

While it is true that, for administrative data, unit and item nonresponse are usually not a problem on core items, it is not clear that administrative data are always more accurate than survey data. An example is the income values reported on IRS tax forms versus the income values reported in survey data; some individuals may intentionally misreport values on tax returns to reduce their tax liabilities. Those same individuals may report the true value in response to a survey question since there is no benefit to misreporting in the survey (via a lower tax liability).

Another content issue for survey data is the timeliness of the data. While many simple surveys are administered quite frequently, such as monthly, most of the more complex surveys occur yearly or even less frequently. Cost and other resource constraints are major factors in the timeliness of the survey data. For example, due to the high cost, complexity, significant data processing, and high respondent burden, the SCF is conducted on a triennial basis.

A final content issue for survey data is validation of the data. While it is sometimes possible to conduct validation studies after a survey is complete, these studies add additional cost to the survey. Validation of some items might require the cooperation of respondents, and requesting such cooperation may trigger suspicions in respondents that might lead to overall lower cooperation with a survey. Sometimes, selected data items are validated against external data sources, such as the Census or administrative data, but, often, no source for validation exists. This is in contrast to some administrative data, such as wages reported on tax forms, where amounts reported by filers are validated against amounts reported by their employers.

Privacy Issues

Any use of administrative records for research purposes must take account of laws protecting data privacy. In the U.S., privacy protections are either spelled out explicitly in agency-specific confidentiality statutes and regulations, or derived from Governmentwide statutes, such as the Privacy Act of 1974 (5 U.S.C. § 552a), and more recently, the Confidential Information Protection and Statistical Efficiency Act of 2002 (44 U.S.C. §

3501) (CIPSEA). In both instances, research uses of administrative data are often restricted to uses within the scope of an agency's mission and must be conducted by persons working for the agency as employees, contractors, or under the Government's Interagency Personnel Act (5 U.S.C. §§ 3371-3375) provisions that allow State government and nonprofit organization employees to work under the same provisions as employees as long as certain conditions are met. Other researchers are usually limited to public-use data sets or data tabulations, for which great care is taken to minimize the possibility of reidentifying data related to specific individuals. Public perceptions of privacy protection are vitally important to maintaining the goodwill required to sustain compliance levels, especially for agencies, like the IRS, which rely heavily on voluntary compliance for the success of their programs.

Government survey data are also often protected by the various privacy and confidentiality laws that apply to administrative data. The confidentiality of the respondent's data is of paramount importance to the current and future success of any survey. If respondents do not believe their data are sufficiently protected, both response rates and the overall data quality in the survey will suffer. Confidentiality and privacy laws provide important safeguards against potential abuse of respondent data by survey sponsors. In addition, surveys that produce publicly available data sets also must engage in a disclosure review to safeguard the identity of the respondents. The data collected during the SCF are protected by the Privacy Act of 1974, CIPSEA, and the Internal Revenue Code through an agreement with SOI. Information on the SCF disclosure review process is detailed in Fries (2003).

▶ Wealth Data

Both the SCF and Federal estate tax return data (ETD) provide important sources from which to study privately held wealth in the U.S. Both data sources collect extensive information on real estate, financial assets, businesses, tangible assets, and debts. The SCF also contains demographic information on household members, as well as extensive income and pension data. Federal estate tax returns provide a more limited demographic profile of the decedent, information on the costs of administering the estate, and data on bequests to charities, the surviving spouse, and other living persons. Figure 1 provides a comparison of data available from both sources.

While there are many similarities between types of data available from the SCF and ETD, there are important structural differences. Some of the most significant include unit of observation, population coverage, and sample size. The SCF is a household survey which uses as its core unit of observation the "primary economic unit," which can consist of a number of different social arrangements, most commonly married or partnered pairs of individuals, and single persons, including those who were widowed, separated, divorced, or never married at the time of the survey, and all others in the household who are considered interdependent with them. Individuals living in institutions, such as nursing homes, are excluded from the area probability portion of the sample but may be in the list sample. All but the very wealthiest households, those with total assets of more than $600 million, are included in sample population [4]. The unit of observation in ETD is always an individual, and the population is limited to individuals with gross estates above the filing threshold applicable on the date of death, $675,000 for 2001 decedents [5].

One of the strengths of ETD is the large sample size. For example, the 2001 estate tax decedent file includes 17,376 records for individual decedents with total assets of at least $675,000. Of these, 9,322 were married, while 8,054 were widowed, single, divorced, or separated. The SCF includes 1,531 households with this level of wealth, only about 200 of which were either headed by widowed, single, divorced, or separated individuals. The large ETD sample size allows reasonably precise estimates for specific demographic groups, as well as geographic estimates by region or state.

While population estimates of wealth from both the SCF and ETD are based on weighted samples, there are significant differences in the method used to calculate the sample weights, which may have an impact on estimates derived from each source. Sample weights for the SCF are calculated using information from the sample design and are constrained using known population totals. Estimates of wealth from ETD rely on a multiplier which

Figure 1: Comparison of SCF and ETD File Content

Variable	Estate Tax Data	Survey of Consumer Finances
Demographic data:	Name, State of residence, year of birth, year of death, marital status, occupation, surviving spouse, (children, others if heirs) previously deceased spouse--year of death, name	State, year of birth, age, marital status, years married, previous marriage information, educational attainment, occupation, household characteristics including age of spouse, number of children, other dependents, age of parents
Real Estate:		
Personal residence	Single family, multiunit, ranch, mobile home; lot size; value (usually from real estate appraisal valued on date of death); mortgage amount	Single family, multiunit, ranch, mobile home; length of time living there; number of acres, value; mortgage type, amount, payment information; rent received
Rental property	Single family, multiunit, ranch, mobile home; lot size; value (usually from real estate appraisal valued on date of death); mortgage amount	Single family, multiunit, ranch, mobile home; length of time owning; value; rent received
Farm property	Value; acreage; mortgage amount	Value; acreage; mortgage type, amount, payment information
Financial Assets:		
Closely held stock	Name of corporation; number of shares; percentage ownership; market value; appraisal	Actively managed: number of businesses, for 3 largest: year formed, type, cost, method of financing, value, income received. For others: total value, cost, income. Nonactively managed: value, cost, type, income received
Publicly traded stock	Number of stocks, market value, name of corporation, brokerage account information	Number of stocks, market value, gain or loss, location (in the U.S. or not) employer stock (yes or no), brokerage account information
U.S. Government bonds	Market value	Face value, market value
Federal Savings bonds	Market value	Face value
Tax-exempt bonds	Market value	Face value, market value
Corporate bonds	Market value	Face value, market value
Mutual funds	Type of fund (stock funds, tax-exempt bond funds, Government-backed bond funds, other bond funds, combination or mixed funds), value	Type of fund (stock funds, tax-exempt bond funds, Government-backed bond funds, other bond funds, combination or mixed funds), type of institution, value, gain or loss since purchase
Noncorporate Businesses	All businesses, active, nonactive. Value at death, appraisals or balance sheets.	Actively managed: number of businesses, for 3 largest: year formed, type, cost, method of financing, value, income received. For others: total value, cost, income. Nonactively managed: value, cost, type, income received
Trusts	Revocable trusts, marital trusts: detailed listing of assets, value. Split Interest trusts: value, assets invested, charitable beneficiary. Other income trusts may not be reported.	Type (income only, equity), amount of annual income, value, indication of how assets are invested
Bank accounts	Type of account (money market, traditional savings, certificate of deposit), current balance, ownership	Type of institution, type of account (money market, traditional savings, certificate of deposit), current balance, ownership
Life insurance	Face value, accrued interest, policy loan amount	Term and whole life: face value, cash value, policy loans (purpose and payment information), premiums

Mortgages and notes	Amount owed to decedent	Amount owed to respondent
Retirement assets:		
Annuities	Equity: value, detailed listing of assets. Income not usually reported unless there is a death benefit or lump sum value.	Type (income only, equity), amount of annual income, value, indication of how assets are invested
401K, Keogh, etc.	Number of accounts, value. Detailed listings of investments are usually provided	Type (education, Roth, Keogh, rollover), number of accounts, type of institution, value
Pensions	Only pensions where surviving spouse is also a recipient so that a portion is included in the taxable estate	Detailed information on pensions from multiple jobs for primary economic unit including type, contribution amount, benefit amount, timing of payments, death benefits, etc.
Social Security Payments	Not reported	Amount received, reason for payment
Other:		
Art/antiques/collectibles; Depletable/ intangible, livestock, proceeds from lawsuits, lottery winnings, futures	Type, amount	Type, amount (up to three different categories)
Vehicles/boats/etc.	Type; value for all vehicles; model and year usually supplied for automobiles; loan amount	Automobiles: first 4--model, year, financing, value, purchased new or used, Others: financing, value. Other vehicles: first 2--type, financing, value, purchased new or used, Others: financing, value.
Debts:		
Consumer debt	Amount owed	Amount of original loan, type, payment information, balance owed, purpose, collateral, type of institution, payment history
Mortgages	Amount owed	Amount of original loan, type, payment information, balance owed, type of institution, payment history

incorporates both the probability of being selected into the SOI sample of estate tax returns and the age and sex-specific probability of being a decedent in a particular year (see Atkinson and Harrison, 1978, for a description of this methodology). Mortality rates, by age and sex, are used to approximate the probability of being a decedent. Because there is no way to control for the weighted population total, the selection of an appropriate mortality rate is important. Research has shown that the wealthy live longer than the general population due to factors such as access to better health care, safer work environments, and better nutrition. While estimates of patterns of wealth holding appear quite robust over a variety of reasonable alternate assumptions about the longevity of the very wealthy, overall aggregate estimates are relatively sensitive to the selection of the mortality rates. Mortality rates calculated for holders of large dollar value annuity policies are used for these estimates.

Valuation Issues

There are significant differences in the determination of asset values in the ETD and SCF. Estate tax returns are generally accompanied by a great deal of documentation to support reported valuations, including tax returns, brokerage account statements, appraisals, business accounting reports, and legal documents. In contrast, only about 32 percent of SCF respondents use such documents when providing valuation data, although extremely wealthy survey respondents often refer to financial documents or seek assistance from their accountants in order to provide accurate data.

While the more systematic presence of valuation documentation may make ETD a potentially more accurate source of wealth data than survey estimates, the administrative nature of ETD imposes important considerations. Unlike questions on the SCF that have been

carefully constructed to capture data needed for specific research purposes, data reported on estate tax returns are influenced by provisions in the tax law, estate planning mechanisms, and the point in the life cycle at which data are collected. For example, the tax code allows certain adjustments in asset values, such as the special valuation of real estate used for farming or certain business purposes, and includes some items, particularly the face value of life insurance and trust property over which a person had a limited power of appointment, that might not ordinarily be considered part of lifetime wealth [6]. In addition, the tax code generally exempts from tax other wealth to which a person has an income interest, but not necessarily actual title, such as defined-benefit pension plans, simple trusts, and Social Security benefits.

A number of other factors can contribute to differences in the values of assets captured in the ETD and those collected on the SCF. While estate tax returns are generally prepared by professionals and are, therefore, likely to be more precise in detail than survey responses, the values are used to compute tax liability; so, there is a natural tendency for the values to be as conservative as legally permissible. This is especially true for hard-to-value assets, such as businesses and certain types of real estate. It should also be noted that the ETD collected by SOI are pre-audit figures. While we believe that the relatively high audit rate for estate-tax returns ensures that complete evasion is relatively rare, the values reported may be subject to underreported and missing values, the later due to informal transfers of small items such as jewelry [7]. In addition, it is common to claim substantial discounts when valuing ownership interests of less than 50 percent in small companies, partnerships, and other nonliquid assets. The creation of family limited partnerships and other estate planning techniques can significantly reduce the asset values included in a decedent's estate by taking advantage of these discounts [8]. Finally, the wealth of some estate tax decedents may differ significantly from that of the general population in the same age cohort, due to expenses related to final illnesses. In addition, when death is anticipated, decedents may have altered the composition of their assets in order to simplify their finances, to provide liquidity to pay for health-related expenses, and to ensure that family-owned business operations are not disrupted by their deaths.

Direct Comparisons Between SCF and ETD Data

The study of wealth includes many goals, only one of which is the determination of point estimates for various populations and subpopulations. The previous section pointed out important structural differences between the SCF and ETD. A key research question then is do these two datasets provide similar analytical results, despite these differences? Focusing on total assets as the measure of wealth, the SCF data show that there were more than 13.4 million households with total assets of $675,000 or more, while the ETD data show that there were more than 6.1 million individuals at or above that wealth threshold. The mean age for heads of household in the SCF was 56, and the median age was 54. For ETD, the mean and median ages were both 60 [9]. Estimates for widowed, single, separated, or divorced persons provide the best opportunity for direct comparisons between the two datasets since the units of observation should be closely aligned. Figure 2 provides a direct comparison of wealth components for the SCF and

Figure 2
Comparisons of SCF and Estate Tax Data Estimates of Wealth, by Marital Status, for Households or Estates with >= $675,000 in Assets
(Money amounts are in thousands of dollars)

	Survey of Consumer Finances				Estate Tax Estimates			
	% reporting	Mean	Median	Total	% reporting	Mean	Median	Total
Single/widowed/div/sep								
Total assets	100.0	2,102	1,099	4,564,262,000	100.0	1,833	1,068	4,822,014,000
Financial assets	100.0	1,122	653	2,435,399,000	100.0	1,189	745	3,108.671,000
Nonfinancial assets	98.5	980	488	2,128,862,000	96.0	678	343	1,713,343,000
Personal residence	85.0	286	230	620,366,000	67.1	320	240	564,534,000
Other real estate	50.7	270	17	586,918,000	36.1	386	215	367,051,000

Note: SCF and ETD estimates are based on samples.

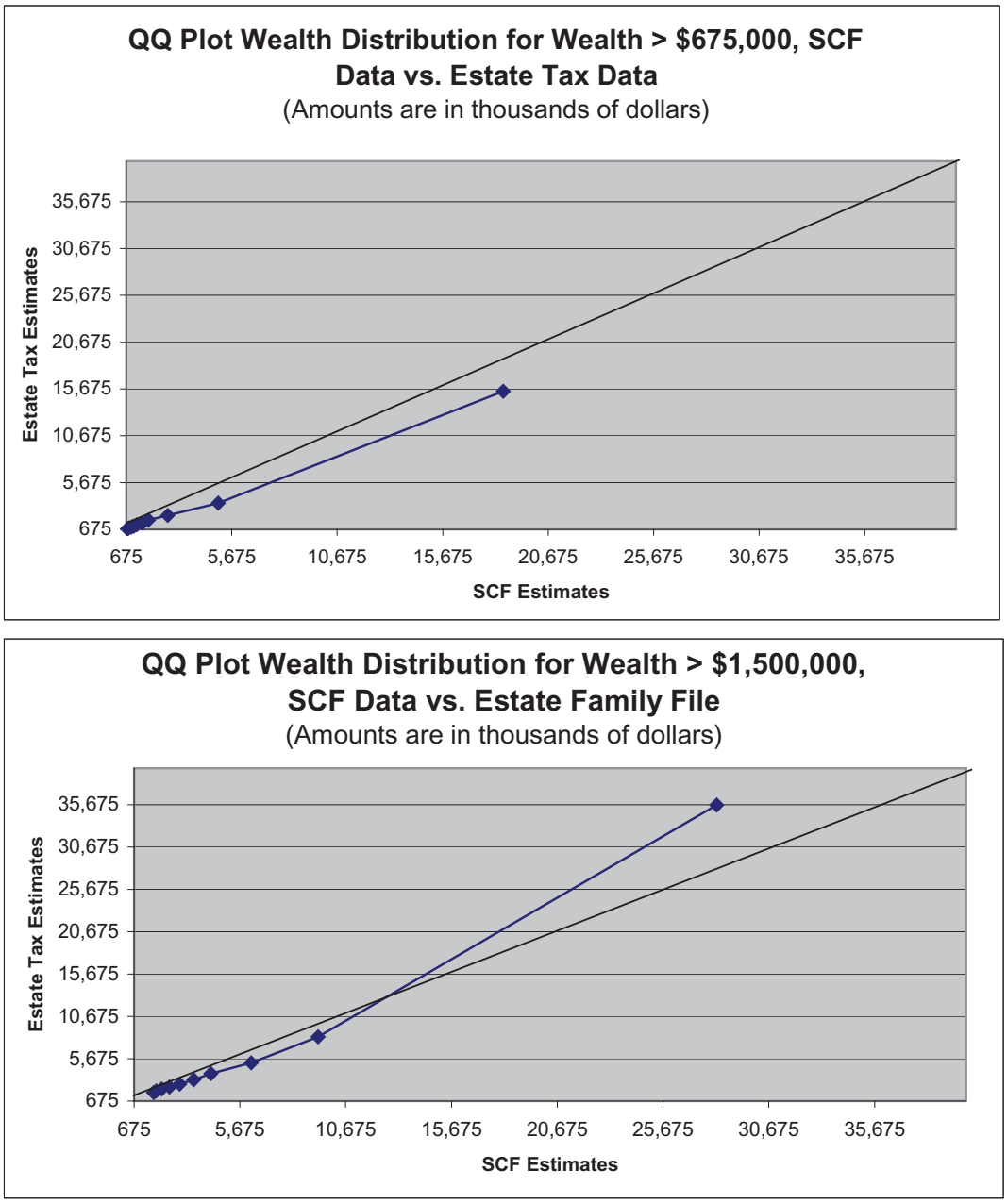
ETD, for unmarried or unpartnered units with at least $675,000 in total assets. The SCF data show that there were 2.17 million single/widowed/divorced/separated households in 2001 with total asset holdings worth nearly $4.6 trillion, while ETD estimates show 2.6 million such individuals with more than $4.8 trillion in total assets. Financial assets compose 53 percent of total assets in the SCF, but account for nearly 65 percent of the total in the ETD estimates. Nevertheless, the mean and median values for financial assets are similar between the two groups, with SCF values somewhat lower than ETD values. Total nonfinancial assets have somewhat higher mean and median values in the SCF estimates. The mean and median values for personal residences in both datasets are remarkably similar, despite the higher incidence of this asset reported in the SCF and the fact that personal residences account for a smaller portion of total assets in the ETD estimates.

Because point estimates for married households in the SCF include assets of both partners while estimates from the ETD are for only one of a pair, direct comparisons are not meaningful. However, it would be useful to know whether differences in the estimates are primarily attributable to differences in the unit of measurement and population coverage, or if these differences are masking more fundamental structural differences between the two data sets. In order to examine these issues, it is necessary either to divide households in the SCF to create individuals, or to impute households from individuals in the ETD. There have been a couple of attempts to simulate the estate tax filing population using SCF data (see for example Poterba and Weisbrenner, 2001; Eller et al., 2001). However, these efforts have been limited by the sample size of the SCF and the sensitivity of the resulting estimates to assumptions about the relative share of household assets attributable to each separate spouse. We choose instead to impute households for married individuals in the ETD. A sketch of the procedures follows (see Johnson and Woodburn, 1994 for a full description of this process).

While estate tax returns provide detailed information on property held jointly with a surviving spouse, they provide virtually no other information on the wealth owned separately by the survivor, making model-based imputation of households infeasible. Instead, hotdeck imputation is used to approximate the wealth of a survivor spouse (see Hinkins and Scheuren, 1986, for a detailed discussion of hotdeck imputation). Married decedents are separated into two groups, based on sex, under the simplifying assumption that decedents on the file, as a group, had characteristics similar to those of the surviving spouses [10]. Adjustment cells are constructed based on the value of jointly held property, within broad age strata, and male decedents were paired randomly with a female decedent, within adjustment cells, to form families. Additional weight adjustments are needed to account for households where the female decedent's wealth is above the estate tax filing threshold, but where the separate wealth of her spouse is below the threshold. Still missing from this simulated household file are households where each partner's independent wealth is below the estate tax filing threshold, but where their combined gross assets exceed $675,000. By choosing a high enough threshold, for example $1.5 million, the effects of these missing households on final estimates should be minimized.

The resulting imputed family data set, while only crudely approximating household wealth for married individuals and ignoring nontraditional households that would be included in the SCF, can nevertheless be used to test whether the two data sources are measuring the same underlying wealth distribution. Figure 3 graphically compares the distributions of total assets using quantile-quantile (QQ) plots. If the distributions implied by the data sets being compared are similar, the plots will form a straight line. Deviation from the 45-degree line indicates variance between the two sets of estimates. The first graph compares the ETD with the SCF. Note that the QQ plot is nonlinear, meaning that the distributions are functionally different. The second graph compares the imputed family data set to estimates from the SCF and truncates the distributions at $1.5 million. In this graph, the plots for the 10th through 90th percentiles are approximately linear and much closer to the 45-degree line than was the case for the untransformed ETD estimates. The values in the SCF are still somewhat larger than ETD, as would be expected. Differences at the 99th percentile, where the ETD estimates are much higher, reflect the sample variance of both datasets, particularly the SCF, which has very few observations at this level of wealth. Overall, these results suggest that the two

Figure 3



data sets produce roughly equivalent measures despite having different units of observation.

If both the ETD and SCF are observing essentially the same population characteristics, they should provide similar estimates of economic trends. One trend that is often considered an important measure of the overall economic well-being in the U.S. is the concentration of wealth, defined here as the share of total wealth owned by a fixed portion of the population. As shown in Figure 4, the SCF estimates reveal that the wealthiest 1 percent of households owned between 30 percent and 35 percent of total household

wealth between 1989 and 2001, with an increase between 1992 and 1995 and a slight decrease after that. Estimates for individuals in the top 1 and top ½ percent of the population constructed from ETD show a similar trend, with a slight increase in the middle of the period, but with concentration in 2001 about the same as in 1989.

▶ Income Data

Both the SCF and the ITF file are important sources of data on the different types of income received by households and tax filers. The main differences between the two sources are the unit of observation, sample size, and the motivations people face in providing data. While much has been said about the differences in the unit of observation in the two data sources, it is also worth noting the difference in the sample size. The ITF file is a sample of approximately 175,000 tax records, but the sample size for the 2001 SCF is a much smaller 4,449 households. Although the SCF has a smaller sample, the detail and scope of the data allow for a broader range of research than is possible with the tax data.

Valuation Issues

The income questions in the SCF are structured to allow the respondents to reference their tax forms when answering the income questions. Figure 5 shows the correspondence between the income questions in the SCF and the line number on IRS Form 1040. The SCF variable numbers that correspond to each line of the IRS Form 1040 are listed on Figure 5. As shown in Figure 5, the SCF income questions were designed to cover most forms of income that a household reports on its tax form. Since the SCF is interested in all sources of household income and not just income subject to taxation, the questions on pensions, IRA/401(k) distributions, annuities, and Social Security payments refer to the total amounts. The SCF also asks about any income from nontaxable investments, such as municipal bonds, and any income received from Government transfer programs (such TANF, SSI, and food stamps). Households are not questioned about any adjustments to total income (lines 23-31a on Form 1040), but households are questioned about their Adjusted Gross Income (AGI, line 33). All

Figure 4: Percentage of Total U.S. Net Worth Held by Top 1 Percent and 1/2 Percent of the U.S. Population, 1989-2001

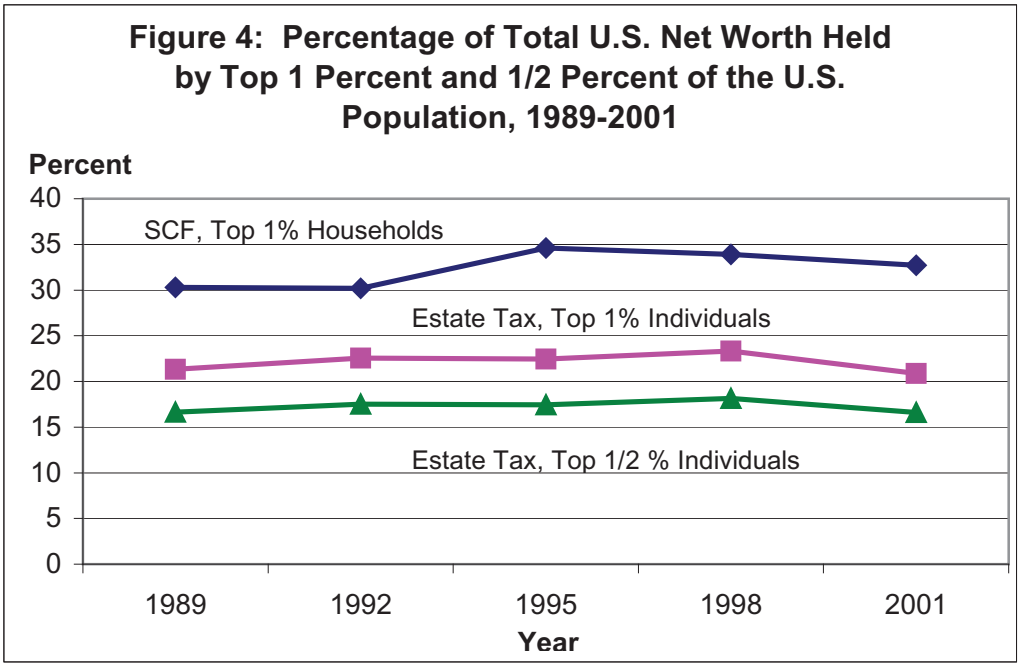


Figure 5

Form **1040** Department of the Treasury—Internal Revenue Service **U.S. Individual Income Tax Return** 2000 (99) IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2000, or other tax year beginning , 2000, ending , 20 OMB No. 1545-0074

Label (See instructions on page 19.) **Use the IRS label.** Otherwise, please print or type. LABEL HERE

Your first name and initial | Last name | **Your social security number**

If a joint return, spouse's first name and initial | Last name | **Spouse's social security number**

Home address (number and street). If you have a P.O. box, see page 19. | Apt. no.

City, town or post office, state, and ZIP code. If you have a foreign address, see page 19.

Important! You **must** enter your SSN(s) above.

Presidential Election Campaign (See page 19.) **Note.** Checking "Yes" will not change your tax or reduce your refund. Do you, or your spouse if filing a joint return, want $3 to go to this fund? ▶ You ☐ Yes ☐ No Spouse ☐ Yes ☐ No

Filing Status Check only one box.

1 ☐ Single
2 ☐ Married filing joint return (even if only one had income)
3 ☐ Married filing separate return. Enter spouse's social security no. above and full name here. ▶
4 ☐ Head of household (with qualifying person). (See page 19.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 ☐ Qualifying widow(er) with dependent child (year spouse died ▶). (See page 19.)

Exemptions

6a ☐ **Yourself.** If your parent (or someone else) can claim you as a dependent on his or her tax return, **do not** check box 6a
b ☐ **Spouse**
c **Dependents:** (1) First name Last name | (2) Dependent's social security number | (3) Dependent's relationship to you | (4) ✓ if qualifying child for child tax credit (see page 20)

If more than six dependents, see page 20.

d Total number of exemptions claimed

No. of boxes checked on 6a and 6b
No. of your children on 6c who:
• lived with you
• did not live with you due to divorce or separation (see page 20)
Dependents on 6c not entered above
Add numbers entered on lines above ▶

Income

Attach Forms W-2 and W-2G here. Also attach Form(s) 1099-R if tax was withheld.

If you did not get a W-2, see page 21.

Enclose, but do not attach, any payment. Also, please use **Form 1040-V.**

7	Wages, salaries, tips, etc. Attach Form(s) W-2		7 X5702
8a	**Taxable** interest. Attach Schedule B if required		8a X5708
b	**Tax-exempt** interest. **Do not** include on line 8a	8b X5706	
9	Ordinary dividends. Attach Schedule B if required		9 X5710
10	Taxable refunds, credits, or offsets of state and local income taxes (see page 22)		10
11	Alimony received		11 X5718
12	Business income or (loss). Attach Schedule C or C-EZ		12 X5704
13	Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ ☐		13 X5712
14	Other gains or (losses). Attach Form 4797		14 X5712
15a	Total IRA distributions	15a X5724	b Taxable amount (see page 23) 15b
16a	Total pensions and annuities	16a X5722	b Taxable amount (see page 23) 16b
17	Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E		17 X5714
18	Farm income or (loss). Attach Schedule F		18 X5704
19	Unemployment compensation		19 X5716
20a	Social security benefits	20a X5722	b Taxable amount (see page 25) 20b
21	Other income. List type and amount (see page 25)		21 X5724
22	Add the amounts in the far right column for lines 7 through 21. This is your **total income** ▶		22

Adjusted Gross Income

23	IRA deduction (see page 27)	23	
24	Student loan interest deduction (see page 27)	24	
25	Medical savings account deduction. Attach Form 8853	25	
26	Moving expenses. Attach Form 3903	26	
27	One-half of self-employment tax. Attach Schedule SE	27	
28	Self-employed health insurance deduction (see page 29)	28	
29	Self-employed SEP, SIMPLE, and qualified plans	29	
30	Penalty on early withdrawal of savings	30	
31a	Alimony paid b Recipient's SSN ▶	31a	
32	Add lines 23 through 31a		32
33	Subtract line 32 from line 22. This is your **adjusted gross income** ▶		33 X5751.X7651.X7652

For Disclosure, Privacy Act, and Paperwork Reduction Act Notice, see page 56. Cat. No. 11320B Form **1040** (2000)

Form 1040 (2000) Page **2**

Tax and Credits

34	Amount from line 33 (adjusted gross income)		34
35a	Check if: ☐ **You** were 65 or older, ☐ Blind; ☐ **Spouse** was 65 or older, ☐ Blind. Add the number of boxes checked above and enter the total here ▶ **35a**		
b	If you are married filing separately and your spouse itemizes deductions, or you were a dual-status alien, see page 31 and check here ▶ **35b** ☐		
36	Enter your **itemized deductions** from Schedule A, line 28, **or standard deduction** shown on the left. **But** see page 31 to find your standard deduction if you checked any box on line 35a or 35b **or** if someone can claim you as a dependent		36
37	Subtract line 36 from line 34		37
38	If line 34 is $96,700 or less, multiply $2,800 by the total number of exemptions claimed on line 6d. If line 34 is over $96,700, see the worksheet on page 32 for the amount to enter		38
39	**Taxable income.** Subtract line 38 from line 37. If line 38 is more than line 37, enter -0-		39
40	**Tax** (see page 32). Check if any tax is from **a** ☐ Form(s) 8814 **b** ☐ Form 4972		40
41	Alternative minimum tax. Attach Form 6251		41
42	Add lines 40 and 41 ▶		42
43	Foreign tax credit. Attach Form 1116 if required	43	
44	Credit for child and dependent care expenses. Attach Form 2441	44	
45	Credit for the elderly or the disabled. Attach Schedule R	45	
46	Education credits. Attach Form 8863	46	
47	Child tax credit (see page 36)	47	
48	Adoption credit. Attach Form 8839	48	
49	Other. Check if from **a** ☐ Form 3800 **b** ☐ Form 8396 **c** ☐ Form 8801 **d** ☐ Form (specify) _____	49	
50	Add lines 43 through 49. These are your **total credits**		50
51	Subtract line 50 from line 42. If line 50 is more than line 42, enter -0- ▶		51

Standard Deduction for Most People
Single: $4,400
Head of household: $6,450
Married filing jointly or Qualifying widow(er): $7,350
Married filing separately: $3,675

Other Taxes

52	Self-employment tax. Attach Schedule SE	52
53	Social security and Medicare tax on tip income not reported to employer. Attach Form 4137	53
54	Tax on IRAs, other retirement plans, and MSAs. Attach Form 5329 if required	54
55	Advance earned income credit payments from Form(s) W-2	55
56	Household employment taxes. Attach Schedule H	56
57	Add lines 51 through 56. This is your **total tax** ▶	57

Payments

If you have a qualifying child, attach Schedule EIC.

58	Federal income tax withheld from Forms W-2 and 1099	58	
59	2000 estimated tax payments and amount applied from 1999 return	59	
60a	**Earned income credit (EIC)**	60a	
b	Nontaxable earned income: amount ▶ and type ▶		
61	Excess social security and RRTA tax withheld (see page 50)	61	
62	Additional child tax credit. Attach Form 8812	62	
63	Amount paid with request for extension to file (see page 50)	63	
64	Other payments. Check if from **a** ☐ Form 2439 **b** ☐ Form 4136	64	
65	Add lines 58, 59, 60a, and 61 through 64. These are your **total payments** ▶		65

Refund

Have it directly deposited! See page 50 and fill in 67b, 67c, and 67d.

66	If line 65 is more than line 57, subtract line 57 from line 65. This is the amount you **overpaid**	66
67a	Amount of line 66 you want **refunded to you** ▶	67a
▶ b	Routing number ▶ **c** Type: ☐ Checking ☐ Savings	
▶ d	Account number	
68	Amount of line 66 you want **applied to your 2001 estimated tax** ▶	68

Amount You Owe

69	If line 57 is more than line 65, subtract line 65 from line 57. This is the **amount you owe**. For details on how to pay, see page 51 ▶	69
70	Estimated tax penalty. Also include on line 69	70

Sign Here

Joint return? See page 19.
Keep a copy for your records.

Under penalties of perjury, I declare that I have examined this return and accompanying schedules and statements, and to the best of my knowledge and belief, they are true, correct, and complete. Declaration of preparer (other than taxpayer) is based on all information of which preparer has any knowledge.

Your signature	Date	Your occupation	Daytime phone number
			( )
Spouse's signature. If a joint return, **both** must sign.	Date	Spouse's occupation	May the IRS discuss this return with the preparer shown below (see page 52)? ☐ **Yes** ☐ **No**

Paid Preparer's Use Only

Preparer's signature		Date	Check if self-employed ☐
Firm's name (or yours if self-employed), address, and ZIP code			EIN / Phone no. ( )

Preparer's SSN or PTIN

Form **1040** (2000)

income amounts reported in the SCF are for the year prior to the survey year.

Even with the close correspondence between the income questions in the SCF and IRS Form 1040, accurate classification and reporting of income amounts are still a potential problem in the SCF. While households are encouraged to reference documents during the interview, in the 2001 SCF, only about 32 percent of households referenced any type of documents. However, of those households that used documents, 43 percent referenced their tax forms. The ability of households that did not reference their tax forms to accurately recall and classify income introduces potential bias or inefficiency into the SCF income estimates. Although the legal penalties for misreporting income provide a strong incentive for filers to report accurate amounts to the IRS, evasion and misclassification may still bias the estimates and introduce inefficiencies.

Direct Comparisons Between SCF and SOI Data

Figure 6 provides a comparison of SCF and SOI income for the 2000 tax year. The first row of Figure 6 highlights the difference in the unit of observation between the two data sources. In the SCF, the unit of observation is the household, which can often contain more than one tax unit. The SCF asks the filing status of the core individual or couple in a household, thus allowing married or partnered households filing separately to be counted as two returns. The SCF underestimates the number of returns, no doubt in large part because the SCF does not ask about the filing status of other individuals within the household. These individuals include dependents who may also file a return and other members of the household who are not financially dependent on the household head or the core couple.

Figure 6
Comparing Components of Total Income from the SCF to the IRS Values, All Returns
(Money amounts in thousands of dollars)

Tax Year	2000		
Data Source	**SCF**	**IRS**	**% Diff**
Number of Returns	102,825,058	129,373,500	-25.8
Components of Total Income			
Wages and salary	4,985,506,700	4,456,167,438	10.6
Business income	651,515,251	213,865,353	67.2
Nontaxable interest	54,929,226	54,511,136	0.8
Taxable interest	138,970,069	199,321,670	-43.4
Dividends	107,561,912	146,987,679	-36.7
Capital gain/loss	492,696,443	630,542,431	-28.0
Rent, royalties, s-corp	180,621,157	238,022,618	-31.8
Unemployment	14,625,905	16,913,305	-15.6
Alimony	26,683,086	6,192,307	76.8
Pensions, annuities, SS	459,542,345	738,596,530	-60.7
Other income	49,438,841	25,370,158	48.7
Total	7,162,090,935	6,726,490,625	6.1
Memo item:			
Broad business income	1,324,832,851	1,082,430,402	18.3

Notes: SCF values are for households who filed or intend to file a tax return.
IRS values from Tables 1.3 and 1.4 in *Statistics of Income—2000, Individual Income Tax Returns*.
Broad business income includes business income, capital gain/loss, and rent, royalties, and S corporation income.

For the components of total income, Figure 6 shows no clear pattern in the comparison of the two data sources; the SCF overestimates five and underestimates six of the income components relative to the SOI estimates. Of the eleven income components, the SCF and SOI estimates are within +/- 30 percent for wage and salary, nontaxable interest, capital gains, and unemployment income. The differences for the seven other income components are quite large; SCF alimony income is 76 percent larger than the SOI estimate, and the amount of SCF pensions, annuities, and Social Security income is 60 percent less than the SOI estimate. The larger differences deserve further investigation.

Some of the differences in the SCF and SOI estimates are due to how each source defines an income component. For example, the SCF question on alimony income instructs the respondent to include child support payments. Since child support payments are nontaxable, such payments should not be included in the SOI estimate. One possible method for removing child support payments from SCF alimony income is to restrict the estimate of alimony income to households who report alimony income but have no children under the age of 25 in the household. This restriction reduces the amount of alimony income to $3.6 billion, which is about 58 percent of the SOI estimate ($6.2 billion).

The SCF underestimates the amount of taxable interest and dividends by 43 percent and 36 percent, respectively. A possible reason for these lower estimates is that households that receive small amounts of taxable interest or dividend income may forget to report these amounts in the SCF questionnaire. Even households with large interest income may find such income less salient if they are not in a phase of life where they would rely on such income for spending. Since the SCF collects extensive information on assets, it is possible to indirectly estimate the amount of income households might receive from their interest and dividend-producing assets. Unfortunately, the estimates of interest and dividend income obtained by applying average rates of return to these types of assets are even lower than the estimates derived from the SCF income questions. Two reasons for this difference are heterogeneity in the rates of return for different households and the sale or consumption of assets during the time prior to the survey interview.

Business income estimated by the SCF is over three times as large as the SOI estimate. However, note that the amount of capital gains and the amount of rent, royalties, and subchapter S corporation income reported in the SCF are about 30 percent lower than SOI estimates. The SCF definition of business income should be analogous to income reported on lines 12 and 18 of SOI Form 1040 (see Figure 5), but it is not unlikely that households may be misclassifying capital gains or rent, royalties, and subchapter S corporation income as business income. This may be partially due to the order of the income questions in the SCF, since the business income question is asked early in the income sequence, while the capital gains and rent, royalties, and subchapter S corporation income questions are asked later in the sequence. A broader definition of business income might include all three of these income measures; summation of the three measures reveals that the SCF estimate is about 18 percent larger than the SOI estimate.

Another large difference between the income estimates is that the SCF understates the total of pension, annuity, and Social Security incomes by 60 percent. By using information reported in other sections of the SCF, it is possible to compute alternative estimates of pension, annuity, and Social Security income. The sum of the three alternative estimates of these components is less than 2 percent larger than the estimate f total pension, annuity, and Social Security income derived from the summary income questions in the SCF. Furthermore, the SCF estimate of Social Security income is about 26 percent larger than the SOI estimate. Thus, the problem appears to be the estimate of pension and annuity income, not the estimate of Social Security income.

The estimate of "other" income, the final income component in Figure 6, is about 50 percent larger using the SCF data than the estimate using the SOI data. One possible reason for the difference is that the SCF definition of other income includes distributions from Individual Retirement Accounts (IRA) or 401(k) plans. If income from these sources is removed, the SCF estimate of other income falls by about $13.3 billion and is now only 30 percent larger than the SOI estimate.

As an attempt to shed further light on the differences between the two data sources, tax units and households are grouped by AGI class. One motivation for this grouping is that households in the SCF with at least $50,000 in AGI are twice as likely to have referenced tax forms during the interview as households with less than $50,000 in AGI (21.5 percent versus 10.3 percent). This suggests that households in the SCF with higher AGI should do a better job of reporting and classifying income. Another motivation for grouping filers or households by AGI is to determine if the differences between the two data source are driven by many small errors throughout the AGI distribution, or one specific segment of that distribution. Figure 7 presents the results of this exercise. For the less $50,000 AGI group, only the estimates of wages and salary and pension, annuity, and Social Security income are within +/- 30 percent. This stands in contrast to the $50,000 plus AGI group, in which all but five income components are within +/- 30 percent.

For the less than $50,000 AGI group, the largest differences are for taxable interest, dividends, and rent, royalties, and subchapter S corporation income. As discussed previously, the differences for taxable interest and dividend income may be due to many households neglecting to report relatively small amounts of these types of income. For example, for households with less than $50,000 in AGI that own interest-bearing assets, about 75 percent of these households do not report any interest income. Furthermore, the median amount of interest-bearing assets for the households that do not report any interest income is only $1,900 [11].

The large difference in the estimates of rent, royalties, and subchapter S corporation income for the less than $50,000 AGI group may be partly due to the treatment of losses in the SCF. Although the SCF allows households to record negative amounts for certain income questions, often households report zero instead of the actual loss. Given the tax treatment of losses, it is not surprising that losses are more likely to be reported to the IRS.

In contrast to the income estimates for all households, the amount of business income reported in the SCF for the less than $50,000 AGI group is lower than the SOI estimate. Again, for business income, it may be more useful to combine business income, capital gains, and rent, royalties, and subchapter S corporation income into one broad measure of business income. For the less than $50,000 AGI group, the SCF estimate of this broad business income measure is less than 1 percent larger than the SOI estimate.

Turning to the bottom panel of Figure 7, for households with $50,000 or more in AGI, the lack of large differences in the estimates for most of the income components is evidence that households referencing tax forms are good for the data. As for the large differences in the estimates of business income and rent, royalty, and subchapter S corporation income, using the broader definition of business income reduces this difference substantially. Under the broad business income definition, the SCF estimate is only 20 percent larger than the SOI estimate. Whether this difference is due to reporting error in the SCF or evasion in the SOI data is unclear.

The most striking result for the $50,000 or more AGI group from Figure 7 is that the SCF estimate of pension, annuity, and Social Security income is less than one-half the SOI estimate. As with the estimates for all households, the summation of the alternative SCF estimates of pension, annuity and Social Security incomes are only about 2 percent less than the SCF estimate derived directly from the income questions. Also, the SCF estimate of Social Security income is only about 17 percent less than the SOI estimate. Thus, the bulk of the difference between the SCF and SOI estimates is due to pension and annuity income. One possible reason for the discrepancy is the treatment of rollovers from one tax-deferred retirement to another tax-deferred retirement account. For example, if a household transfers the balance of one IRA account to another IRA account, the transfer is not taxable, but the transfer amount should appear on line 16a of Form 1040 (see Figure 5). Often households neglect to report these rollovers on their tax forms since there are no tax implications. However, the SOI estimate will include these rollovers, even if the household does not include them on its tax form [12]. Since households in the $50,000 or more AGI group are about twice as likely to have some sort of tax-deferred retirement account, these households may have more rollovers.

Figure 7
Comparing Components of Total Income from the SCF to the IRS Values, By AGI Class, All Returns
(Money amounts in thousands of dollars)

Tax Year	2000		
Data Source	**SCF**	**IRS**	**% Diff**
AGI < $50,000			
Number of Returns	63,504,207	77,370,713	-21.8
Components of Total Income			
Wages and salary	1,495,908,100	1,514,257,995	-1.2
Business income	71,562,974	94,459,352	-32.0
Nontaxable interest	6,367,893	7,253,787	-13.9
Taxable interest	27,735,062	60,487,940	-118.1
Dividends	17,297,297	41,826,985	-141.8
Capital gain/loss	22,558,717	37,621,491	-66.8
Rent, royalties, s-corp	17,365,370	-21,255,979	222.4
Unemployment	9,033,543	12,204,865	-35.1
Alimony	14,568,265	4,357,077	70.1
Pensions, annuities, SS	272,705,769	294,763,093	-8.1
Other income	17,835,043	7,616,376	57.3
Total	1,972,938,034	2,053,592,982	-4.1
Memo item:			
Broad business income	111,487,061	110,824,864	0.6
AGI >= $50,000			
Number of Returns	39,320,851	32,798,001	16.6
Components of Total Income			
Wages and salary	3,489,598,600	2,941,909,441	15.7
Business income	579,952,277	119,406,001	79.4
Nontaxable interest	48,561,333	47,257,350	2.7
Taxable interest	111,235,007	138,833,728	-24.8
Dividends	90,264,615	105,160,694	-16.5
Capital gain/loss	470,137,727	592,920,941	-26.1
Rent, royalties, s-corp	163,255,787	262,335,219	-60.7
Unemployment	5,592,363	4,708,441	15.8
Alimony	12,114,821	1,821,107	85.0
Pensions, annuities, SS	186,836,576	443,833,436	-137.6
Other income	31,603,798	17,753,782	43.8
Total	5,189,152,905	4,675,940,140	9.9
Memo item:			
Broad business income	1,213,345,791	974,662,161	19.7

Notes: SCF values are for households who filed or intend to file a tax return.
IRS values from Tables 1.3 and 1.4 in *Statistics of Income—2000, Individual Income Tax Returns*.
Broad business income includes business income, capital gain/loss, and rent, royalties, and S corporation income.

A final item to note from Figure 7 is that the SCF and SOI estimates of total income for each AGI group are remarkably close. This provides evidence that, although households may misclassify the components of income, the aggregate level of income is fairly consistent.

▶ Conclusions

Our research has shown that, while ETD and SCF data seem to be capturing very similar portfolio data for the wealthiest people in the U.S, differences in population coverage and the unit of observation make it very difficult to declare estimates from one source superior to the other. There is a great deal of evidence that the financial characteristics of the very wealthy are sufficiently heterogeneous to require quite large samples to make meaningful estimates for small subpopulations. It is also clear that the increasingly complicated financial and business arrangements practiced by the very wealthy require a great deal of attention to the definition of data variables when attempting any sort of analysis. Here, we are thinking about the proliferation of nontraditional investment instruments, such as derivatives, strips, options, and futures, as well as complex ownership arrangements, such as trusts, family limited partnerships, and holding companies. Lifecycle effects are also an important consideration; the portfolios of working individuals are different from those of the retired, which are also going to be different from individuals who face the end of their lives.

For studying broad trends in the population or for an overview of the top of the wealth distribution, the SCF provides more complete coverage than ETD. By focusing on households, the SCF data are uniquely suited for answering many complex economic questions and provide comparability with other publicly available national datasets. The availability of extensive savings, income, debt, work history, and demographic data also makes the SCF a much richer source of data than ETD for many research purposes. In addition, the sample design ensures that individuals at all phases of the lifecycle are included in the sample, thus providing a broad measure of the economic behavior of all households.

Data from U.S. estate tax returns provide a unique source of data on wealthy individuals. For many purposes, such as the study of intergenerational wealth transfers, they are the only viable data source. The large sample size permits detailed study of individuals at the highest levels of the wealth distribution. ETD can also support detailed study of the wealthy in various demographic groupings, particularly by age, marital status, and sex, while these groups are not sufficiently represented in the SCF to allow reliable estimates. These demographic characteristics seem to be key determinants of behaviors such as portfolio choice, charitable giving, and bequest decisions. In addition, the abundance of valuation documentation provided with ETD provides unique opportunities to study in detail the financial planning and business arrangements employed by the wealthy to both minimize tax liability and to ensure that a legacy of wealth accumulation is preserved beyond their lifetimes.

Estimates for households made up of single, widowed, divorced, or separated individuals in the ETD and SCF were remarkably similar, and our simulations suggest that data for married or partnered households are likewise comparable. Overall, values reported on estate tax returns appear to be conservative relative to those in the SCF, reflecting the difficulty of valuing some assets, especially businesses; practical considerations, such as the difficulty of finding a willing buyer for a fractional interest in a basket of market goods; and the natural desire to minimize tax liability to the great extent possible within the constraints of the tax code. In addition, differences between the mean and median ages reported in the ETD and those in the SCF suggest that the use of mortality rates that reflect the longevity advantages enjoyed by the wealthy in constructing wealth multipliers may not completely compensate for overrepresentation of the elderly in the decedent population, perhaps introducing a slight bias. The ETD may also be biased by effective financial and estate planning, by expenses associated with a long final illness, and by changes in asset holdings made in anticipation of death.

In terms of the comparison between the SCF and SOI income data, our research has shown that, although there are differences in the unit of observation and issues with the definition of certain income types, the two data sources compare quite favorably. One reason for this is the close correspondence between the SCF income

questions and the income categories on IRS Form 1040. While it appears that households often misclassify income, the total amount of income reported by households in the SCF is only 6 percent larger than the SOI estimate. Due to the detail and scope of the SCF data, it is often possible to use data from other sections of the survey to make adjustments to better align the SCF and SOI income definitions. The detail and scope of the other data collected in the SCF also allow for a broader range of research than the SOI tax data. However, the large sample size and administrative nature of SOI tax data make it an appealing source for certain types of research, such a tax policy.

The direct comparison of the SCF and SOI income data reveals that encouraging households to reference their tax forms is critical for the accuracy of the SCF income data. Households with lower AGI may feel it is unnecessary to check their tax forms given the few types of income they receive, but it clearly makes a difference, as Figure 7 demonstrates. Households with higher levels of AGI are more likely to receive more types of income due to the increasing complexity of their financial situations. Thus, it is potentially even more difficult for these households to correctly report and classify their incomes without referencing their tax forms.

Overall, the message for researchers is that the SCF and SOI data are complementary sources of data on both wealth and income. The goal of our research is not to declare one data set superior to the other; that is a difficult judgment to render. What we have attempted to show in this paper is that there are many important issues to understand when comparing administrative and survey data. The key, then, is that each data source has strengths and weaknesses that need to be understood and carefully considered before attempting to use them to answer any set of research questions.

▶ References

Aizcorbe, Ana; Kennickell, Arthur B.; and Moore, Kevin B. (2003), "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances," *Federal Reserve Bulletin*, Volume 89, pp. 1-32.

American Statistical Assocation (1977), "Report of the Ad Hoc Committee on Privacy and Confidentiality," *The American Statistician*, Volume 31, pp. 59-78.

Atkinson, A.B. and Harrison, A.J. (1978), *Distribution of Personal Wealth In Britain*, Cambridge University Press, Cambridge, England.

Brackstone, C.J (1987). "Statistical uses of Administrative Data: Issues and Challenges," *Statistical Uses of Administrative Data Proceedings*, pp. 5-26.

Brown, Robert E and Mazur, Mark J. (June 2003), "IRS' Comprehensive Approach to Compliance Measurement," 2003 National Tax Association Spring Symposium, http://www.irs.gov/pub/irs-soi/mazur.pdf.

Cartwright, David W. and Armknecht, Paul A. (1979), "Statistical Uses of Administrative Records," *Proceedings, Section on Survey Research Methods*, American Statistical Association, pp 73-76.

Eller, Martha Britton (2001), "Audit Revaluation of Federal Estate Tax Returns," *Internal Revenue Service Statistics of Income Bulletin*, Winter 2000-2001, Washington, D.C., pp. 100-139.

Eller, Martha Britton; Erard, Brian; and Ho, Chih-Chin Ho (2001), "Noncompliance with the Federal Estate Tax," in *Rethinking Estate and Gift Taxation*, William G. Gale, James R. Hines, and Joel Slemrod, editors, Brookings Institution Press, pp. 375-421.

Frankel, Martin and Kennickell, Arthur B. (1995), "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fries, Gehard (2003), "Disclosure Review and the 2001 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Garnick, Daniel H. and Gonzalez, Maria Elana (1979), "Statistical Uses of Administrative Records: Where Do We Go From Here?," *Proceedings, Section on Survey Research Methods*, American Statistical Association .

Hinkins, Susan and Scheuren, Frederick (1986), "Hot Deck Imputation Procedure Applied to a Double Sample Design," *Survey Methodology*, Volume 12, pp. 181-196.

Jensen, Paul (1987), "The Quality of Administrative Data From a Statistical Point of View, Some Danish Experience and Considerations, *Statistical Uses of Administrative Data Proceedings*, pp. 291-300.

Johnson, Barry W. and Mikow, Jacob M. (Spring 2002), "Federal Estate Tax Returns, 1998-2000," *Statistics of Income Bulletin*, Internal Revenue Service, Washington DC, pp. 113-186.

Johnson, Barry W. and Woodburn, Louise (1994), "The Estate Multiplier Technique, Recent Improvements for 1989," in *Compendium of Federal Estate Tax and Personal Wealth Studies*, Barry Johnson, editor, Internal Revenue Service, Publication 1773, pp. 391-400.

Kennickell, Arthur B. (2003), "A Rolling Tide: Changes in the Distribution of Wealth in the U.S., 1989-2001," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (2001), "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (1999), "Revisions to the SCF Weighting Methodology: Accounting for Race/Ethnicity and Homeownership," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (1998a), "List Sample Design for the 1998 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Kennickell, A. (1998b), "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association.

Kennickell, A. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kennickell, Arthur B. and McManus, Douglas (1993), "Sampling for Household Financial Characteristics Using Frame Information on Past Income," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. and Woodburn, R. Louise (1999), "Consistent Weight Design for the 1989, 1992, and 1995 SCF's, and the Distribution of Wealth," *Review of Income and Wealth*, Series 45, 2, pp. 193-215.

Lessler, Judith T. and Kalsbeek, William D. (1992), *Nonsampling Error in Surveys*, John Wiley and Sons, New York.

Menchik, Paul (1991), "Economic Status as a Determinant of Mortality Among Nonwhite and White Older Males: or, Does Poverty Kill?," Institute for Research on Poverty, Discussion Paper Number 93891.

Petska, Tom and Strudler, Mike (1999), "The Distribution of Individual Income Taxes: A New Look at an Old Issue," *Turning Administrative Systems Into Information Systems, Internal Revenue Service*, pp. 7-22.

Poterba, James and Weisbenner, Scott (2001), "The Distributional Burden of Taxing Estates and Unrealized Capital Gains at Death," in *Rethinking Estate and Gift Taxation*, William G. Gale, James R. Hines, and Joel Slemrod, editors, Brookings Institution Press, pp. 422-456.

Sailer, Peter and Weber, Michael (1996), "Creating Household Data from Individual Income Tax Returns," *Proceedings, Section on Survey Research Methods*, American Statistical Association .

Sailer, Peter and Weber, Michael (1999), "The IRS Population Count: An Update," *Proceedings, Section on Survey Research Methods*, American Statistical Association.

Scheuren, Fritz (1994), "Historical Perspectives on IRS Wealth Estimates With a View to Improvements," *Compendium of Federal Estate Tax Data and Personal Wealth Studies*, Department of the Treasury, IRS Publication 1773, pp. 355-362.

Statistical Policy Working Paper 6--Report on Statistical Uses of Administrative Records (1980), Federal Committee on Statistical Methodology, United States Office of Management and Budget.

Tourangeau, Roger; Johnson, Robert A.; Qian, Jiahe; Shin, Hee-Choon; and Frankel, Martin R. (1993), "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago.

Wilk, M. B. and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, Volume 55, pp. 1-17. (as described in Hoaglin, David C. (1985), "Using Quantiles to Study Shape," *Exploring Data Tables, Trends, and Shapes*, John Wiley and Sons, New York.

Wilson, Robert (Fall 1988), "Statistics of Income: A By-Product of the U.S. Tax System," *Statistics of Income Bulletin*, Department of the Treasury, Internal Revenue Service, Volume 8, Number 2, pp. 103-114.

▶ Footnotes

[1] In nine U.S. States, nearly all property acquired by a married couple is considered owned equally by both parties. Property acquired separately by gift or bequest is generally exempted.

[2] In States where there are no community property rights, assets titled legally as joint tenants are considered owned equally by both partners in a marriage, usually without regard to how much consideration each party contributed to purchase the asset.

[3] Details of the SCF list sample design are provided in Kennickell, 2001.

[4] Due to the difficulty of gaining cooperation from the wealthiest individuals, the SCF uses as its upper sample threshold the minimum amount of wealth required for inclusion in the listing of the wealthiest 400 individuals in the U.S., as estimated by *Forbes* magazine. Kennickell (2001) discusses the methodology used for selecting the SCF list sample.

[5] Gross estate is a measure similar to total assets, but which includes the full face value of life insurance, certain gifts made prior to death, and certain assets placed in trust.

[6] Where possible, we modify the data to compensate for these reporting anomalies. For example, the full face value of life insurance is included in the decedent's total gross estate for tax purposes; however we impute a cash value using data from the SCF.

[7] Examination rates vary by size of estate. In 2003, about 6.4 percent of all returns were examined, while 27.5 percent of those reporting estates of $5 million or more were subject to examination. A recent Statistics of Income (SOI) study, based on the results of IRS audits of estate tax returns filed in 1992, estimated that detected undervaluation of assets was about 1.2 percent of total asset holdings for all audited returns (Eller, et al., 2001).

[8] A family limited partnership is a business arrangement in which a wide array of business and market assets are transferred to a partnership, with general partner interests held by parents

and limited partner shares distributed to children through annual tax-exempt gifts. This results in fractured ownership interests in the individual assets, qualifying them for large valuation discounts for tax purposes.

[9] The mean and median ages for heads of households with total assets of $1,500,000 from the SCF were both 57, virtually the same as for individuals in the ETD with this level of wealth, for whom the mean and median ages were 58.

[10] This approach will tend to overpredict wealth since some surviving spouses would in reality have less wealth than those available for matching in the ETD.

[11] For households with $50,000 or more in AGI that own interest-bearing assets, about 53 percent do not report any interest income. Median interest-bearing assets for these nonreporting households is $6,200.

[12] A rollover transaction generates a Form 1099-R that SOI matches to Form 1040. If a filer neglects to report the rollover on his or her tax form, the value from Form 1099-R is added to the filer's Form 1040.

The Effect of Late-Filed Returns on Population Estimates: A Comparative Analysis

by Brian Raub, Cynthia Belmonte, Paul Arnsberger, and Melissa Ludlum, Internal Revenue Service

The Statistics of Income (SOI) division of the Internal Revenue Service (IRS) collects and disseminates detailed data based on samples of administrative records, including tax and information returns. Estimates for populations of interest for SOI studies are produced by drawing stratified, random Bernoulli samples of tax and information returns as they are filed, over periods that span a predetermined timeframe. While this methodology results in the inclusion of the majority of targeted returns, a small number of returns for each study are received beyond the data collection period. These "late-filed" returns may introduce non-response bias into the population estimates, which might be mitigated by post-stratification or weighting adjustments. (The term "late-filed return" as used in this paper does not address the compliance, or lack thereof, of return filings with statutory requirements.) Using three SOI studies with varying sampling frames, this paper will function as a case study on the effects of truncated sampling periods on population estimates.

The data presented in this paper are derived from two sources—sample data produced by SOI and administrative data obtained from the IRS Masterfile for the population of returns filed. SOI sample data typically include detailed, error-perfected financial and other information about the tax filing entity. SOI sample data are used to produce population estimates that are used in statistical studies and for analysis of tax policy. Data obtained from the IRS Masterfile include limited information for the population of filers. This information is generally used for a variety of purposes related to tax administration.

SOI conducts annual studies of a wide range of filers, including individuals, corporations, partnerships, estates, trusts, tax-exempt charitable organizations, and many other filers. This paper focuses on three SOI studies—the Estate Tax study, the Private Foundation study, and the Exempt Organization study.

▶ The Estate Tax Study

With its annual Estate Tax study, SOI extracts demographic, financial, and asset data from Federal estate tax returns. The annual study allows production of a data file for each filing, or calendar, year. By focusing on a single year of death for a period of 3 filing years, the study allows production of periodic year-of-death estimates. A single year of death is examined for 3 years, as over 98 percent of all returns for decedents who die in a given year are filed by the end of the second calendar year following the year of death. Data included in this paper are for Year of Death 2004 and were obtained from returns filed in Calendar Years 2004-2006.

The estate of a decedent who, at death, owns assets valued in excess of the estate tax applicable exclusion amount, or filing threshold, must file a Federal estate tax return, Form 706, *U.S. Estate (and Generation-Skipping Transfer) Tax Return*. For decedents who died in 2004, the exclusion amount was $1.5 million. Alternate valuation may be elected only if the value of the estate, as well as the estate tax, is reduced between the date of death and the alternate date. The estate tax return is due 9 months from the date of the decedent's death, although a 6-month filing extension is allowed. In some cases, longer filing extensions may be permitted.

For the Year of Death 2004 Estate Tax study, there were 11,817 Form 706 returns in the sample selected from a population of 42,424. The SOI Estate Tax study is classified into strata based on year of death, the size of total gross estate, and age of the decedent. For the Year of Death 2004 study, there were a total of 57 sam-

This paper was originally presented at the 2009 annual meetings of the American Statistical Association held in Washington, D.C., on August 2-6, 2009.

pling strata, with sampling rates ranging from 4 percent to 100 percent.

▶ The Private Foundation and Exempt Organization Studies

The annual SOI studies of private foundations and exempt organizations collect detailed financial data, as well as information on charitable and grant-making activities and compliance with IRS regulations from information returns filed by exempt organizations. Studies are conducted for a single tax year and include samples of returns filed and processed during the 2 calendar years immediately following the target tax year. Data discussed in this paper for the Private Foundation and Exempt Organization studies were obtained for Tax Year 2004 returns filed in Calendar Years 2005 and 2006. The Tax Year 2004 samples include organizations with accounting periods beginning in Calendar Year 2004 (and ending between December 2004 and November 2005), for which returns were filed and processed to the IRS Business Masterfile during Calendar Years 2005 and 2006. While this 2-year sampling period ensures almost complete coverage of the target population, there are still a number of returns processed after the close of the second year (i.e., December 31, 2006 for the Tax Year 2004 study), which are generally excluded from the samples.

Private foundations and nonexempt charitable trusts are required to file Form 990-PF (*Return of Private Foundation or Section 4947(a)(1) Nonexempt Charitable Trust Treated as Private Foundation)* annually. Similarly, certain exempt organizations are required to file Forms 990 *(Return of Organization Exempt from Income Tax)* or Form 990-EZ *(Short Form Return of Organization Exempt from Income Tax)*. SOI conducts annual studies based on samples of Forms 990-PF, 990, and 990-EZ filed for a given tax year. These information returns are due 5 months after the close of the organization's accounting period, although a 3-month filing extension is allowed. In some cases, additional filing extensions may be granted.

For the Tax Year 2004 Private Foundation study, there were 7,805 Form 990-PF returns in the sample, selected from a population of 80,570. The SOI Private Foundation study is classified into strata based on the size of end-of-year fair market value of assets, with each stratum sampled at a different rate. Sampling rates ranged from 1 percent for private foundations with total assets less than $125,000 to 100 percent for private foundations with total assets of $10 million or more.

The Tax Year 2004 exempt organization sample of section 501(c)(3) filers comprised 15,070 Forms 990 and 990-EZ, selected from a population of 279,415. End-of-year book value of assets was the stratifying variable for the exempt organization study. Sampling rates ranged from 1 percent for exempt organizations with total assets less than $500,000, to 100 percent for those with total assets of $50 million or more.

▶ Late-Filed Returns

To examine the effect of late-filed returns on each of the studies, an augmented sampling frame, which incorporates 2 years of additional return filings, was constructed from IRS Masterfile data. The following tables show the number of late-filed returns received within the current and augmented sampling frames, as well as the percentage of selected financial variables represented by returns received inside and outside of the sampling period.

Table 1, below, shows the percentage of Year of Death 2004 Forms 706 filed, total gross estate, and net estate tax reported for returns filed over a 5-year collection period (2004–2008), by size of gross estate and by age of the decedent More than 98 percent of all Year of Death 2004. Forms 706 filed over the 5-year period were received within the 3 years, 2004 through 2006, from which returns were sampled. However, the estates of younger decedents filed returns outside of the 3-year sampling frame proportionately more often than the estates of their older counterparts. For example, nearly 4 percent of returns filed for decedents under 40 were received in 2007 and 2008. The percentage of total gross estate represented by late-filed returns was 1.1 percent, with the corresponding figure for net estate tax only 0.5 percent. These smaller percentages are attributable to the fact that late-filed returns were smaller on average

than other returns and were proportionately more often nontaxable, as shown in the following tables.

Table 1: Estate Tax Returns Filed for 2004 Decedents, IRS Masterfile Data by Age of Decedent, 2004–2008

Calendar Year	*Returns*	*Total gross estate*	*Net estate tax*
2004-2006	**98.4%**	**98.9%**	**99.5%**
Under 40	96.4%	97.0%	100.0%
40 under 50	97.0%	97.5%	98.9%
50 under 65	97.2%	97.6%	98.7%
65 and over	98.6%	99.0%	99.6%
2007-2008	**1.6%**	**1.1%**	**0.5%**
Under 40	3.6%	3.0%	0.0%
40 under 50	3.0%	2.5%	1.1%
50 under 65	2.8%	2.4%	1.3%
65 and over	1.4%	1.0%	0.4%

Table 2 examines the same population as the previous table, classified by size of total gross estate. The table shows that returns for the smallest estates, those with between $1.5 and $2 million in gross estate, were filed in the 2 years immediately following the sampling period twice as frequently as were returns for the largest estates.

Table 2: Estate Tax Returns Filed for 2004 Decedents, IRS Masterfile Data by Size of Total Gross Estate, 2004–2008

Calendar Year	*Returns*	*Total gross estate*	*Net estate tax*
2004-2006	**98.4%**	**98.9%**	**99.5%**
$1.5 million<$2.0 million	98.2%	98.2%	98.9%
$2.0 million<$3.0 million	98.3%	98.3%	99.1%
$3.0 million<$5.0 million	98.4%	98.4%	99.1%
$5.0 million<$10.0 million	98.8%	98.8%	99.5%
$10 million and over	99.1%	99.6%	99.7%
2007-2008	**1.6%**	**1.1%**	**0.5%**
$1.5 million<$2.0 million	1.8%	1.8%	1.1%
$2.0 million<$3.0 million	1.7%	1.7%	0.9%
$3.0 million<$5.0 million	1.6%	1.6%	0.9%
$5.0 million<$10.0 million	1.2%	1.2%	0.5%
$10 million and over	0.9%	0.4%	0.3%

Table 3 examines the same population as Tables 1 and 2, classified by tax status of the return. It shows that nontaxable returns were filed outside of the sampling period more than twice as often as taxable returns.

Table 3: Estate Tax Returns Filed for 2004 Decedents, IRS Masterfile Data by Tax Status, 2004–2008

Calendar Year	*Returns*	*Total gross estate*	*Net estate tax*
2004-2006	**98.4%**	**98.9%**	**99.5%**
Taxable	99.1%	99.4%	99.5%
Nontaxable	97.9%	98.2%	N/A
2007-2008	**1.6%**	**1.1%**	**0.5%**
Taxable	0.9%	0.6%	0.5%
Nontaxable	2.1%	1.8%	N/A

Table 4 illustrates the extent to which estimates based on Form 990-PF data collected from the current 2-year sampling period might be enhanced by using the augmented sampling frame. More than 98 percent of the Tax Year 2004 private foundation returns included in the augmented sampling frame were processed in the 2 years immediately following the close of the tax year. A closer examination reveals that the percentage of returns received and processed during the first 2 years increases with asset size. For example, 97.9 percent of returns filed by small organizations (those with assets less than $1,000,000) were processed during the 2005-2006 period, compared to 99.2 percent of the returns of medium-sized foundations (those with assets between $1 million and $50 million), and 99.7 percent of the returns of the largest foundations (those with assets of $50 million or more).

Table 4: Tax Year 2004 Private Foundation Information Returns, IRS Masterfile Data by Calendar Year and Size of Organization, 2005–2008

Calendar Year	*Returns*	*Assets*	*Revenue*	*Charitable Disbursements*	*Excise Tax on Net Investment Income*
2005-2006	**98.3%**	**99.5%**	**99.4%**	**99.5%**	**99.5%**
Small	97.9%	98.8%	98.7%	98.9%	98.8%
Medium	99.2%	99.3%	99.3%	99.4%	99.5%
Large	99.7%	99.6%	99.6%	99.7%	99.6%
2007-2008	**1.7%**	**0.5%**	**0.6%**	**0.5%**	**0.5%**
Small	2.1%	1.2%	1.3%	1.1%	1.2%
Medium	0.8%	0.7%	0.7%	0.6%	0.5%
Large	0.3%	0.4%	0.4%	0.3%	0.4%

Table 5 shows the breakdown of data from Forms 990 and 990-EZ returns by filing period and size of assets. As with private foundations, the vast majority of

Tax Year 2004 returns were filed in the first two years after the end the tax year. Again, a large portion of the returns filed in the final 2 years of the augmented sampling frame are from small organizations – those with total assets less than $100,000. Consequently, late filers of Forms 990 add little to the aggregate totals for most of the financial variables collected: less than 1 percent of total assets, revenue, and net worth.

Table 5: Tax Year 2004 Exempt Organization Information Returns, IRS Masterfile Data by Processing Year and Size of Organization, 2005–2008

Calendar Year	*Returns*	*Assets*	*Revenue*	*Net Worth*
2005-2006	**97.3%**	**99.3%**	**99.3%**	**99.2%**
Small	95.7%	96.4%	96.2%	96.1%
Medium	98.3%	98.8%	98.8%	97.8%
Large	99.2%	99.3%	99.3%	99.3%
2007-2008	**2.7%**	**0.7%**	**0.7%**	**0.8%**
Small	4.3%	3.6%	3.8%	3.9%
Medium	1.7%	1.2%	1.2%	2.2%
Large	0.8%	0.7%	0.7%	0.7%

▶ Current Treatment of Late Filers

Although the Estate Tax, Private Foundation, and Exempt Organization studies share a common challenge in addressing the effect of late-filed returns on population estimates, each of the three studies currently uses a different approach in dealing with this challenge.

Year of Death population estimates for the Estate Tax study include weight adjustments for late-filed returns. Weight adjustment factors are calculated based on past late filing patterns using historical data from the IRS Masterfile and are updated periodically. The aim of using these weight adjustments is to improve the overall population estimates as well as estimates for the subpopulations of returns that have historically filed late with greater frequency. As shown in Table 6, weight adjustment factors varied by size of estate, tax status of return, and age of decedent. For each size of estate and age combination, non-taxable returns received a higher adjustment factor than taxable returns. Estates with $10 million or more in gross estate received weight adjustment factors based on tax status regardless of age, as illustrated in the top portion of the table. For estates with less than $10 million in gross estate, weight adjustment factors were assigned based on tax status and age.

Table 6: Weight Adjustment Factors for Year of Death 2004 Estate Tax Population Estimates

Total gross estate ≥ $10 million		
	Taxable	*Nontaxable*
All ages	1.004	1.013
Total gross estate < $10 million		
Age	*Taxable*	*Nontaxable*
Under 40	1.036	1.052
40 under 50	1.019	1.035
50 under 65	1.018	1.028
65 and older	1.009	1.020

Table 7—shows the aggregate effect of weight adjustment factors on the Year of Death 2004 estate tax estimates. The number of returns increased about 1.5 percent compared to a 1.2 percent increase in total gross estate and less than a 1 percent increase in net estate tax. The differences in the impact of weight adjustments on these three variables is consistent with the fact that late-filed returns comprised proportionately more small returns and non-taxable returns than the population as a whole.

Table 7: Effect of Weight Adjustments on Estimates of Year of Death 2004 Estate Tax Population

[Money amounts are in millions of dollars]

	Returns	*Total Gross Estate*	*Net Estate Tax*
Unadjusted estimate	41,599	183,657	22,075
Estimate with weight adjustment	42,239	185,921	22,220
Percentage increase	1.54	1.23	0.66

In contrast, population estimates for the Private Foundation study do not include standard adjustment factors to account for returns filed after the close of the two-year sampling period. Instead, during file closeout, efforts are made to identify and include late-filed returns of private foundations that would have been sam-

pled at the 100-percent rate (i.e., organizations with fair market value of assets of $10 million or more). These include returns of organizations sampled in previous study years, as well as returns of organizations posting to the IRS Masterfile for the first time. Potentially, this can extend the 2-year sampling frame by four to five months, the typical length of time between the end of the sampling period and the creation of the final study file. Table 8, shows population estimates for selected variables from SOI's Tax Year 2004 Private Foundation study. The table includes population estimates from returns processed during the regular 2-year sampling period, as well as enhanced population estimates including adjustments for late-filed returns. Only 11 large-case, late-filed returns were added to the Tax Year 2004 sample. These returns represented 100th of 1 percent of the population estimate, and about a one-fifth of 1 percent addition to total revenue, charitable disbursements, and net investment income excise tax.

Table 8: Tax Year 2004 Private Foundation Data from SOI Estimates, Including Added Late-Filed Returns

[Money amounts are in millions of dollars. Detail may not add to totals because of rounding.]

Calendar Year	*Returns*	*Assets*	*Revenue*	*Charitable Disbursements*	*Excise Tax on Net Investment Income*
SOI two-year estimate	76,886	509,471	58,539	32,071	467
Additional data from late-filed returns	11	453	129	54	1
Enhanced SOI estimate	76,897	509,924	58,668	32,125	469
Additional data as percentage of total	0.01	0.09	0.22	0.17	0.21

The Exempt Organization study includes no weight adjustments and no attempt is made to add returns to the sample that are filed outside of the two-year sampling frame. Adjustments to the sample are made for certain organizations that file returns within the 2-year sampling period. Examples of these adjustments include rejecting short-year returns and those that file with an incorrect subsection code; and adding returns that have posted incorrectly to the Masterfile as duplicate, below the filing threshold, or with incorrect total assets.

Using IRS Masterfile data as a proxy, we can mimic the Private Foundation study's technique of processing returns from the certainty strata that are filed within five months after the close of the normal sampling period. Based on the Masterfile data, 21 large-case returns would have been added to the Tax Year 2004 sample. These returns would have accounted for a one-third of 1 percent addition to the aggregate totals for assets, revenue and net worth.

▶ Strengths/Weaknesses

These analyses reveal a number of strengths and weaknesses for each of the three approaches to the late-filer problem. The weight adjustment approach, as employed for the Estate Tax study, potentially improves the overall population estimates. It also may improve estimates for subpopulations for which returns have historically been filed late with the greatest frequency. The adjustments seem to be an effective means of counteracting any bias that may result from the existing sampling period. To the extent that late filers create bias in the Estate Tax study estimates, the weight adjustment approach may mitigate the bias.

On the other hand, the weight adjustment approach may not always be an effective method of predicting filing habits. The weight adjustments are developed from observed trends in historical data; this information may not always reliably predict future filing patterns. Although the characteristics of late filers have been relatively stable over time, significant changes to the estate tax law could alter these patterns.

The inclusion of large, late-filed returns in the Private Foundation study provides for more complete coverage of the target population by including returns that would have been selected with certainty within the defined sampling period. Additionally, this approach ensures that files are suitable for time-series analysis of a specific organization or panel of organizations. This strength may be unique to data for tax-exempt organizations, whose information returns, in most cases, are not subject to the same disclosure and confidentiality rules as data obtained from tax returns filed by other types of organizations and individuals.

The primary weakness of including large, late-filed returns only in the enhanced estimate is the inconsistency that it introduces. Slight variances in tax return processing, sample file creation or review, or sample file delivery date can affect the sampling period from which the enhanced estimate is drawn from year to year. Further, the method fails to address late-filed returns of smaller organizations, which account for the largest share of the late-filing population.

The "no-adjustment" approach that is used in the Exempt Organization study ensures a consistent sampling frame with a well-defined sampling period. This approach employs the Bernoulli sample over a 2-year period and does not include arbitrary additions or discontinuations. Because the population is framed as the estimate of filers of the 2-year period and not as the "universe" of filers, the bias does not exist.

Because, unlike the weight adjustment method used in the Estate Tax study, the "no-adjustment" approach does not attempt to account for late filers, it could consistently underestimate the number of returns filed by smaller organizations. By omitting some large case returns that are received outside of the defined sampling period, this approach also provides a somewhat less complete dataset for time-series panel analysis than does the Private Foundation study.

▶ Conclusions/Future Research

Late-filed returns present a common challenge for studies of data obtained from tax returns, such as the Estate Tax, Private Foundation, and Exempt Organization studies. Although, for each of the studies, the number of late-filed returns is modest in comparison to the number of returns filed within the defined sampling period, the absence of these returns may introduce bias into the population estimates.

Currently, each of the three studies discussed in this paper uses a unique approach to mitigate the potential bias introduced by late filers. The weight adjustment method, employed for the Estate Tax study, improves some aspects of the study's estimates, but could become distorted if filing patterns observed in historical data do not continue into the future. The enhanced Private Foundation estimate, which is obtained by including targeted returns received after the end of the sampling period, benefits time-series analysis. However, it creates inconsistencies in the year-to-year sampling period. The "no adjustment" method used for the Exempt Organization study provides a distinct sampling period, but does not address the exclusion of relatively small filers from the estimates.

The unique characteristics of late filers in each of the three studies discussed in this paper, as well as the benefits and shortfalls of using each of the three approaches to address the later-filer problem, provide a number of opportunities for further research.
This analysis will be expanded to research additional tax years and years of death in order to explore historical filing patterns. This effort will attempt to isolate an optimal sampling period that balances population coverage with timeliness of completion of the estimates. Additionally, weighting adjustments, similar to those in use for the Estate Tax study, will be developed for the Private Foundation and Exempt Organization studies. The adjustments will be examined for accuracy, as well as their effect on organization-level data from year-to-year.

Differences in Income Estimates Derived from Survey and Tax Data

by Barry W. Johnson, Internal Revenue Service and Kevin Moore, Board of Govenors, Federal Reserve System

The Statistics of Income Division of the United States Internal Revenue Service collects statistical data from samples of most major federal tax and information returns. Among these are annual studies of *U.S. Individual Income Tax Returns* (Form 1040). These data are used by both the U.S. Congress and the Executive Branch of the Government to evaluate and develop tax and economic policy, and by other government agencies and the general public for a variety of different purposes.

Form 1040 is filed annually by individuals or married couples to report income, including wages, interest, dividends, capital gains, and some types of business income. Also reported are data on deductions, expenses, and tax credits. The SOI sample of these returns is stratified based on: (1) the larger in absolute value of positive income or negative income; (2) the size of business and farm receipts; (3) the presence or absence of specific forms or schedules; and (4) the usefulness of returns for tax policy modeling purposes (see Internal Revenue Service, 2005).

The Survey of Consumer Finances is a survey of household balance sheets conducted by the Board of Governors of the Federal Reserve System in cooperation with the SOI. Beginning with 1983, the survey has been conducted triennially, with data collected by the Survey Research Center at the University of Michigan in 1983, 1986, and 1989, and by NORC, a national organization for social science and survey research at the University of Chicago, from 1992 forward. In addition to collecting information on assets and liabilities, the SCF collects information on household demographics, income, relationships with financial institutions, attitudes toward risk and credit, current and past employment, and pensions (for more details on the SCF, see Bucks, Kennickell, and Moore, 2006).

The SCF uses a dual-frame sample design to provide adequate representation of the financial behavior of all households in the United States. One part of the sample is a standard multistage national area probability sample (Tourangeau et al., 1993), while the list sample uses the SOI individual income tax data file to oversample wealthy households (Kennickell, 2001). This dual-frame design provides the SCF with efficient representation of both assets widely held in the population, such as cars or houses, and assets more narrowly held by wealthy families, such as private businesses and bonds. Wealth data from the SCF are widely regarded as the most comprehensive survey data available for the United States.

Sample weights constructed for the SCF allow aggregation of estimates to the U.S. household population level in a given survey year (Kennickell and Woodburn, 1999; Kennickell, 1999). Missing values in the 1989-2004 SCF were imputed using a multiple imputation technique (Kennickell, 1991, 1998b).

▶ Income Data

Both the SCF and the SOI file are important sources of data on the different types of income received by households and tax filers. There are a number of differences between the two sources, including the population covered, unit of observation, available data, and the motivations people face in providing data. It is also worth noting the difference in the sample size. The 2004 SOI file is a sample of approximately 200,000 tax records out of a population of about 130 million, while the sample size for the 2004 SCF is much smaller, about 4,500 households. Although the SCF has a smaller sample, the detail and scope of the data allow for a broader range of research than is possible with the tax data.

This paper was originally presented at the American Statistical Association's 2008 Joint Statistical Meetings in Denver, CO.

The population of Federal income tax filers includes only those U.S. citizens and resident aliens whose gross income, a concept defined by statute, was above legislatively prescribed thresholds. Nonresident aliens are subject to different filing requirements, based on income earned in the U.S. Income tax filers represent roughly 61 percent of the U.S. individual population (see Sailer and Weber, 1999). In addition, recent income tax filing gap estimates for Tax Year 2000 suggest that as many as 11 million taxpayers, or about 9 percent of the potential income tax filing population, either file returns late or not at all (see Brown and Mazur, 2003). In contrast, the SCF sample design ensures coverage of the entire U.S. population.

The unit of observation in the case of federal income taxes can vary according to current filing regulations. Married couples may file returns jointly, but they are also allowed to file separately when marginal tax rates favor treating the two incomes separately. Dependent children and others living in a home may also be required to file separate returns to report both earned and unearned income. Differences in the economic unit reported on income tax returns limit the data's usefulness for some types of research.

In the SCF, the area-probability sample uses a sampling frame in which the household is the unit of observation, but, for the list sample, the unit of observation is the tax-filing unit. Often the tax-filing unit is analogous to the household, but, for certain households, such as households where a married couple files separately and those with multiple subhouseholds located within a household, there are differences. While there is the possibility of frame errors in the list sample, adjustments are made during the construction of the frame and during the sampling stage to limit these distortions (see Kennickell and McManus, 1993; Frankel and Kennickell, 1995; Kennickell, 1998a; and Kennickell, 2001).

Because income tax reporting requirements are established by legislation, data concepts and definitions may not necessarily coincide with those required for economic analysis. For example, income is combined for couples who file a joint income tax return, however, for some research purposes, it would be useful to know the amounts earned by each individual. Another consideration is that, while a precise geographic location is often useful for analytical purposes, mailing addresses present on tax records may not always be the appropriate location, as when a post office box number is supplied rather than a street address. Addresses on tax returns might also be those of paid preparers rather than the filers. In some instances, a filer who owns multiple residences may even file from the address that provides the best tax advantages, rather than the address that he or she would consider 'home.'

An important aspect of data content is continuity over time, both in the items included and in the data definitions. SOI goes to great lengths to ensure both in its annual data files. However, coverage and content are subject to discontinuities resulting from changes to laws, regulations, administrative practices, and program scope. For example, income tax law revisions in 1981, 1986, 1990, and 1993 all made significant changes, both to the components of income subject to taxation and to the allowable deductions from income, that had significant impact on the statistical uses of tax return data (see Petska and Strudler, 1999).

Since surveys have more flexibility than administrative systems to specify a conceptual framework, many issues related directly to the definition and scope of the data are less pressing. However, content and valuation issues of a different sort are present in survey data. Unit and item nonresponse are two important sources of nonsampling error in surveys, though there are methods to help deal with both these issues, such as sample weight adjustments and imputation. For respondents who agree to participate and answer all the survey questions, measurement error is still a concern in survey data. Respondents may "guestimate" answers to questions; even if respondents' guesses overall are not biased, such approximation reduces the estimation efficiency of the data. Respondents may also have difficulty recalling past events. Other typical measurement errors include rounding dollar amounts, misunderstanding questions, and altering responses due to stigma or prestige attached to certain behaviors or a desire to protect privacy. A large volume of research exists on measurement error and its effects on survey

data (see Lessler and Kalsbeek, 1992 and the references therein).

While it is true that, for administrative data, unit and item non-response are usually not a problem on core items, it is not clear that administrative data are always more accurate than survey data. For example, some individuals may intentionally misreport values on tax returns to reduce their tax liabilities–it is estimated that underreporting may have resulted in underpayment of as much as $120 billion in income taxes for Tax Year 1998 (Brown and Mazur, 2003). Those same individuals may report the true value in response to a survey question since there is no benefit to misreporting in the survey.

The income questions in the SCF are structured to allow respondents to reference their tax forms when answering the income questions. Figure 1 shows the correspondence between the income questions in the SCF and the line number on IRS Form 1040. The SCF income questions were designed to cover most forms of income that a household reports on its tax form. The figure shows that there is much overlap between the two data sources, although there are some differences. Since the SCF is interested in all sources of household income and not just income subject to taxation, the questions on pensions, IRA/401(k) distributions, annuities, and Social Security payments refer to the total amounts. The SCF also asks about any income received from government transfer programs (such TANF, SSI, and food stamps). Households are not questioned about any adjustments to total income, but households are questioned about their Adjusted Gross Income (AGI). All income amounts reported in the SCF are for the year prior to the survey year.

Even with the close correspondence between the income questions in the SCF and IRS Form 1040, accurate classification and reporting of income amounts are still a potential problem in the SCF. To improve comparability, respondents are encouraged to reference documents, including tax forms, during the interview. Figure 2 shows that, for the 2004 SCF, almost 21 percent of all households referenced their tax forms. This represents a significant increase over earlier surveys. Higher income respondents were more likely to use tax returns during their interviews. Almost 25 percent of those reporting at least $50,000 in adjusted gross income referenced tax forms in answering the income module of the SCF in 2004, compared to fewer than 18 percent of those with lower incomes.

▶ Comparisons Between SCF and SOI Estimates

Figure 3 provides a comparison of SCF and SOI estimates for Tax Years 1988, 1991, 1994, 1997, 2000, and 2003 and highlights the difference in the unit of observation between the two data sources. In the SCF, the unit of observation is the household, which can sometimes contain more than one tax unit. The SCF asks the filing status of the core individual or couple in a household, thus allowing married or partnered households filing separately to be counted as two returns. The SCF consistently underestimates the number of returns in the tax filing population, no doubt in large part because the SCF does not ask about the filing status of other individuals within the household. These individuals include dependents who may also file a return and other members of the household who are not financially dependent on the household head or the core couple. Estimates of the income tax filing population produced using the SCF have improved over time and differed from the actual total by less than 23 percent for Tax Year 2003. Despite significant differences in filing population estimates, the SCF and SOI estimates of total income differ by no more than approximately 10 percent in each Tax Year shown, with the SCF estimate larger in each case.

SCF estimates of wages and salaries, unemployment and alimony payments, and other income are consistently larger than those produced by SOI. The difference between the estimates of alimony income is due to definitional differences; the SCF question on alimony income instructs the respondent to include child support payments. Since child support payments are nontaxable, such payments should not be included in the SOI estimate. The differences between the SCF and SOI estimates of "other income" are difficult to pinpoint, given the wide range of types of income

Figure 1

Form **1040** Department of the Treasury—Internal Revenue Service **U.S. Individual Income Tax Return** **2000** (99) IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2000, or other tax year beginning , 2000, ending , 20 OMB No. 1545-0074

Label (See instructions on page 19.) Use the IRS label. Otherwise, please print or type.

LABEL HERE

Your first name and initial | Last name | Your social security number

If a joint return, spouse's first name and initial | Last name | Spouse's social security number

Home address (number and street). If you have a P.O. box, see page 19. | Apt. no.

City, town or post office, state, and ZIP code. If you have a foreign address, see page 19.

Important! You **must** enter your SSN(s) above.

Presidential Election Campaign (See page 19.) **Note.** Checking "Yes" will not change your tax or reduce your refund. Do you, or your spouse if filing a joint return, want $3 to go to this fund? . . . ▶ You ☐Yes ☐No Spouse ☐Yes ☐No

Filing Status Check only one box.

1 ☐ Single
2 ☐ Married filing joint return (even if only one had income)
3 ☐ Married filing separate return. Enter spouse's social security no. above and full name here. ▶
4 ☐ Head of household (with qualifying person). (See page 19.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 ☐ Qualifying widow(er) with dependent child (year spouse died ▶). (See page 19.)

Exemptions

6a ☐ **Yourself.** If your parent (or someone else) can claim you as a dependent on his or her tax return, **do not** check box 6a . . . . . No. of boxes checked on 6a and 6b
b ☐ **Spouse** . . . . . No. of your children on 6c who:
c **Dependents:** (1) First name Last name | (2) Dependent's social security number | (3) Dependent's relationship to you | (4) ✓ if qualifying child for child tax credit (see page 20)
- lived with you
- did not live with you due to divorce or separation (see page 20)

Dependents on 6c not entered above

If more than six dependents, see page 20.

Add numbers entered on lines above ▶

d Total number of exemptions claimed . . . . .

Income

Attach Forms W-2 and W-2G here. Also attach Form(s) 1099-R if tax was withheld.

If you did not get a W-2, see page 21.

Enclose, but do not attach, any payment. Also, please use Form 1040-V.

7	Wages, salaries, tips, etc. Attach Form(s) W-2		7 X5702
8a	**Taxable** interest. Attach Schedule B if required		8a X5708
b	**Tax-exempt** interest. **Do not** include on line 8a	8b X5706	
9	Ordinary dividends. Attach Schedule B if required		9 X5710
10	Taxable refunds, credits, or offsets of state and local income taxes (see page 22)		10
11	Alimony received		11 X5718
12	Business income or (loss). Attach Schedule C or C-EZ		12 X5704
13	Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ ☐		13 X5712
14	Other gains or (losses). Attach Form 4797		14 X5712
15a	Total IRA distributions . 15a X5724	b Taxable amount (see page 23)	15b
16a	Total pensions and annuities 16a X5722	b Taxable amount (see page 23)	16b
17	Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E		17 X5714
18	Farm income or (loss). Attach Schedule F		18 X5704
19	Unemployment compensation		19 X5716
20a	Social security benefits . 20a X5722	b Taxable amount (see page 25)	20b
21	Other income. List type and amount (see page 25)		21 X5724
22	Add the amounts in the far right column for lines 7 through 21. This is your **total income** ▶		22

Adjusted Gross Income

23	IRA deduction (see page 27)	23	
24	Student loan interest deduction (see page 27)	24	
25	Medical savings account deduction. Attach Form 8853	25	
26	Moving expenses. Attach Form 3903	26	
27	One-half of self-employment tax. Attach Schedule SE	27	
28	Self-employed health insurance deduction (see page 29)	28	
29	Self-employed SEP, SIMPLE, and qualified plans	29	
30	Penalty on early withdrawal of savings	30	
31a	Alimony paid b Recipient's SSN ▶	31a	
32	Add lines 23 through 31a		32
33	Subtract line 32 from line 22. This is your **adjusted gross income** ▶		33 X5751.X7651.X7652

For Disclosure, Privacy Act, and Paperwork Reduction Act Notice, see page 56. Cat. No. 11320B Form **1040** (2000)

Figure 2: Percent of Households Refering to Tax Forms During Field Interviews, 1989-2004 SCF

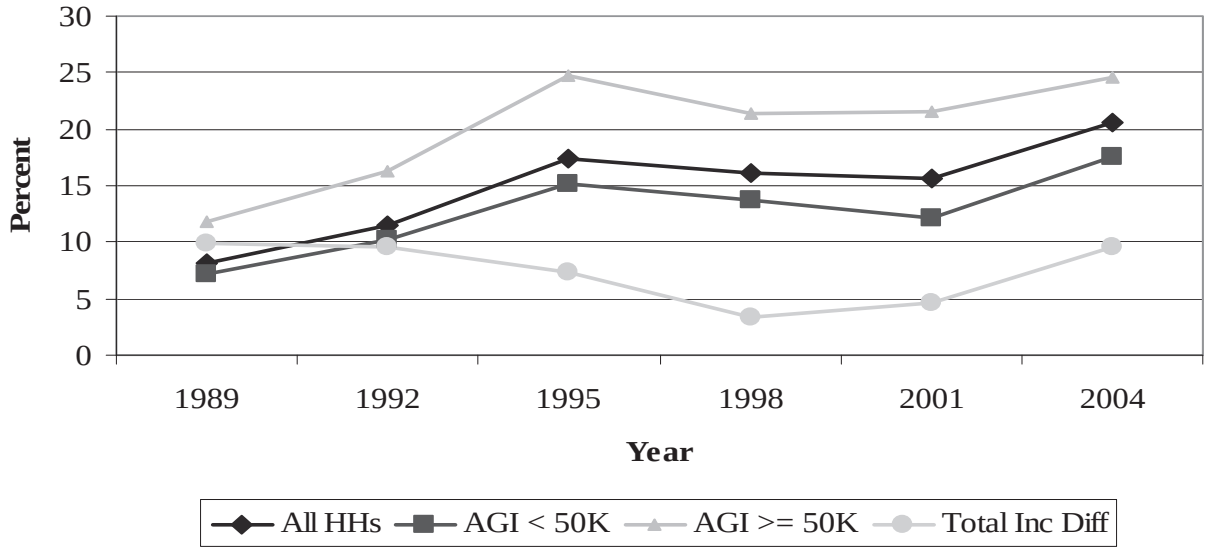
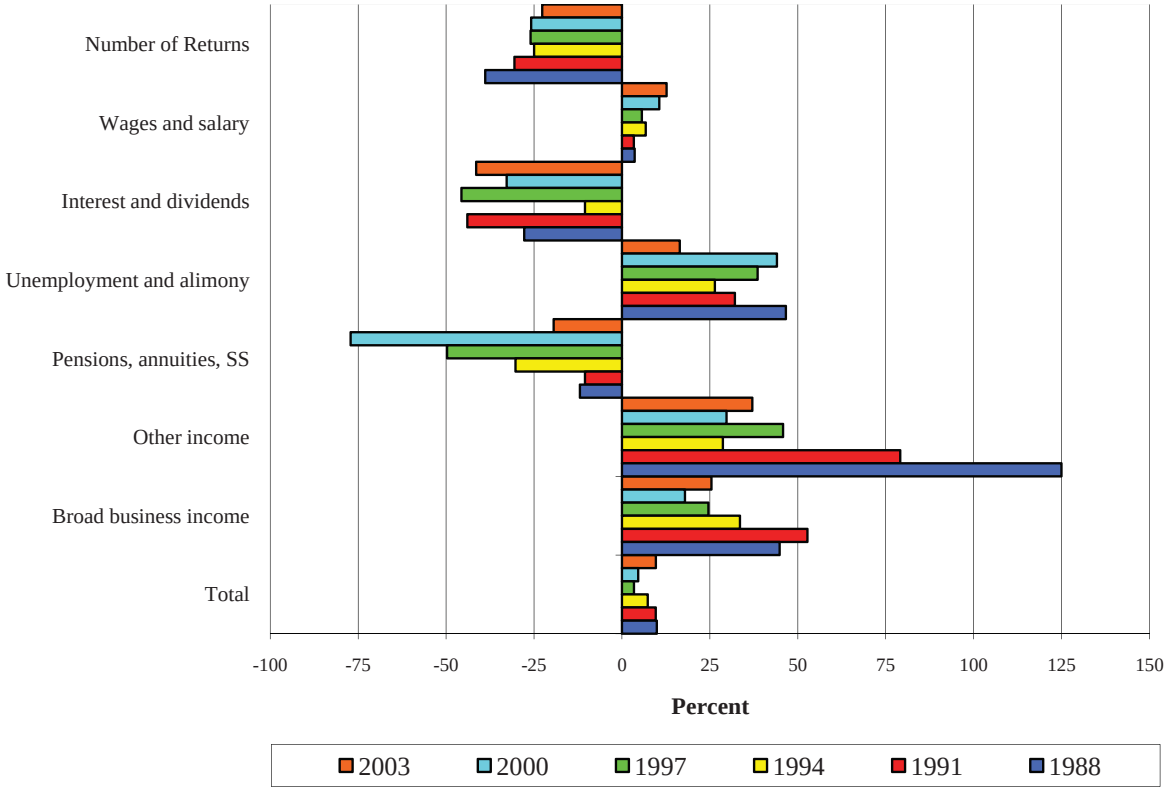


Figure 3: Percent Differences Between Aggregate SCF and SOI Estimates, Selected Tax Years



potentially included in that category. Of the income categories shown, estimates for wages and salaries derived from the two data sources are relatively small, increasing from just 3.6 percent for Tax Year 1988 to 12.7 percent for Tax Year 2003.

The SCF estimates of broad business income are also consistently larger than the SOI estimates. Broad business income combines sole proprietorship and farm income, capital gains, and rent, royalties, and subchapter S corporation income. These components are combined because households in the SCF may misclassify capital gains or rent, royalties, and subchapter S corporation income as sole proprietor income. This could be partially due to the order of the income questions in the SCF, since the sole proprietor and farm business income questions are asked early in the income sequence, while the capital gains and rent, royalties, and subchapter S corporation income questions are asked later in the sequence. Constructing a broader measure of business income eliminates some of these classification issues and reduces the differences substantially, especially for the three most recent tax years shown.

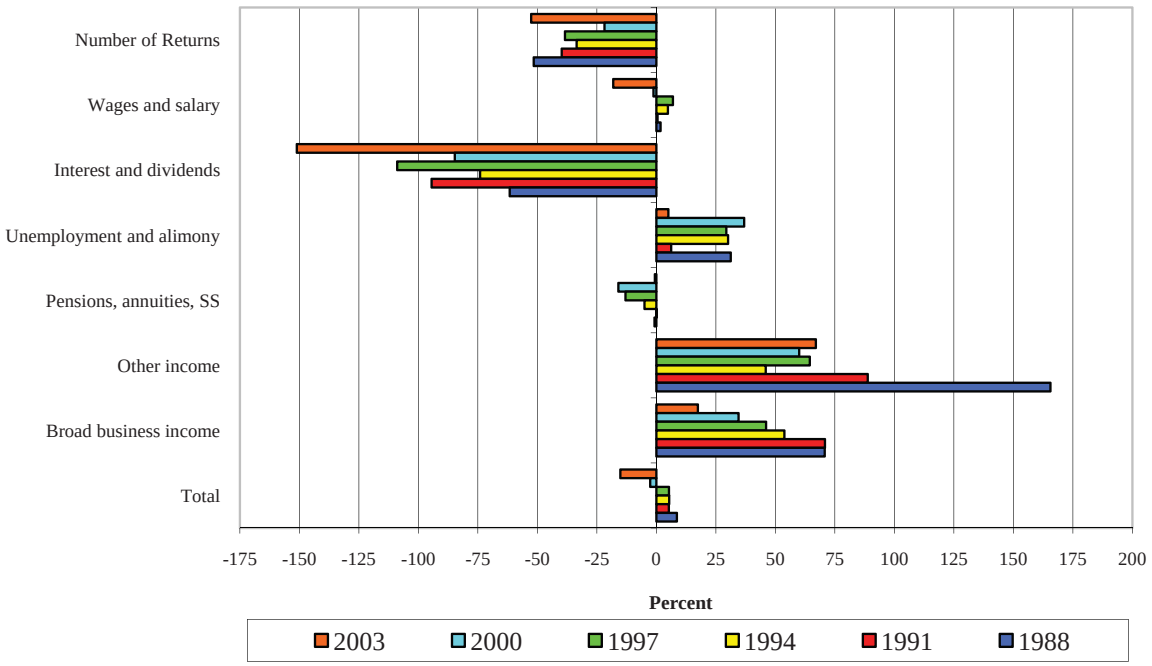
The SCF consistently underestimates the amount of interest (taxable and nontaxable) and dividends, as well as income from pensions, annuities, and Social Security. Differences between the SOI and SCF estimates of interest and dividends range from -10.5 percent to as much as -45.6 percent. One possible reason for these lower estimates is that households that receive only small amounts of taxable interest or dividend income may forget to report these amounts in the SCF questionnaire. Another possible reason is that households may not think they have "received" this income, particularly in the case of interest earned on bank accounts and money market funds. Even households with relatively large dividend and interest incomes may underestimate these values, due to the inherent variability of annual earnings, especially if they are not in a phase of life where such income is an important source of disposable income. The SCF understates the total of pension, annuity, and Social Security incomes by -10.5 percent to as much -77.1 percent, depending on the year. Using information reported in other sections of the SCF, it is possible to compute alternative estimates of pension, annuity and Social Security income. This computation reveals that (1) information in other sections of the survey corresponds closely with information provided in the income module of the SCF and (2) the SCF estimates of Social Security income are consistently similar to, but larger than the SOI estimates, while the SCF estimates of pension and annuity income are substantially less than the SOI estimates.

As noted previously, households in the SCF with at least $50,000 in AGI were much more likely to have referenced tax forms during the interview than lower income households. This suggests that households in the SCF with higher AGI should do a better job of reporting and classifying income. Data for respondents in these two AGI classes are shown in Figures 4 and 5.

For respondents in the less than $50,000 AGI group, estimates derived from SCF and SOI data for wages and salaries, unemployment and alimony, pensions, annuities and Social Security, and total income are all reasonably close. In contrast, estimates for interest and dividends are substantially different between the two sources. Again, this may be due to a large number of households neglecting to report relatively small amounts of interest income on the SCF. For example, in the 2004 SCF, only about a quarter of households with less than $50,000 in AGI that owned interest-bearing assets reported any interest income The median amount of interest-bearing assets for these households was only $1,200, suggesting that unreported interest would have been very small.

Figure 4 also shows that there is a sizeable difference in the estimate of broad business income for the less than $50,000 AGI group, although the difference has declined over time. Much of this difference is due to much larger estimates of rent, royalties, and subchapter S corporation income in the SCF and may be partly due to the treatment of losses in the survey. Although the SCF allows households to record negative amounts for certain income questions, households often report zero instead of the actual loss. Given the potentially favorable tax treatment of losses, actual losses are more likely to be reported to the IRS.

Figure 4: Percent Differences Between Aggregate SCF and SOI Estimates, AGI < $50,000, Selected Tax Years

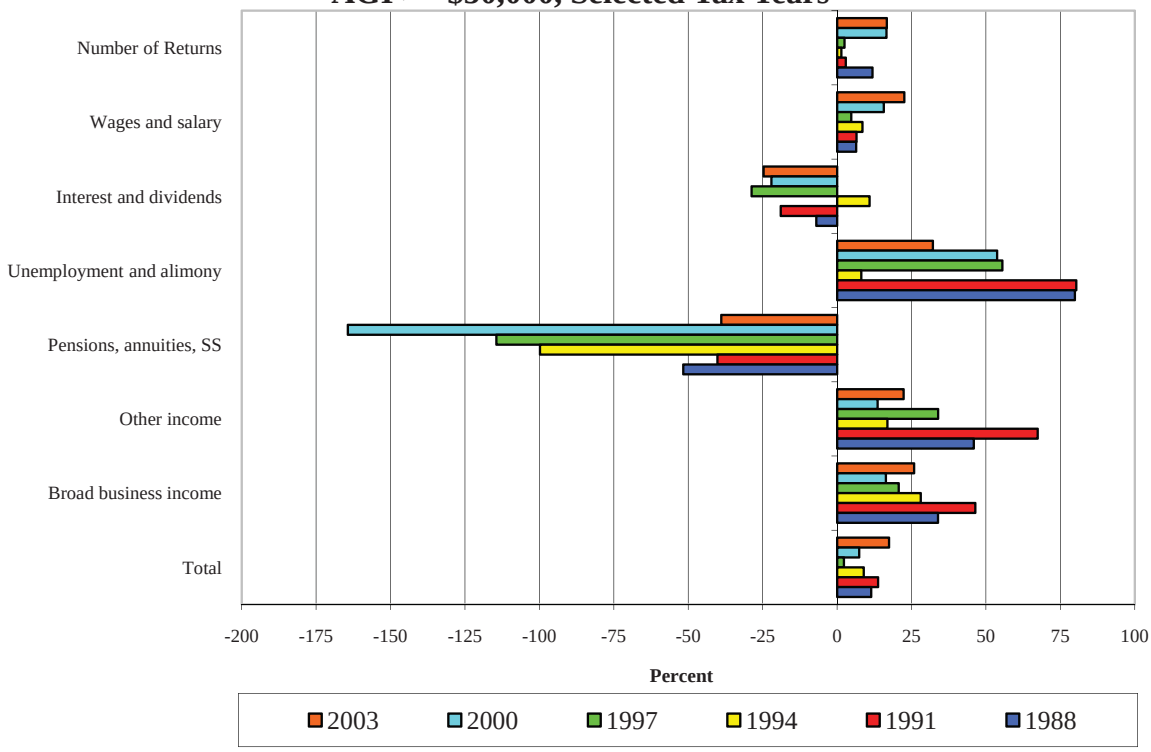


Turning to households with $50,000 or more in AGI, there is some evidence that the increased use of tax forms as references by members of this group improves the comparability between SCF and SOI estimates (see Figure 5). Estimates from these two data sources for the number of tax returns filed, as well as total income, wages and salaries, and interest and dividends differ by less than 30 percent. Also, the percentage differences in the broad business income estimates are smaller for households with $50,000 or more in AGI than for the lower income group. The SCF estimate for interest and dividends is less than the SOI amount in all but one year. Here again, only about 44 percent of households with $50,000 or more in AGI that owned interest-bearing assets reported any interest income, suggesting that even these respondents may neglect to report relatively small amounts. The median value of interest-bearing assets for these nonreporting households was about $6,000.

Most striking for the $50,000 or more AGI group are differences between the SCF and SOI estimates of pension, annuity, and Social Security income for all tax years shown. As with the estimates for all households, the summation of the alternative SCF estimates of pension, annuity, and Social Security incomes are very similar to the SCF estimate derived directly from the income questions. Also, the SCF estimates of Social Security income are typically fairly close to the SOI estimates. Thus, the bulk of the difference between the SCF and SOI estimates is due to pension and annuity income.

One possible reason for this discrepancy is the treatment of rollovers from one tax-deferred retirement account to another tax-deferred retirement account. For example, if a household transfers the balance of one IRA account to another IRA account, the transfer is not taxable, but the transfer amount should ap-

Figure 5: Percent Differences Between Aggregate SCF and SOI Estimates, AGI >= $50,000, Selected Tax Years



pear on line 16a of Form 1040 (see Figure 1). Often, households neglect to report these rollovers on their tax forms since there are no tax implications. However, the SOI estimate will include these rollovers, even if the household does not include them on its tax form.[1] Since households in the $50,000 or more AGI group are about twice as likely to have some sort of tax-deferred retirement account, these households are likely to have more rollovers. In published SOI estimates, a rough measure of the amount of rollovers is the difference between total and taxable pension and annuity income. For filers with $50,000 or more in AGI, about 60 percent of pension and annuity income is taxable, compared to about 80 percent for filers with less than $50,000 in AGI. If households in the SCF are not reporting their rollovers in the pension income question, this could explain most of the difference between these SCF and SOI estimates.

▶ Conclusion

In summary, the Survey of Consumer Finances contains an income module that is designed to capture information comparable to that reported on IRS Form 1040 for the tax year prior to the year in which the survey is conducted. Estimates produced from these data should closely match those produced by the Statistics of Income Division of the IRS. Indeed, taking into account differences in the reporting unit between the two data sources and sample variance, estimates of total income for each AGI group and tax year examined are remarkably close. Disaggregating total income into more detailed categories, however, reveals important differences.

Differences between estimates produced using SOI and SCF data are due in part to the idiosyncrasies of

[1] A rollover transaction generates a Form 1099-R that SOI matches to Form 1040. If a filer neglects to report the rollover on his or her tax form, the value from Form 1099-R is added to the filer's Form 1040.

the Tax Code. Some income items, including a portion of Social Security income, certain components of payments from a divorced spouse, and interest earned on some investments are exempt from taxation and are therefore excluded from SOI estimates. However, for the purpose of studying a household's economic condition, these items are necessarily included in estimates produced by the SCF. Other items, such as the allocation of depreciation on rental properties or the carryforward (or even backward) of business losses, are an important part of good tax planning, but are not easily captured within the structure of a household survey. The relative consistency of differences between SCF and SOI estimates over time, as shown in Figures 3, 4, and 5, suggests that they may be attributed primarily to these types of inherent disparities.

Figures 3, 4, and 5 do show significant improvements in the comparability of SCF and SOI estimates over time, which suggests that households sometimes classify income items differently in their survey responses than on tax returns. Some of this improvement is due to changes in the structure of the SCF over time. Cognitive testing and experience have led to some changes in both question design and the order in which questions are asked. An important change was the transition from a paper survey instrument to computer-assisted personal interviewing (CAPI) after the 1992 SCF. The CAPI instrument helps improve the quality of data collected by performing real-time tests intended to ensure that all dollar values are entered as reported by the respondent. CAPI also facilitates online tools, such as definitions and code lists, which improve the quality of data collected in the field. The research presented here also suggests that encouraging households to reference their tax forms is critical for improving the comparability of data between the SCF and SOI. Where classification differences persist, it is it is often possible to use information from other sections of the survey to make adjustments in order to better align the SCF and SOI income definitions. Ultimately, these classification differences may highlight the challenges some taxpayers face in classifying their incomes according to IRS reporting requirements. It is clear that, for some taxpayers, IRS distinctions between certain forms of income are blurred.

The goal of the research presented in this paper has not been to declare either the SCF or SOI data superior. Instead, we have attempted to document important similarities and differences between the two data sources. The detail and scope of the data collected in the SCF allow for a broader range of research than in the SOI tax data. The large sample size and administrative nature of SOI tax data make them appealing for certain types of research, such as studying some aspects of tax policy. The key, then, is that both data sources have strengths and weaknesses that need to be understood and carefully considered before attempting to use them to answer any set of research questions.

▶ Acknowledgements

The authors gratefully acknowledge the insightful comments, editing suggestions and encouragement provided by Arthur Kennickell, Martha Gangi, and James Dalton.

▶ References

Bucks, Brian , Kennickell, Arthur B., and Moore, Kevin B. (2006), "Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 92, A1-A38.

Brown, Robert E and Mark J. Mazur, (June 2003) "IRS' Comprehensive Approach to Compliance Measurement," 2003 National Tax Association Spring Symposium http://www.irs.gov/pub/irs-soi/mazur.pdf.

Frankel, Martin and Arthur B. Kennickell (1995) "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kennickell, Arthur B., (2001), "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B., (1999), "Revisions to the SCF Weighting Methodology: Accounting for Race/ Ethnicity and Homeownership," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (1998a) "List Sample Design for the 1998 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B., (1998b), "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association.

Kennickell, Arthur B. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kennickell, Arthur B. and Douglas McManus (1993) "Sampling for Household Financial Characteristics Using Frame Information on Past Income," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. and Woodburn, R. Louise, (1999), "Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth," *Review of Income and Wealth*, Series 45, 2, 193-215.

Lessler, Judith T. and William D. Kalsbeek (1992) *Nonsampling Error in Surveys*, New York: John Wiley & Sons.

Petska, Tom and Mike Strudler, (1999) "The Distribution of Individual Income Taxes: A New Look at an Old Issue," *Turning Administrative Systems Into Information Systems, Internal Revenue Service*, pp. 7-22.

Sailer, Peter and Michael Weber, (1999) "The IRS Population Count: An Update," Proceedings, Section on Survey Research Methods, American Statistical Association.

Statistics of Income—2003 Individual Income Tax Returns, Internal Revenue Service, Washington, DC, 2005.

Tourangeau, Roger, Johnson, Robert A., Qian, Jiahe, Shin, Hee-Choon, and Frankel, Martin R., (1993), "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago.

Selected Additional Readings

Eller, Martha B. and Barry W. Johnson (1999) "Using a Sample of Federal Estate Tax Returns to Examine the Effects of Audit Revaluation on Pre-Audit Estimates," American Statistical Association, *Proceedings of the Section on Government Statistics.*

Gale, William G., James R. Hines Jr., and Joel Slemrod (eds.) (2001) *Rethinking estate and gift taxation*. Washington, D.C.: Brookings Institution Press.

Gangi, Marth E. and Brian G. Raub (2006) "Utilization of Special Estate Tax Provisions for Family-Owned Farms and Closely Held Businesses" *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Summer pp 128-145.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen (1993). "The Carnegie Conjecture: Some Empirical Evidence," *The Quarterly Journal of Economics*, MIT Press, vol. 108(2), pp 413-35.

Jacobson, Darien B. (2004) "Federal Estate Tax Returns Filed for Nonresident Aliens, 2001 and 2002," *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Summer, pp 187-202.

_____.(2002) "Federal Estate Tax Returns Filed for Nonresident Aliens, 1999 and 2000," *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Summer, pp 66-81.

Johnson, Barry W. (ed.) (1994) *Compendium of Federal Estate Tax and Personal Wealth Studies Volume 1*, Washington DC, Internal Revenue Service.

Johnson, Barry W. and Martha Britton Eller (1998) "Federal Taxation of Inheritance and Wealth Transfers," in Miller, Robert K. Jr. and Stephen J. McNamee, (eds.) *Inheritance and Wealth in America*, Plenum Press New York, pp. 61-90.

Joulfaian, David (2000). "Estate Taxes and Charitable Bequests by the Wealthy," NBER Working Papers 7663, National Bureau of Economic Research, Inc

_____ (2005) "Estate Taxes and Charitable Bequests: Evidence from Two Tax Regimes" OTA Working Paper 92

_____ (1997) "The Federal Estate and Gift Tax: Description, Profile of Taxpayers, and Economic Consequences" Department of the Treasury OTA Working Paper 80

_____ (2007) "The Federal Gift Tax: History, Law, and Economics," US Department of the Treasury OTA Working Paper 100

_____ (2004) "Gift Taxes and Lifetime Transfers: Time Series Evidence," *Journal of Public Economics*, Volume 88, pp 1917-1929.

_____ (2000) "Taxing Wealth Transfers and Its Behavioral Consequences," *National Tax Journal* Vol. 53 no. 4 Part 1, pp 933-958.

Joulfaian, David and Kathleen McGarry (2004) "Estate and Gift Tax Incentives and Inter Vivos Giving," *National Tax Journal*, Vol. LVII, No. 2, Part 2, pp 429-444.

Joulfaian, David and Mark O. Wilhelm (1994) "Inheritance and Labor Supply," *Journal of Human Resources*, University of Wisconsin Press, vol. 29(4), pp 1205-1234.

Mikow, Jacob M. (2000) "Fiduciary Income Tax Returns 1997" *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Winter pp77-99.

Raub, Brian (2007) "Recent Changes in the Estate Tax Exemption Level and Filing Population," *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Summer, pp 114-119.

Rosenmerkel, Lisa S and Joseph Newcomb (2010) "Fiduciary Income Panel, Tax Years 2002-2006" *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Spring, pp 90-96.

Schreiber, Lisa M. (2005) "Fiduciary Income Tax Returns, Filing Years 2003 and 2004," *Statistics of Income Bulletin,* Washington DC, Internal Revenue Service, Fall, pp. 130-161.