Report on the State of Publicly Available Data and Statistics

For the Study of Public Economics

Subcommittee: Len Burman, Dan Feenberg, Austan Goolsbee, Charles Hulten, Bruce Meyer, John Karl Scholz, Joel Slemrod (chair)[1]

Preliminary draft[2]

December 15, 2005

1.   Introduction and Motivation

In his review of the challenges (and new approaches) to the release of IRS data, Greenia (2002) calls the tax data a "national treasure of vital information about the nation." He goes on to say that:

> "The uses of tax data go well beyond that of tax administration. The nation has long
> recognized their value not only for the formulation of tax policy and other administrative
> uses (such as Social Security) but also statistical purposes. Tax data, mostly in statistical
> non-identifiable form, are broadly used to assist the nation's decision makers both within
> the federal government such as Congress and the Administration and outside the federal
> government such as businesses, policy think tanks, academic research groups, and state
> entities."

As Abowd and Lane (2003) elaborate on, public access to micro data has many social benefits, including a) the analysis of complex policy questions that cannot be answered otherwise, b) calculating marginal rather than average effects of programs, c) providing a scientific safeguard and checks and balances in a democratic society through the possibility of independent replication, and d) assuring continually improving data quality through research use of the data to identify anomalies and interaction between data providers and data users in order to continually improve data quality.

[2] Although earlier versions have been circulated to the subcommittee, as of December 15, 2005 this draft has not been, so it does not fully reflect the full subcommittee's input.

To be sure, there are also costs to providing better data and better public access to the existing data. Including the resource cost, concerns about privacy and the potential disclosure of confidential information, as well as reputation costs if outsider use of the data is misconstrued as being official or if the research results cast doubt on existing government policies.

The subcommittee has been charged by the American Economic Association's Committee on Economic Statistics to prepare a study of the needs for data improvement in public economics, both from a research and public policy point of view. We have decided to focus our inquiry on federal tax-and-transfer issues, and to stay away from sub-fields in public economics—such as education, environment, and health—that tend to have their own particular institutions and therefore data issues, and, for the most part, from state and local public finance data issues.

To achieve this goal, the following three questions must be addressed:

- What are the major gaps in statistical data that hinder either basic or policy research in the economic analysis of tax and transfer programs? To place the answer to this question in proper context, we offer a brief summary of what data are already in the public domain.

- What initiatives could be taken to fill these gaps—either using existing surveys or administrative data, by merging existing data sets or by undertaking new data-gathering exercises?

- What obstacles—legal, bureaucratic, or other—now restrict access of academics and other researchers outside the government to relevant data?

- What institutional initiatives might overcome these obstacles?

2. Summary of Key Existing Public-Use Data Sets

2.1 Tax Return Data

Individual Income Tax Return Data: Cross-Section

The tax research community is fortunate to have, for (nearly) every year since 1960, a large sample of actual tax returns (at least 80,000 records) prepared for public use by the Statistics of Income Division (SOI) of the Internal Revenue Service. This file, dubbed the "Tax Model" by the IRS (although it contains only data and no modeling capability), includes almost 200 variables for each taxpayer. Virtually everything on the Form 1040 is available, together with several items from many supporting schedules.

The tax model file is a stratified random sample, with high sampling rates for wealthy taxpayers and small states. Certain easily traceable data items - such as alimony and real estate taxes - have been eliminated or obscured to protect taxpayer identity. (Strudler et al., 1987 and Winglee et al, 2002). Generally this is not a substantial hindrance to use. However, the data are as poor in demographic information (e.g., age, race, sex, occupation, and hours worked are not available) as they are rich in reported income information. State of residence is not available for taxpayers with incomes over $200,000. What is provided is nonetheless sufficient to calculate federal tax liabilities to a high degree of accuracy and state liabilities to a somewhat lower standard (where state is identified).

A complete collection of the public use files from 1960 to 1991 is available from the National Archives, and the 1992 to 2001 files can be obtained on CD-ROM from the SOI, and are available at numerous research centers. The non-public use files prior to 1979 were discarded at both SOI and the Office of Tax Analysis (OTA) at the Treasury Department, and apparently no other copies existed.

Individual Income Tax Return Data: Panel

Panel data are especially critical for understanding the behavioral responses to tax changes. In a cross-section, marginal tax rates are fully determined by income and other household characteristics (albeit, in a complex, nonlinear way), and so separately identifying the impact of income and net-of-tax price on a given behavior is exceedingly difficult, and often depends on tenuous assumptions that the researcher knows the functional relationship that links income, price, and the behavior under study. With panel data that span a tax change, with less (but not trivial) stringent assumptions one can more confidently associated changes in behavior to the change in tax structure that occurred.

Panel data facilitates the measurement of the lifetime resources of a household, which is particularly important for understanding the determinants of household behavior. It enables analyses that distinguish between the temporary and permanent responses to tax changes, and for analyses that attempt to identify the timing of the response to anticipated tax changes, both

those that are known (or suspected to be) short-lived, and those that are perceived to be long-lasting.

The provision of a single year's cross-section versus a panel also has a profound effect on the measure of the progressivity of a tax system. Only an annual income tax can track annual income perfectly, so that all deviations of reported annual income from lifetime income whether from survey error, imputation error, or life-cycle smoothing tend to make the annual income tax look better, and consumption-based taxes (even ones with the same lifetime incidence as an annual income tax) look worse when measured against a standard of annual income progressivity.

Only one panel dataset of tax return data has been made available for public use. This is the Continuous Work History (CWHS) panel, which includes all the variables from the annual cross sections for tax years 1979 to 1990. Indeed, the records are included in the annual cross sections. Taxpayers are selected for inclusion in the CWHS panel if their primary social security number has one of a small number of 4 digit endings. For budget reasons the number of included endings varied from 2 to 5, so that the sampling rate varies between 2 and 5 out of 10,000.

With this design taxpayers drop out of the panel randomly, or when they die or (if female) when they marry, and importantly return to the panel when they divorce or are widowed, or randomly. This contrasts with more common panel designs where lost respondents never rejoin. In fact, non-random attrition behavior in the CWHS panel mirrors the attrition in the universe, and cannot be characterized as bias or error. The CWHS has been used in several tax-price regression studies, including the influential recent studies of the elasticity of taxable income. However, given its unstratified nature, the CWHS panel is not well suited for studies of capital income, or the behavior of high income taxpayers.

The CWHS sample has continued to be collected in recent years, but is available only to government analysts. It has not been released for fear of disclosure. Partly this is the result of fear that an unrecognizable person might become recognizable through time, in which case it is too late to drop identifying information such as the state code. Another source of risk would be a divorce. Unless the primary taxpayer was dropped from the sample, a former spouse who has retained old tax return information could identify his or her subsequent returns. The SOI has also prepared numerous "prospective panels". Typically these start with a subset of the stratified cross section in the annual Tax Model, and collect all subsequent tax returns for those taxpayers over the next 5 to 10 years. No such panel has ever been released for public use, however.

Business Data

Most of the business information used in public finance comes from the aggregated information released by the IRS through the Statistics of Income. These data are divided by broad categories

such as by firm size or by industry but there is no means of creating cross-tabulations or special requests of data, the way that one can do with the economic censuses, for example.

There is substantial business and industry information collected by the Census Bureau through the censuses of manufacturing, services, retail, and so on and they take place every five years. There is also firm-level detailed information in the COMPUSTAT data that gives the entire financial reports of every publicly traded company in every year (and often includes some international information, as well).

The Longitudinal Business Database is a longitudinally linked business register from the Census Bureau allows measurement of dynamics for all businesses by industry, geography, size, age. It also allows measurement of activity at both the establishment and firm (enterprise) level. Currently, LBD only includes establishments with at least one employee.

The Longitudinal Database on Businesses, maintained by the Bureau of Labor Statistics. contains longitudinally linked data on employers, by state, which allow measurement of job creation and reduction by employer characteristics (including establishment size, age, industry, location). The source data are State 202 files from the Unemployment Insurance (UI) system.

Longitudinal research data on manufacturing establishments, based on the Census Bureau's Annual Survey of Manufacturing and Census of Manufacturing, provides data from 1963 to the present, and allows measurement of productivity (labor and total factor productivity) and dynamics (entry/exit, job creation and destruction)

The Longitudinal Employer-Household Dynamics data longitudinally links UI wage records and combines Census business and household data to produce matched employer-employee data so that both worker flows (accessions/separations) and job flows (job creation/reduction) can be measured at sub-state geographic levels.

A problem for public finance research is that these extensive business data sources cannot be matched to data on taxes in any systematic way. The difference between financial and tax accounting in the U.S. means that financial income, for example, does not match taxable income. Researchers have had to impute taxable income or approximate the match by industry in an effort to come up with sensible data. One of the frustrating things about this is the fact that in many cases the census bureau derives its sample from the IRS database of firms. In other words, all of the firms are naturally matched by their tax-ID numbers, but there is nothing reported at even an aggregate level that allows researchers to do the same type of matching.

In order of detail, the following would greatly help our ability to research and understand the operation of the economy.

2.2 Non-tax-return Data

Consumer Expenditure Survey (CES)

The Consumer Expenditure (CES) survey is a nationally representative survey conducted by the Bureau of Labor Statistics (BLS) that is designed to provide a continuous summary of the spending habits of U.S. households. The survey gathers expenditure data at the consumer unit level. The BLS estimates that the survey accounts for up to 95 percent of all household expenditures, making it the most comprehensive survey of expenditures for U.S. households. The CES also reports detailed information on demographic characteristics as well as employment and income information for each member of the consumer unit age 14 and over. The CES includes two separate survey instruments–the Interview Survey and the Diary Survey. The Interview Survey is a rotating panel survey that interviews approximately 5,000 households quarterly and follows each household for up to five consecutive quarters. The survey asks comprehensive questions about a wide variety of expenditures. From these questions, the BLS provides data on more than 600 unique expenditure categories. The BLS also conducts a separate diary survey that provides more detailed information on smaller or more frequent expenditures that tend to be more difficult to recall.

The CES is especially valuable for the analysis of the distributional and behavioral implications of excise taxes, but can also shed light on the effects of proposals for broad consumption taxes. It has also been used to examine the macroeconomic implications of tax cuts to combat recessions by analyzing the magnitude and timing of spending responses to the tax cuts.

Survey of Consumer Finances (SCF)

The Surveys of Consumer Finances (SCF) have been fielded every three years (beginning in 1983) by the Board of Governors of the Federal Reserve System, in cooperation with other federal agencies. The SCF gathers detailed information on the assets and liabilities of a random, stratified cross-section of American families. The SCF intentionally oversamples very high-income taxpayers, which allows the SCF to capture the top tail of the wealth distribution in a way other surveys with wealth information cannot. They are widely regarded as the most reliable sources of information on the financial status of U.S. households. Although this data set does not

provide any federal tax information, it does in fact employ such information in the process of selecting the sample; this is enabled through the fact that this work is considered to be beneficial to tax administration beneficial.

All of the SCFs fielded between 1989 and 2001 share a similar structure. Each gathers extensive information on assets, liabilities, and demographic characteristics. The creators of the SCF provided five separate imputed values (replicates) for each missing variable. They also selected 999 sample replicates from the final data in a way that allows users to capture important dimensions of sample variation. The sample replicates are particularly useful because confidentiality concerns prohibit the release of information on the survey's stratification design.

The SCF has information on only roughly 4,000 households each wave. The relatively small sample size makes it (mostly) impossible to look at interesting subsamples of the data. For example, one strategy for better understanding factors affecting wealth accumulation is to examine differences in wealth between exogenously defined groups. For example, black families appear to have less wealth than white families, even after conditioning on observable characteristics. Understanding the factors accounting for those differences might illuminate general factors that are important determinants of the vast cross-sectional differences in wealth. But the small SCF sample sizes preclude thorough analyses that might adopt this strategy.

The SCFs have been used to examine many issues in public economics, including the effects of tax incentives on saving, factors affecting housing markets, the role of pensions in wealth accumulation, and studies examining factors affecting the evolution of wealth inequality.

Current Population Survey

The Current Population Survey (CPS) is a nationally representative monthly survey of approximately 60,000 households. The CPS is the most commonly used source of nationally representative income data and the source of our official unemployment rate and poverty rate data. The March CPS files include the Annual Income Supplement data, which are the most extensively used by researchers. In the March interview respondents are asked to provide detailed retrospective information including usual hours worked, weeks worked during the previous year, and income for the previous year from a variety of sources including earning and transfer income.

CPS data play a prominent role in studies of expenditure programs. A central design issue is how program parameters affect program takeup, incomes, and employment. CPS data capture these outcomes fairly well.

HRS/AHEAD

The Health and Retirement Study (the HRS) is a national panel study, with an initial sample of over 12,600 persons in 7,600 households. It oversamples Hispanics, Blacks, and Florida residents. The baseline interviews were conducted in-home, face-to-face in 1992 for the 1931-41 birth cohort (and their spouses, if married, regardless of age). Follow-ups interviews occur by telephone every second year, with proxy interviews after death.

The survey collects extensive information on health and cognitive conditions and status; retirement plans and perspectives; attitudes, preferences, expectations, and subjective probabilities about salient life events; family structure and financial transfers; employment status and job history; job demands and requirements; disability; demographic characteristics; housing; income and net worth; health insurance and pension plans; and some experimental modules. The Study of Asset and Health Dynamics of the Oldest Old (AHEAD) data are a national panel study with an initial sample of 7,447 respondents aged 70+ (and their spouses, if married, regardless of age) taken from the HRS household screening. The study has a supplemental sample of respondents aged 80+ from the Medicare Master Enrollment File (HCFA). Like the HRS, it oversamples of Blacks, Hispanics, and Florida residents. The baseline survey was in-home, and had face-to-face interviews conducted in 1993, including spouses, if married, regardless of age. Follow-ups by telephone with respondents every other year, including proxy interview after death. Wave 1 data collection was completed in February 1994 and Wave 2 data collection ended in May 1996.

In 1998 the HRS and AHEAD studies were merged into a single data collection effort. The new study included Wave 4 of HRS and Wave 3 of AHEAD. Baseline information was added for two new sub-samples: Children of the Depression (CODA)--persons born in 1924 through 1930; War Baby (WB)--persons born in 1942 through 1947. The HRS is now a panel study, representing all persons over 50 years of age in the United States.

The HRS has been and will continue to be used in a large number of public finance, aging, and studies in other fields. The data are particularly well suited for examining the relationship between health, income, and wealth over time; examining life cycle patterns of wealth accumulation and consumption; monitoring work disability; examining the timing, magnitude and

correlates of intergenerational transfers; and examining how the mix and distribution of economic, family and program resources affect key outcomes, including retirement, dissaving, health declines and institutionalization.

Survey of Income and Program Participation (SIPP)

The Survey of Income and Program Participation (SIPP) is a nationally representative panel survey covering roughly 40,000 households. Recent panels were initiated in 1993, 1996, and 2001. Each panel is divided into waves; each wave covers four months. Within a panel, interviewers contact survey respondents at the end of every wave and collect monthly data on a variety of topics, including demographic characteristics, labor force attachment, income, and program participation. The SIPP also include topical modules in each wave focusing on information such as assets, child well-being, and taxes.

The tax topical module has been included on an approximately annual basis in the SIPP for nearly twenty years. Although it has also been referred to as the "Earnings and Benefits" module and the "Property Income and Taxes" module in past panels, the content has remained comparable. The module contains information relating to federal taxes and property taxes. Examples include adjusted gross income, tax liabilities, filing status, tax forms filed, exemptions, deductions, and credits. More specifically, these modules are of particular value because they address a number of unique and interesting research questions. Within the modules it is possible to study the relationships of a variety of demographic and economic characteristics with respect to tax related information. Potentially, researchers could investigate topics such as how tax deductible contributions vary with whether individuals file taxes, how daycare arrangements compare with child and dependent care expense credits, or how many families rise above the poverty line by claiming the earned income tax credit.

By merging the tax topical modules with core interview questions it becomes possible to provide a greater breadth of analyses. Furthermore, the tax topical modules of multiple years can be merged to incorporate a panel component into questions. For completed SIPP panels, each individual will potentially have responded to four years of tax topical modules as well as twelve waves of core questions.   The added breadth of information allows a single merged dataset to encompass individual tax data as well as participation in programs such as social security, food stamps, public housing, Medicaid, school lunch, and pensions. Additionally, it allows for the individual tax data to be used in conjunction with a wider variety of income and demographic information. Although the panels are only four years long, comparable data for time series analyses are available in the form of past panels.

Panel Study of Income Dynamics (PSID)

The Panel Study of Income Dynamics (PSID) is a longitudinal survey of a nationally representative sample of US families begun in 1968. 34 waves of data have been collected on the same families and their descendents as of 2005. Through its long-term measures of economic and social well being the survey has been the key source of information to study the dynamism inherent in social and behavioral processes. The PSID's innovative design and long-term panel have been central to the fundamental understanding of many key social science issues: intergenerational relations, income, poverty and wealth; cyclical behavior of wages, labor supply and consumption; savings, wealth accumulation, and transfers; demographic events such as teen childbearing, marriage, divorce, living arrangements, mortality; labor market behavior; health behaviors including exercise, smoking and eating; and the effect of neighborhoods. Through most of its history, the PSID has been annual, but in 1997 it moved to every-other-year interviewing.

As a consequence of low attrition rates, the success in following young adults as they form their own families, and efforts to contact sample members who declined an interview in prior years, the sample size grew from 4,800 families in 1968 to 8,450 in 1996. Had the sample not been reduced to 6,168 in 1997, it was projected that by 2006 the number of families would have reached the 10,000 mark. Today the sample has grown to 7,776 families, and over its 35 year history the PSID has collected information on more than 60,000 individuals and 40,000 variables. The PSID provides detailed information on employment, earnings and transfer income. The PSID data files provide a wide variety of information about both families and individuals collected over the span of the study. The central focus has been the collection of economic, demographic, and health data, with substantial detail on income sources and amounts, employment, family composition changes, residential location, health status, health insurance, health expenditures and health behaviors. Content of a more sociological or psychological nature is also included in some waves of the study. Although the PSID does not survey families about all expenditures, it does collect data on food and housing expenditures at the family level.

3. Uses of Data

Micro data sources are regularly put to a variety of uses, which can generally be categorized as (1) revenue scoring (2) revenue forecasting (3) distributional analysis and (4) behavioral response studies.

Program Evaluation

A considerable amount of public finance research seeks to understand the effects of public programs on individual and household behavior. For example, the creation of the Temporary Assistance for Need Families (TANF) program under the Personal Responsibility and Work Opportunities Reconciliation Act (PRWORA) of 1996 and the expansions of the Earned Income Tax Credit (EITC) over the last decade, increased attention to the employment experiences, labor market earnings, and transfer income received by disadvantaged individuals and households in the U.S. Studies of the EITC and TANF/PRWORA tend to focus on whether low-income households can achieve self-sufficiency without resorting to welfare or other public assistance programs. While income and employment levels are only partial indicators of the well-being of households, they continue to be ones most often used to assess the consequences, intended and unintended, of welfare reform.

More broadly, good measures of income and employment and broader measures of family well-being for low-income families are necessary to (1) assess the circumstances and labor market attachment of low-income and welfare populations at the national, state and local levels; (2) evaluate welfare reform and learn the effects of other programs adopted by state and federal governments; and (3) meet reporting requirements under welfare and other programs and to aid in the administration of these programs.

The EITC and the TANF programs are just two of the many examples of programs that have been examined using survey or administrative data. There are substantial bodies of work examining other programs including Social Security retirement, Social Security disability, unemployment insurance, workers' compensation, supplemental security income (SSI), Food Stamps, Medicaid, Medicare, and housing assistance. Research on these programs seeks to determine who receives these program benefits, the consequences of these programs for the distribution of income, the effects of the programs on employment and hours worked, on savings, family formation, living arrangements and other outcomes.

Distributional Analysis:

Numerous authors are concerned with the progressivity of the income tax, or the extent to which various provisions of the tax code affect the progressivity of the tax system. The Tax Model is by far the best dataset for this purpose, as other datasets typically exclude most high-income taxpayers, and do not have information about tax avoidance measures taken. The incidence of the corporate income tax is of continuing interest, and might be informed by information about the ownership of stock from data on dividends received and capital gains and losses.

Behavioral Response Analysis:

The non-linear income tax, and the variety of state income taxes create non-experimental variation in the after-tax price of leisure, charitable giving, savings, home ownership, medical care and other goods. This has been exploited by numerous authors to measure demand elasticities. Where goods are differentially treated by the tax law, the estimates obtained show the behavioral effect of the tax subsidy, and can sometimes be used to inform the tax expenditure estimate for an expansion of such subsidies. Panel data facilitates separating out temporary and permanent responses to a given tax change, and the behavioral response to tax changes perceived to be temporary and those perceived to be permanent.

Revenue Scoring:

These files are typically used to prepare estimates of the revenue effect of proposed changes in the tax law in total and by income class. These are commonly referred to by practitioners as "scores." Provided that additional data are not required, the procedure is as simple as calculating tax liabilities under the old and proposed regimes, and tabulating the result, possibly accounting for the effect of changes in marginal tax rates on behavior. The microsimulation overcomes the difficulty of estimating revenue effects from aggregate data when the tax law is nonlinear and depends in a complex way on a large number of variables with strong correlations.

While revenue scoring is chiefly a government activity, done by agencies with full access to non-public use data, a number of private entities, including the National Bureau of Economic Research, the Urban-Brookings Tax Policy Center and various accounting firms, have created similar models, and their estimates often match the results of OTA and JCT fairly closely. This suggests that most of the limitations affecting outside researchers also affect the official estimators.

Revenue Forecasting:

Forecasts of aggregate income tax revenue are chiefly influenced by economic growth and the business cycle but can also be informed by micro data, even in years when the tax law itself is constant. For example, approximately 25 percent of the large revenue surprises of 1996-2000 were explainable by real bracket creep. Without micro data, the size of that effect would be difficult to estimate due to the non-linearity of the tax system that produces marginal tax artes that differ across taxpayers. Much of the remainder of the revenue surprise is now attributed to rapid growth in income among those in the highest tax bracket, especially due to capital gains on stock

options, which are reported as wages on the 1040, and therefore not available as a separate category in the Tax Model, or indeed in any administrative dataset available to the revenue forecasters, again emphasizing that certain important difficulties of analysis are not presently caused by disclosure avoidance, and affect insiders as much as others. (Employers were required to report exercised stock options as a separate line on Form W-2 starting in 2003, which should help remedy this particular data gap.)

4. Gaps in Existing Public-Use Data Sets

The foregoing discussion suggests that, ideally, analysts of tax-and-transfer policy would have access to stratified panel micro dataset with extensive tax return and transfer receipt information as well as information on income, consumption, wealth, and demographics. Obviously it is a long way from where we are now to such an ideal dataset. But there are places where incremental improvements could be made that would be important.

Adding Demographic Information to Tax Return Data

Because tax return data are administrative data, information that is superfluous to the task at hand—documenting tax liability—is not collected.   Thus, for example, the age, race and sex of the taxpayer is not collected, nor is the number of hours worked or even the breakdown by individual of income and income-related activities for married coupled filing jointly, rendering the data inadequate for many research questions in labor economics. Without wealth or taxable basis information, studies of capital taxation are bound to be of limited reliability. Some of these problems can be ameliorated from other administrative sources, and the IRS receives from the Social Security Administration some demographic information, particularly gender and date of birth, and matches this to the tax return information. No recent matches have been made public, however. Internal files with detailed transaction data for capital gains have been made repeatedly, but except for 1973 have not been released for public use. Even those files have no wealth information.

Information returns filed by employers to the IRS include much valuable information, including the number of jobs held, the industry of each job, the division of wage income between spouses, sources of property income by industry, etc. These data are routinely matched for internal government purposes, but have not been made public in the recent past.

Some users of the Tax Model have compensated for the absence of demographic data with statistical merges of the Tax Model and the CPS. While convenient for revenue estimation and

forecasting as well as for distributional analysis, they are not of much analytical use. Obviously one cannot regress an item sourced from one file on regressors from the other file, as any relationship would, except by construction, be spurious.

The requirements for revenue scoring are weaker because the merge doesn't have to be perfect, or even to reflect all the covariances in the real world. It does have, though, to measure accurately the covariances of the relevant variables with the marginal tax rate. As far as we know, no study of how well extant statistical merges achieve this objective, although analysts at the Census Bureau could do this kind of validation study by making use of an exact match of tax records to (for example) a CPS or SIPP survey.

A potential solution to the lack of demographics on tax-related data might be found with the SIPP tax topical modules. The SIPP contains extensive information on household characteristics, program participation and income. If the tax topical module contained accurate information on filing status and tax-related variables, it would overcome the problems described above. Moreover, the data would facilitate study of factors affecting the decision to file, since SIPP, unlike federal tax returns, collects information on filers and non-filers.

The SIPP tax topical modules are intended to be representative of the civilian, non-institutionalized U.S. population. In each core wave, individuals are assigned weights such that the sample is representative of the population. These weights take into account factors such as "differential response, differential coverage, and differential attrition" within the sample. The problem lies in the fact that the topical module individual weights are not calculated; rather the corresponding core weights are used. Although the SIPP's extensive editing process is designed to impute missing values and adjust for internal inconsistencies, the topical modules are all but unusable prior to editing. This is due to the fact that a large number of survey participants answer core questions relative to those who answer questions in the topical modules. Moreover, the response rate within the tax topical module varies significantly between questions. As an illustration of these issues, there were approximately 70,000 individuals who responded to core questions in wave four of the 2001 panel of the SIPP. Taking weights into account, those respondents represented 282 million people, compared to 288 million in the population. Within the tax topical module, however, eligible respondents (all individuals over the age of 15) representing approximately 52 million people did not answer the question, "did you file federal income taxes for 2001?" Imputing responses for spouses in which one partner responded and the other partner did not respond makes it possible to decrease the non-respondents to 5451 participants representing approximately 20 million people. This is still 8 percent of the target population, and could lead to selection bias issues. As an added complication, this question had a relatively high response rate compared to many other questions in the tax topical module. For example, after imputing responses for spouses, only 12,612 participants representing 52 million

people responded to the question, "what was your adjusted gross income in 2001?" Comparing this number to the IRS estimate of 181 million tax filers in 2001 clearly illustrates the potential for selection bias.

The cause of widespread attrition in the transition from core to topical questions and the decade long lag in editing are problems that can only be addressed properly within the Census Bureau or another federal statistical agency. From the perspective of a researcher outside the Census Bureau, however, the choice is to use an unedited tax module with its associated selection bias risks or to use an older tax module (the most recent edited module refers to the 1996 tax year). For researchers tempted to use more recent modules, IRS data presents a possible means to address selection bias. By comparing tax topical module responses to IRS return data it would be possible in many cases to evaluate the accuracy and value of the SIPP data for a particular research question. Indeed, despite the problem mentioned above, the tax topical module may represent particular subsets of the population very well. In addition, in some applications it may only be necessary that it represents subsets of the population in accurate proportions instead of absolute numbers. A preliminary investigation of the adjusted gross income question addressed above suggests that the proportions of tax returns in each income bracket do match much more closely than the absolute numbers.

Matching Consumption Data with Tax Data

Matching consumption data with tax return data would be very valuable for many purposes. For example, theory suggests that, more so than current income, consumption should reflect lifetime income--both achieved and prospective. This correspondence is perfect only under very strong assumptions, such as perfect capital markets but, as mentioned earlier, it would nevertheless be insightful for studying the lifetime progressivity of the tax system to classify taxpayers by current consumption. The income tax return contains no useful consumption data other than, arguably, charitable contributions. Furthermore, arguably consumption is a better measure than income of economic well-being, especially for low-income families, and there is compelling evidence that income is badly underreported in household surveys at the bottom of the income distribution. Consequently, research on antipoverty programs should also focus on consumption, if suitable data were available. There is also evidence about consumption drops at retirement, which is sometimes taken as evidence of inadequate saving of households during pre-retirement years

Faced with this problem, many users of the data have made consumption imputations taken from the Consumer Expenditure Survey. Such imputations are less useful than might appear, and not only because the CES accounts for only 70 percent of NIPA consumption. Consider the effect of matching a low-income, high-consumption record from the CES to a low-income taxpayer in the Tax Model. There are many such records in the CES, possibly genuine, but many probably reflect

incomplete income reporting. Because income reporting is tangential to the main purpose of the CES, the BLS interviewer may not be aggressive in following up income questions. The incorrect income data will not affect calculations of the progressivity of the income tax (a low income tax will be attributed by calculation to a low income record), but will make a consumption tax look more regressive than it really is because the high consumption will imply a high consumption tax on a low income taxpayer. The nature of this argument is not affected if the income data is taken from the CES rather than the matched 1040 -- the match is on the basis of income (among other items).

One avenue of improvement is to improve the income and wealth data in the CES. Another is to add and improve the quality of the consumption data in large household (and preferably panel) surveys such as the SIPP, PSID, and CPS. Even the PSID consumption questions would likely prove useful. The HRS/AHEAD has experimented with a module in 2001 and in 2003 asking a small set of consumption questions. There is some evidence that the responses to the 32-question consumption module do a very good job of tracking consumption patterns in the consumer expenditure survey (Hurd and Rohwedder, 2005). This raises the possibility that other surveys might be able to elicit useful information on consumption in a considerably less costly way than many had thought.

Adding Tax Information to Non-Tax-Return Data

Currently, publicly available survey data sets often impute very basic tax variables and ignore others, and so may be completely mischaracterize households' tax liability and marginal tax rate on publicly available survey data. Matched tax return and survey data would also be useful for understanding the error in income reported in survey data. Currently, some government agencies have access to selected years of the CPS and the SIPP that are matched to tax return data, but those files are not available to the public. (There are also many households in the survey databases that cannot be matched to tax returns, often because of an invalid social security number or an unwillingness to report a social security number. It is unlikely, but not inconceivable, that the match rate is uncorrelated with other variables of interest, such as tax liability.) Putting aside the issue of improved access, publicly available studies that evaluate the quality of imputation procedures would be valuable.

With respect to business returns where we presume that public access to micro data is not an option, establishing an automated tabulating/cross-tabulating system for use on the SOI data, like the one used at the Census Bureau, would be a great boon to research and would have, essentially, no danger of disclosure (the Census system automatically withholds any information that relates to too small a cell that individual information could be identifiable). This automatic tabulation system would allow researchers to at least better match the existing non-tax

information on businesses to the information in the tax returns. An alternative would be for the Census Bureau to ask for tax information in some of their business surveys, so there would be less or even no need to match anything to tax data. As an example, the *Characteristics of Business Owners* survey done in 1992 asked extensive questions about the demographics and backgrounds of the business owners, as well as many questions about the operation of their business. Importantly, though, they also asked some questions like "how much income did your business report on its tax return last year," and some other questions relating to their tax position.

5. Privacy and Disclosure Concerns

Particularly sensitive issues surround the privacy and disclosure of information from tax returns. (This is not to say that the some of the same issues do not apply to survey data.) Current statutes require that federal tax information (FTI) must be protected in perpetuity, compared to a 30-year statute of limitations for Census business data, and 72 years for demographic data. The protected return information includes not only such obviously sensitive items as income and tax liability, but also variables relating to the taxpayer's identity and address, whether a taxpayer did or did not file a return, as well as the nature and source of income, and information regarding the actual or possible investigation of a return. Given the proliferation of information about individuals and organizations in the public domain and the technology-facilitated access to that information, protecting the confidentiality of micro data is a daunting challenge.

Section 6103 of the Internal Revenue Code prohibits disclosure of tax returns and tax return information to federal and state agencies or employees (or to anyone else) except for exceptions provided in the code section. It also provides criminal penalties for anyone, including IRS employees, who make unauthorized disclosures or inspections of federal tax information. The prohibition applies to officers, employees, and agents who have access to returns or return information by virtue of permitted disclosures of such returns or information under Code Section 6103. Return information includes the taxpayer's identity, whether a taxpayer did or did not file a return, the nature, source, and amount of income, payments, receipts, deductions, net worth, liability, deficiencies, closing (or similar) agreements, and information regarding the actual or possible investigation of a return.

Although the law requires the IRS (in Section 6108c) and authorized statistical recipients (in Section 6103(j)(4)) to make anonymous any federal tax information that is publicly released, the practical operational position taken by the IRS is that tax data must be protected from potential intruders who, using "reasonable means, might attempt to identify the taxpayer." Reasonable

means includes the use of available computer technology, public or privately available data, mathematical and statistical techniques, and a working knowledge of the subject matter to which the data apply. What the implications of that are for public access to micro data is evolving, and researchers must take the privacy concerns seriously. Moreover, because the "reasonable means" standard is technology-relative, the confidentiality protection of federal tax information is of particular concern given the lack of a statute of limitations.

An additional concern for statistical uses of tax data concerns the fundamental conflict in administrative and statistical agency mandates regarding the use of administrative, especially tax, data. On the one hand, IRS is mandated to provide FTI only for authorized purposes and to the minimum extent necessary. Conversely, statistical agencies like Census are mandated to use existing data systems, including administrative records, to the maximum extent possible. Reconciling such a difference needs to be undertaken carefully and any reconciliation effected needs to be demonstrably credible.

The public release of data either in tabular form or in public-use files is produced by the IRS Statistics of Income Division under either Section 6108(a) or 6108(b) of the Internal Revenue Code. 6108(a) mandates the publication, at a not less than annual basis, of statistics "including classification of taxpayers and their income, the amounts claimed as allowed as deductions, exemptions, and credits, and any other facts deemed pertinent and valuable." 6108(b) says that the Secretary of Treasury "may, upon written request by any party or parties, make special statistical studies and compilations involving return information…and furnish to such party or parties transcripts of any such special statistical study or compilation. 6108(c) says that any publication or disclosure authorized under 6108(a) or 6108(b) shall not disclose…or otherwise identify, directly or indirectly, a particular taxpayer.

One recent development is the Criteria Agreement signed into effect by the IRS and the Census Bureau in September 2000. This agreement was largely prompted by IRS concerns about whether a new form of confidential data access created by Census for outside researchers—the Research Data Centers—was legal. The agreement included the understanding that any new data use or access of FTI had to be authorized by an explicit dual-agency approval process involving both Census and the IRS and that the predominant purpose of the access, including for outside researchers, had to be the benefit of Census under Title 13, Chapter 5, so that statistical use was a necessary but not sufficient condition for authorized access. In addition, a similar and explicit dual-agency post-project certification of project results completes the authorization process.

The Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) provided uniform confidentiality standards and penalties across agencies, and enables sharing of business data among BEA, BLS, and Census. These agencies may grant access to outside researchers for statistical purposes only under the control and supervision of the agency, and subject to the same limitations and penalties as agency employees. Access sites might include the Census Research data centers. A companion bill (known as the "J bill" because it would amend IRC section 6103(j))) would add BLS to the list of agencies that the Internal Revenue Code authorizes access to tax data, extend BEA access to non-corporate business data, and provide uniform, validated data on the business population to BLS, BEA, and Census has stalled. The potential significance of the J bill is considerable, as the Census business register consists of FTI inextricably "commingled" with data covered under Title 13. Thus, any meaningful sharing of business data among the three agencies designated by CIPSEA must begin with the population of business data across legal ownership forms maintained by Census (probably including non-employer data, which are maintained on a separate file). Even the basic building blocks of data sharing (i.e., the entity data of employer identification numbers, etc.) are largely dependent upon annual infusions of population FTI.

In response to increased fears that micro tax return data sets might be becoming more vulnerable, the SOI has increased the degree to which disclosure protection methods such as subsampling and micro aggregation/blurring are applied to the annual cross-sectional income tax return data, and stopped providing a public-use version of the individual panel data after tax year 1990.

From time to time it has been suggested that confidentiality problems could be mitigated by releasing cross-product matrices rather than micro-data. The suggestion has little merit, as the number of studies that do not require (1) dropping certain taxpayers, such as the self-employed, or non-itemizers or (2) making non-linear combinations of variables, such as tax liability under an alternative regime or (3) creation of instruments such as first-dollar tax rates, all of which render pre-packaged cross-product matrices of limited use, is very small. These same difficulties apply to a lesser extent to most schemes designed to make data anonymous. Even a scheme that retains all the linear relations among the variables in the full sample (Sande 2001) does not retain the non-linear relationships, nor the relationships for selected samples. We note and welcome the work done by, e.g. Abowd and Woodcock (2001) on synthetic data—a sample composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data, but have not yet seen these methods applied successfully to standard tax-and-transfer related data sets. As stressed by Abowd and Lane (2003), such synthetic data sets can be used by researchers at a remote site to develop an understanding of the structure of

the confidential data, develop analysis code, and even estimate basic relationships before sending the code to the secure site to estimate the underlying relationships on the original confidential data.

6.Proposals

In this final section we collect some suggestions for initiatives that would address some of the problems we have already discussed, plus some that deal with other issues.

Data Centers
We welcome improved researcher access to existing datasets via institutions like the Census Research Data Centers (RDCs). Apparently matches of SIPP and CPS with Social Security Earnings Histories are available at the RDCs for researchers. This would be something of a breakthrough, since the Earnings Histories are IRS data, and SOI has always maintained that they were not available under any conditions of use.

Under current rules for authorized work (providing necessary Census benefits, etc.) an external researcher can bring in an outside microdata set (confidential or otherwise), match to confidential data, and leave with the original external dataset (Census would retain no copy). The Census-researcher work agreement would simply need to stipulate that, upon conclusion of the project, the matched internal/external data would be destroyed. Thus, Census would retain no identifiable data from the external dataset.

Under an ideal arrangement, a researcher could bring in a micro data set for matching to micro data from Census (and/or to micro data from BLS, BEA, SSA, etc.) for an authorized purpose, do so under Special Sworn Status, as an IPA (i.e., covered by the Intergovernmental Personnel Act Mobility Program, discussed below) or whatever satisfactorily ensures confidentiality, and then leave with the external micro data, which would not be retained in identifiable form on any database in the BRDC. The research purpose would, of course, have to be authorized by the agencies whose data are matched for this project. Agency mandate or mission would be key to authorization, in keeping with not only the Census-IRS Criteria Agreement, but the proposed J bill.

Such an arrangement might operate under the auspices of OMB/NSF/Census, or involve a combination of reputable non-government institutions along the lines of the sponsorship of Census RDCs by the NBER, FRB, and various universities. Under such an arrangement, it might be possible eventually for an outside researcher with tax-and-transfer interests to match an entire

SOI file to Census data, instead of just the limited items that are currently authorized, as long as the project purpose is demonstrably tied to an agency's mandate/mission.

Mediated Access to Tax Return Data

One possible solution to the disclosure problems concerning individual tax returns would be to automate the process of conducting statistical analysis on confidential data without any possibility of disclosure. One costly and labor-intensive way to do this, which has been used by some researchers, is to submit, e.g., SAS programs to the government agency, which assigns a staffer to run the program and examine the output to detect and prevent disclosure. We envision a biannual or annual process where researchers would submit proposals, and an outsider would rank them for merit. This process would be greatly facilitated by having a staffer who is dedicated to working on these research projects. Given the tight budget at statistical agencies, this is unlikely to happen without earmarked funding coming from the outside.   In addition to the funding issue, the success of such a program depends on the availability of dedicated, experienced staff. The selection process would also be important, e.g., that it be perceived as "politically neutral."

This procedure could include consideration of requests for release of limited sets of variables from samples of tax returns selected by social security number endings. For example, one might be able to get half a dozen variables for a special study, even if some of those variables were not in the public use file for confidentiality reasons. With so few variables in the file, confidentiality concerns would be significantly lessened. We note that SOI does make customized tables available on a fee basis to academics and accounting firms. The preparation of such tables is time-consuming, and it isn't clear that under the current system SOI fully recovers its costs.

Allowing Researchers Access to Data on a Temporary Government Employee Basis

A program to bring outstanding students and researchers into the agencies for a semester or year, to perform analytic work relevant to the agency function, would be valuable. Such a program would allow an academic to make a proposal and, if accepted, perform the work at the agency as a government employee could bring new methods to the agency, and knowledge to its employees. Such an assignment can (and has, in the past) occur under the auspices of the Intergovernmental Personnel Act Mobility Program (IPA), which provides for the temporary assignment of personnel between the Federal Government and state and local governments, colleges and universities, Indian tribal governments, federally funded research and development centers, and other eligible organizations. One problem is that current interpretation requires the IRS to pay IPA personnel on some fair market basis, thus making difficult mutually advantageous

arrangements at a (lower-than-market) rate of compensation that reflects the value of access to data and people familiar with the data. Outside funding to pay IPA personnel is a possibility, but one that may be have problematic legal aspects.

Validation Studies

It would be extremely useful to develop a systematic program to improve the quality of data available for analyses of taxation and program evaluation. The following recommendations use the problem of evaluating welfare reform as an example (for further detail, see Hotz and Scholz, 2002), although many of the same recommendations would apply to the analysis of taxation.

Several studies, including Moore, Stinson and Welniak (1997) and Roemer (2000) conclude that incomes, especially labor earnings, are measured well in the CPS and SIPP. But no study, to our knowledge, focuses on income reporting by disadvantaged, welfare-eligible, and/or welfare-prone populations. Moreover, participation in welfare programs is underreported in the CPS (and SIPP) and this underreporting appears to have increased over time. This is a troubling problem, especially as one looks to the future when TANF programs become state-specific, with different names. Given these trends, we think further work on the sources of antipoverty program underreporting and its origins in nationally representative survey data. This work also needs explore the degree to which income is understated for low-income families. The most promising way to do this is to match CPS (or SIPP) data to unemployment insurance data (or tax data for those who file) to examine the consistency of income reports across datasets.

Data from state unemployment insurance systems are a central source of information for evaluations of welfare reform. UI data are generally readily available and, given they are based on employer reports of payments, would appear to provide accurate information on an individual's earned income. It would be extremely useful to examine the coverage and trends in coverage of low-income populations with UI data (following Blakemore et al., 1996). Such an examination could be aided by using a match of UI data with respondents in a national survey, such as the SIPP, so that one could learn more about the demographic characteristics of individuals (and households) that report labor market earnings on a survey that are not recorded in UI wage records data.

States may be able to augment UI data used for evaluation of welfare reform by collecting supplemental information on the degree to which employers are designating workers as independent contractors. Additional work at the state level to assess the overall coverage of UI data would also be valuable.

More work is also needed to understand the extent to which UI wage data provide a misleading measure of the earnings available to low-income *households.* This problem arises in short- and long-term follow-up analyses of earnings for welfare samples drawn from state caseloads. One can use UI data to measure subsequent earnings for individuals that were in assistance units as long as they remain on welfare. However, it is difficult to accurately measure household income *after* assistance units leave the rolls because it is difficult to keep track of the identities of household members. Tax returns data may also be useful to learn more about whether the discrepancies between UI wage data and income measures from tax returns are the result of differences in family composition and the "composition" of income reported on tax returns.

To learn more about family well-being, it will be necessary to continue to rely on targeted, follow-up surveys to monitor samples of welfare leavers. Unfortunately, surveys are expensive. We recommend that a pilot study be undertaken to devise a survey that is designed to just get SSNs of other adults in a household, which can then be used to obtain UI wage earnings for these family members.

When examining transfer programs implemented at the state level (such as welfare), a major concern is that wage earnings from UI data are missed because individuals move out of state or that workers earn part of their income in other states. Again, comparisons of UI wage data with data from federal tax returns may help us to assess the importance of this problem and, more importantly, the biases that it imparts on measures of individual and household income.

Linking Demographic and Tax Data

There are two primary ways that scholars can have access to tax-related information with detailed demographic information. The Census Bureau regularly matches CPS data (and occasionally matches SIPP) data with federal tax returns. If the proposed project has an "essential tax administrative purpose," we believe provisions can be made that would allow tax researchers to have access to these data, on site at the Census Bureau.

The other opportunity is to improve the SIPP tax topical modules. Two actions are necessary to pursue this agenda. First, systematic study needs to be made of current SIPP tax topical module research to carefully assess its strengths and its limitations. The committee's preliminary examination of recent years indicates that nonresponse is a serious issue, but the distributions of adjusted gross income and filing status, conditional on responding to the questions is similar to the distributions found in the tax returns filed at the IRS. But much more needs to be done to assess the strengths, weaknesses, and potential of these data. Second, it would be exceptionally useful to have a commitment by the Census Bureau and IRS, in conjunction with academic

researchers, to systematically improve the SIPP tax topical module. Ideally, the tax topical module and core SIPP data would be matched to the tax returns filed in that year. The accuracy of tax topical module responses could be assessed making direct comparisons to information reported on tax returns. Equally important, systematic exploration could be conducted of the characteristics associated with non-response. The goal of this work would be to develop alternative weights that would allow the tax topical modules to be used for tax policy research. Because BLS has no confidentiality statute other than CIPSEA, it is worth considering whether it might be fruitful to match the CES to Title 13 data (including dependent information, earnings, etc.) as a project for the Census RDC process.

Archiving

Steps should be taken to ensure that data, once collected, are systematically archived somewhere rather than destroyed. Much valuable data that the government has collected at great expense have been destroyed (for example, historical data on individual estate tax returns and BLS surveys of employee benefits). Although in principle all data collected by the government are supposed to be archived in the National Archives, in practice this is rarely done. But it suggests that a legal structure is already in place and that greater enforcement of this rule could ensure that valuable data are maintained for the research community. The archival process must guarantee not only that the data are stored, but done so in such a manner that the media do not degenerate over time. For example, the IRS has hundreds or thousands of old tapes that are probably unreadable at this point.

Clearing House for State and Local Administrative Data

States collect a considerable amount of administrative data, including on UI, TANF, and Medicaid. A few places around the country are trying to match data from different sources to create interesting new data sources. The Poverty Institute at Wisconsin and Chapin Hall at Chicago are developing such data sources, and the Upjohn Institute for Employment Research has begun to establish an administrative data repository. Developing a Federal clearing house / coordination effort, perhaps along with resources, might help states be more willing to participate in these efforts and may make the data more available for research. The clearing house could also contain data collected from policy experiments. Depositing data could be made a requirement by funding agencies. A report by Hotz et al., 1998 made several worthwhile recommendations regarding fostering institutions to aid in the development and use of administrative data, establishing a centralized and ongoing repository for administrative data, and

developing guidelines and standards for administrative data to make them comparable across states and linkable within states.

The Role of the AEAStats Committee

A consistent theme of this report is the ubiquity of spillover effects in data collection and dissemination. Data collected for one purpose could often be expanded, at limited cost, in a way that is immensely valuable for other purposes. Data collected for one audience can often, at limited cost, be made available to other audiences. Often these spillovers are not considered, to the detriment of the potential contributions of data dissemination and analysis discussed in this report.

We suggest that the AEAStats Committee view one of its major roles as recognizing these spillover effects, and advocating and facilitating policies that would cause the spillover effects of data collection and dissemination to be appropriately internalized. Most of the suggestions we have made in this report are of this nature. We also recommend that the Committee maintain an ongoing, formalized conversation with officials at major statistical agencies regarding the status of data collection efforts and to regularly alert the academic community about these efforts. In addition, the Committee might serve as a clearing house for suggestions about acquiring additional information. Because some of the proposals we have made to internalize data spillovers require resources, we also suggest that the AEAStats committee is the natural organization to seek funding that would enable the kinds of data collection and dissemination efforts this report has detailed.

References

Abowd, John W. and Julia I. Lane (2003). "The Economics of Data Confidentiality." Mimeo, October 13.

Abowd, John W. and Simon D. Woodcock. (2001). "Disclosure Limitation in Longitudinal Linked Data." In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.) *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.*, Amsterdam: North-Holland.

Blakemore, Arthur E., Paul L. Burgess, Stuart A. Low, and Robert D. St. Louis (1996), "Employer Tax Evasion in the Unemployment Insurance Program," *Journal of Labor Economics*, 14(2), 210-230

Browning, Martin, Thomas F. Crossley and Guglielmo Weber (2003): "Asking Consumption Questions in General Purpose Surveys," *Economic Journal* 113, F540-F567.

Greenia, Nicholas H. (2004). "Developing Adoptable Disclosure Protection Techniques: Lessons Learned from U.S. Experience." *Privacy in Statistical Databases*, CASC Project Final Conference, Barcelona, Catalonia, Spain, June 9-11, Proceedings, Springer Press,

Greenia, Nicholas H. (2002). "The Release of IRS Data: Challenges and New Approaches." Paper presented at the 2002 IRS Research Conference. July 8.

Hotz, V. Joseph et al. (1998). "Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development," Joint Center for Poverty Research.

Hotz, V. Joseph and John Karl Scholz, (2002). "Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data," in *Studies of Welfare Populations: Data Collection and Research Issues*, National Research Council: National Academy Press, 275-315.

Hurd, Michael D. and Susann Rohwedder, (2005). "The Consumption and Activities Mail Survey: Description, Data Quality, and First Results on Life-Cycle Spending and Saving," mimeo, Rand, February.

Internal Revenue Service. (2002). *Tax Information Security Guidelines for Federal, State, and Local Agencies: Safeguards for Protecting Federal Tax Returns and Return Information*. Publication 1075. June.

Moore, Jeffrey C., Linda L. Stinson, and Edward J. Welniak Jr. (1997), "Income Measurement Error in Surveys: A Review," U.S. Bureau of the Census, Statistical Research Report.

Roemer, Marc, (2000). "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990-1996," mimeo, Income Statistics Branch, Bureau of the Census, June 16.

Sande, Gordon. (2001). "Methods for Data Direction Microaggregation in One or More Dimensions." Presented at the Federal Committee on Statistical Methodology Research Conference, November 15.

Strudler, Michael, H. Lock Oh, and Fritz Scheuren. (1986). "Protection of Taxpayers' Confidentiality with Respect to the Tax Model." *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 375-81.

Winglee, Marianne, Richard Valliant, Jay Clark, Yunhee Lee, Michael Weber, and Michael Strudler. (2002). "Assessing Disclosure Protection for a SOI Public Use File." Presented at the 2002 American Statistical Association meeting.