CAVEATS TO THE RESEARCH USE OF TAX-RETURN ADMINISTRATIVE DATA

Joel Slemrod

The expanded availability of tax-return administrative data, especially at the population level, has triggered an explosion of scholarly research addressing tax policy, tax administration, and non-tax questions. In this paper I discuss the tremendous promise of these data to inform tax policy and other economic questions, while reminding readers that use of the data is not without its issues. In what follows I offer a user’s guide that stresses the caveats a research user should bear in mind.

Keywords: administrative data, tax policy, data collection

JEL Codes: H20, H21

I. INTRODUCTION

The expanded availability of tax-return administrative data, especially at the population level, has triggered an explosion of scholarly research addressing tax policy, tax administration, and non-tax questions. Chetty (2015) lists 92 researchers who have used tax-return data as of 2015, including 55 people at academic institutions, compared to just four as of 2010, and documents that the percentage of publications at leading journals that use administrative data of all kinds has been rising steadily since 1990.

The wider availability of such data is an exciting development. In this paper I discuss its advantages, but remind readers that use of these data is not without its issues. In what follows I offer a user’s guide that stresses the caveats a research user should bear in mind.

II. WHAT ARE ADMINISTRATIVE DATA, WHAT ARE SURVEY DATA, AND HOW ARE THEY DIFFERENT?

Although it would be natural to begin with some definitions to clarify the issues at hand, this task is made difficult by the paucity of widely accepted definitions. Indeed, in an authoritative treatment of survey methodology, Groves (2004) admits at the outset that the word “survey” is not well-defined. In the literature (e.g., Rodgers, Brown, and
Duncan, 1993), one sees reference to “administrative records of survey data,” which obviously is troublesome for making clear distinctions. I don’t think this is a big issue for tax-return administrative data, as in this case one knows it when one sees it.

What survey and administrative data generally have in common is that at a crucial point some person writes down or enters data, which may be numeric or text. There are, though, two crucial differences. Administrative data are collected and maintained by an official government agency for some policy purpose. The American Statistical Association (1977, p. 60) 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.” Second, in some cases, there may be consequences, negative or positive, for misstatements of the truth in submitting administrative data, while there are no such consequences in completing a survey.

III. SURVEY DATA

There is a vast literature on survey methodology, occupying scores of volumes and articles. The origin of modern surveys can be traced to Charles Booth’s 1903 study of the London working classes. However, it was the 1930s that saw the birth of mass survey research, with companies such as Gallup and Roper surveying the public on upcoming elections as well as planned products or services. Since its beginning, the mode of data collection has also changed. Face-to-face or mail interviews were used at the outset, before telephone surveys became more common in the late 1960s, and, most recently, web-based surveys have become prominent.

One important issue is the accuracy of the data in measuring the underlying construct, or relationship, in which the researcher is interested. Groves (2004) considers four kinds of errors: (1) coverage error, (2) nonresponse error, (3) sampling error, and (4) measurement error from inaccuracies in responses recorded on survey instruments, due either to the effect of interviewers, respondents, the wording of survey questionnaires, or the mode of data collection.

Survey researchers worry about biased answers due to acquiescence (the tendency to agree) and social desirability (the tendency to present oneself in a favorable light). They also worry about response-order effects. On web-based surveys when the respondent is sitting in front of a screen rather than in front of an interviewer, visual syntax becomes an issue: the layout of the question and answer elements on the screen is known to affect the answers provided. Clearly the relevance and importance of all of these effects is context-specific.

Survey researchers often tout that they learn over time and improve the questions, but this raises the issue of the comparability over time of the data collected. Furthermore, if members are participating in a longitudinal survey, there may be time-in-sample bias:

---

1 I say “generally” because, in a business-tax-return setting, there may be automatic transfer of information from accounting systems to tax reports.
given their experience with the survey over time, responses may begin to differ from the responses given by people answering the same survey for the first time (Couper, 2000).

Survey researchers are aware that nonresponse is systematically related to certain demographics; for example, it is usually larger for people living in urban areas and among males. Meyer, Mok, and Sullivan (2015) caution that household surveys have suffered recently not only from rising rates of unit nonresponse, but also increased item nonresponse and measurement error. As web surveys have proliferated, researchers are concerned with the special kinds of bias introduced by the nonrandom penetration of computer access, the so-called “digital divide” (e.g., Couper et al., 2007).

IV. ADMINISTRATIVE DATA IN LABOR ECONOMICS

Both survey and administrative data have long been studied in labor economics, and the relative strengths and weaknesses of the two sources have been extensively discussed and documented. The administrative data studied include state unemployment insurance (UI) wage records, Social Security Administration records, and, very occasionally, income tax returns.

For example, Hotz and Scholz (2002) compare income and employment data of the low-income population as measured in national surveys and three kinds of administrative sources: state UI wage records, Social Security Administration records, and income tax returns. They stress that neither surveys nor administrative sources provide the true values for any outcome at the individual level, and focus on the relative differences in outcome measures across data sources. They note several dimensions along which survey and administrative records may differ: (1) population coverage, (2) definition of reporting units, (3) sources of income, (4) inadvertent measurement error, and (5) incentives associated with data gathering mechanisms.

Hotz and Scholz (2002, p. 283) stress that “there is little or no ‘cost’ to [survey] respondents of misreporting of income, employment, or other circumstances.” One common finding in this literature (e.g., Moore, Stinson, and Welniak, 2000; Roemer, 2000) is that there is underreporting of many types of income in surveys, with complex underlying reasons. Kornfeld and Bloom (1999) note that, while employers have an incentive to underreport earnings to the UI system (and hence avoid being subject to UI taxes), they have no incentive to conceal earnings when reporting to the Internal Revenue Service (IRS), because wages are a business expense that will generally lower tax liability. Most recently, Hurst, Li, and Pugsley (2014) argue that there is substantial willful understatement of income in surveys, given there is no cost to misstatement but some positive chance that income reports will be checked against tax returns.

There is also evidence that some other types of administrative data are not immune to misreporting. For example, Blank, Charles, and Sallee (2009) compare administrative data from Vital Statistic records to data from the U.S. Census and find that, in states where the minimum age of marriage laws are binding, younger individuals appear to have lied about their age to government officials when applying for their marriage license.
V. INCOME-TAX ADMINISTRATIVE DATA

I turn now to consider administrative tax-return data. I will focus on the U.S. income tax, mostly on the personal income tax, and will consider both the returns themselves (Form 1040, etc.) as well as the information reports (Form 1099s, Form W-2, etc.). The ability to match up the returns with the information reports is particularly valuable in many contexts, as this provides detailed information on, for example, sources and payers of income. The information on the Form W-2 wage and tax statement filed by employers can also, in principle, be matched to other data sources, even for non-filers. I will also focus on the use of tax data to study tax-policy-related questions, although the returns can be used as a source of longitudinal data to study many other important issues such as geographic mobility, geographic dispersion in income mobility, and the return to high-quality elementary education. Such studies reflect the fact that in the United States there is no real-time list of names and addresses of all residents (a register), so the file of personal tax returns is as close as we get to an annual register.

Let me clarify which tax-return data I will be referring to. First of all, the Statistics of Income (SOI) Division of the Internal Revenue Service maintains a structured mechanism for transforming administrative data into statistical files, using its own data collection systems that are completely autonomous of the main IRS tax-return processing. 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. Since 1960, the IRS has made publicly available to researchers a random sample of anonymized individual income tax returns that have been edited by SOI. These days the sample is stratified by income, the presence of a large amount of business receipts, the presence or absence of special forms or schedules such as Schedule C and F, and the potential usefulness of the return for tax policy modeling. For the 2014 file, out of a total of 147,759,485 returns filed, 332,040 were selected for the sample. There is also a public-use panel data (but not stratified) version of these data, but it is now quite outdated, spanning tax years 1979 to 1990.

The signal development of the last few years is that selected researchers have had access to a population database of unedited tax returns (sometimes referred to as the Compliance Data Warehouse) that includes individual income tax returns, business income tax returns, information returns, and other relevant forms including records of

---

2 Similar data are available to selected researchers in Nordic countries. Because there is certainly a systematic relationship between some country characteristics (e.g., attitudes toward privacy, trust in government) and data availability, we must be especially cautious about generalizing the conclusions to countries that differ on key dimensions.

3 The SOI also constructs and maintains scores of specialized data sets based on tax-return information. Johnson and Rib (2015) refer to the fact that SOI is currently conducting approximately 110 different projects involving data collection from returns and information documents.

4 This description is drawn from Johnson and Rib (2015).
Caveats to the Research Use of Tax-Return Administrative Data

births and deaths from the Social Security Administration. These data are “rawer,” as they do not go through the checking procedures of the SOI program, so at the moment there is a tradeoff between accuracy and sample size and scope. The data are accessible to researchers via application to the Joint Statistical Research Program of the SOI Division of IRS, whose objective is to increase use of its tax microdata by researchers outside the Federal government to “advance the understanding of how existing taxes affect people, businesses, and the economy, and provide new understanding of taxpayer behavior that can aid in the administration of the U.S. tax system.” In the latest round of this program in 2014, over 80 applications were received and 12 projects were approved; the 27 researchers from the approved projects span 14 different academic institutions, two government agencies, and a non-profit policy shop. Raj Chetty, Emmanuel Saez, and their colleagues, especially John Friedman and Danny Yagan, have organized much of the population data into an “SOI Databank,” which provides a more research-ready structure by pulling information from many disparate population data sets and organizing it into an accessible panel structure.

When the population of returns is available, the notion of sampling error disappears, or at least changes. One might think that, if the entire population is observed, the standard errors of estimates should be zero, although researchers continue to report non-zero standard errors in such settings. Abadie et al. (2014) explore the meaning of standard errors for a model intended to capture causal effects and suggest that, even if outcomes are observed for all units of a population, there still exist missing potential outcomes for each unit for the treatment levels to which they were not exposed. In other words, a potential outcome framework allows for uncertainty from conditional random assignment. There is also no issue of social desirability with tax returns, except perhaps in personal interactions with tax professionals, at least as long as the United States does not have public disclosure of income tax information.

The issues of coverage, nonresponse, and measurement error raised by Groves (2004) in connection with survey data still remain, however. There are two separate accuracy-related issues. The first is whether the data available to researchers accurately represent what was on the tax form as filed, what I will call transcription inaccuracies. The second is whether the reports on tax returns accurately represent the underlying construct of interest, known as the issue of validity in the survey methodology literature. Some of these issues are less important for the SOI data, due to the error-correction processes described above.

A. Transcription Inaccuracies

Even for raw e-filed data, data processing errors occur. Templates might change and not be correctly considered. When working with third-party provided documents, such as Forms W-2 or 1099, there can be duplicate filings that must be sorted out to avoid

---

6 The 2014 round was the second, the first having occurred in 2012.
double counting as well as amended forms. Some of the information on a tax return, such as occupation and address, is text, and the coding of the text can misrepresent what the taxpayer intended, and what the taxpayer wrote may be garbled. It may be that the standard editing rules sometimes cause the elimination or adjustment of a “correct” value, known as “adjustment error” in the survey methodology literature. Finally, any of these data sets must address the issue of amended returns and returns filed after the “processing” year of the data. Including the late filings with the subsequent tax year, as the SOI file does, is a particular issue when the tax code changes, as some (usually small) fraction of the returns will have been filed under a different tax law, muddying the identification of behavioral responses. The extent of transcription inaccuracies may vary depending on whether the return is filed electronically or not; the share of each is changing over time and, within a given year, the method of filing is likely correlated with observed and unobserved characteristics of the taxpayer.

B. Validity: Laziness and Sloppiness

Some taxpayers are lazy, some keep poor records, and others are sloppy. Some supply only a bottom-line number and attach a spreadsheet, PDF, or a handwritten balance sheet instead of the line-item detail the tax form requests. Even some corporate tax returns are filed with the front page blank, with a “see work documents” note. Others group a number of unrelated items together and report them on a line labeled “other.” More generally, those in the know say that the quality of the data provided on supporting schedules is on average much lower than the values directly used to calculate tax liability.

Taxpayers make “honest” mistakes. A recent survey of businesses revealed that more than a quarter of firms “struggle” with manual, incorrect tax data entry (Bloomberg BNA, 2015). Sometimes the tax code and/or instructions allow taxpayers to legitimately make choices as to how information is reported. For example, why bother reporting an income item that is exactly offset by an expense item, such as reimbursements for travel? There is evidence that, under stress, Schedule C businesses pad “other expenses.” Data items used primarily as background information, for example an address or occupation, may be of particularly low quality or even incomplete, as there is no incentive for accuracy and some taxpayers may perceive an incentive for misdirection. Although not a transcription error, note that taxpayers may file tax returns using their business or tax preparer’s address, instead of their home address. Notably, if the issue is just laziness or sloppiness, errors should be symmetric around zero unless, for example, more effort is

---

7 The IRS has the right to assign additional charges or penalties when a person files incomplete or incorrect tax paperwork, but it usually only does so when the mistake leaves the individual owing money.

8 This makes it difficult to run these returns through the standard discriminant formula that affects the chance of audit, although the IRS could (and maybe does — the formula is a secret) add “lack of detail” as an input to the DIF formula, and could even announce this.

9 Both of these phenomena arise in Slemrod et al. (2015), which is an evaluation of the impact of the new 1099-K form that requires credit-card companies to report on the receipts of their customers.
required for documenting tax additions compared to tax subtractions, and the variation would be higher for items that require more effort.

C. Validity: Rational Overstatement of Tax Liability

I will get to evasion in a moment. But, before I do, note that some taxpayers file returns that overstate their liability — they do not minimize (legal) tax liability. For example, they do not bother claiming every credit or deduction for which they are eligible. This behavior could be due to lack of knowledge, for which there is plenty of evidence. For example, a survey conducted by the National Public Radio, Kaiser Family Foundation, and the Harvard Kennedy School (2003) that addresses taxpayer knowledge generally find that 28 percent of respondents did not know whether they were eligible for the Earned Income Tax Credit, and 5 percent of those that were eligible did not take advantage of it.

Alternatively, a taxpayer may be aware of the tax law but optimally decide to forego some tax savings because of some cost of so doing. Pitt and Slemrod (1989) and Benzarti (2015) find that substantial compliance costs of itemizing deductions on individual tax returns keep many people from itemizing who would save taxes by doing so. Benzarti (2015) finds that the burden of tax filing is larger for richer households, consistent with the fact that the value of time increases with income. It may be that some taxpayers do not want to draw IRS attention, so they estimate things like the value of donated items in a very conservative way: how conservatively might depend on noncompliance on other parts of the tax return. This phenomenon does not lead to mismeasurement of taxable income, but would lead to errors in calculating, for example, net income from employment or the distribution of medical expenses.

Bhargava and Manoli (2015) examine the determinants of incomplete take-up with the Earned Income Tax Credit (EITC) in the United States using a randomized controlled trial (RCT) of alternative mailings that vary the role of program information (regarding benefits, costs, and rules), informational complexity, and stigma. They find that simply receiving a mailing increases the take-up rate, suggesting that routine contact from the tax authority can have a significant effect on taxpayer behavior, at least in the short run. In addition, both simplification and the visual display of benefits increase take-up. Notably, a follow-up study by Manoli and Turner (2014) found that such informational interventions have little to no effect in the long-term. Guyton et al. (2016) extend this work to non-filers using an RCT to induce filing among non-filers who are eligible to receive credits, even if they owe tax on net. They find similar results: there is a concurrent effect, but one that does not persist in future years when the mailed reminders stop.

Many people do not have to file an income tax return, a key reason why a file of taxpayers does not match the population at large. The filing threshold changes over time and sometimes there are special reasons to file — such as to claim the 2008 stimulus payment — so that the importance of this issue also changes over time. The IRS is currently engaged in a project comparing tax information documents to Census data to learn more about the non-filing population.
D. Question Ordering

As discussed, the order and precise wording of questions is known to matter in surveys. One approach taken in survey research, randomizing question order, is in principle feasible in tax returns but is not likely to be pursued extensively, or at all. Note, though, that while tax returns come in a particular order, the taxpayer (or preparer) need not complete the return in that order, unlike in an interview survey. The ability to enter information in any sequence is present in some, but not all, tax software. Software could be designed to impose a particular sequence of issues to be addressed.

Changes in the filing process could affect the validity of the information provided on the returns. Perhaps the most striking example of this was the introduction as of tax year 1986 of the requirement that taxpayers who claim dependents must list their children’s Social Security numbers, after which seven million dependents “vanished” from the tax rolls, some incalculable combination of real pets and phantom children.

E. Willful Noncompliance

The key distinguishing characteristics of tax-return administrative data, especially compared to survey data, are that what one reports generally affects tax liability and, relatedly, there may be negative consequences for detected understatements of tax-return data. The penalty for willfully attempting to evade tax is either up to five years’ imprisonment, a maximum fine of $250,000 ($500,000 for corporations), or both, plus the costs of prosecution. Of course, these penalties are only applied if the noncompliance is detected (and the IRS claim is upheld through an appeal process), and the chance of such detection is the linchpin of the deterrent effect of enforcement, which remains the leading positive model of tax evasion. According to the latest IRS tax gap estimates (for tax year 2006), the gross tax gap is $450 billion, or 16.9 percent of the estimated actual (remitted plus unremitted) tax liability. Notably, the noncompliance rate varies greatly by the source of information reporting. When there is little to no third-party-reported information (such as self-employment income), the IRS estimates the noncompliance rate to be 56 percent, compared to 1 percent when there is both withholding and substantial reporting (such as for wages and salaries). Most, but not all, line items on a tax return have implications for taxable income and tax liability, so that evasion will in general affect all of the line items. An especially problematic kind of willful noncompliance is the failure to file a return at all when tax is due.

Examining post-audit returns can in principle address some of these problems. The micro data from the stratified random audits of the National Research Program, used

---

10 Bankman, Nass, and Slemrod (forthcoming) discuss a different perspective on this issue, exploring whether question ordering and wording can be purposely chosen to reduce noncompliance. For example, they discuss moving the attestation to the beginning of the form, and adding a yes-or-no question about cash receipts. See Tourangeau, Couper, and Conrad (2004), who conduct experiments to investigate the impact of visual features of survey questions. They find that varying the spacing of both substantive and non-substantive answers can affect a respondent’s perception of the midpoint of the scale.
to estimate the tax gap, are available, although not publicly. But even these data do not perfectly capture and correct for noncompliance, as indicated by the fact that in calculating the aggregate “tax gap” the IRS multiplies the noncompliance found by the NRP auditors by a factor of between two and four.

F. Constructing a Panel

One of the tremendous accomplishments of the IRS Databank is the linking of the tax returns of taxpayers and their dependents over time. This is by no means a trivial task, as it must deal with issues such as the ordering of names on a joint return, marriage and divorce, as well as tracking dependents after they are no longer claimed as dependents by their parents or guardians.

G. Constructing a Time Series

Tax-return data feature information from tax forms, and what is required on these forms changes occasionally, and can change a lot in a major tax reform. The best example is the Tax Reform Act of 1986, which increased the standard deduction, eliminated the deductions for personal interest payments and sales taxes, subjected miscellaneous deductions to a floor, and increased the floor on deductible medical expenses, all of which reduced the fraction of taxpayers who itemized their deductions. Furthermore, the increase in the threshold income for filing reduced the number of taxpayers who needed to file at all.\(^\text{11}\) This is obviously a problem when constructing a time series for a line item that need not be reported in some years. It is also a problem, and a potentially big one, when that item is a component of an income measure. The researcher’s conceptual definition of income may not change, but the tax code’s definition of income can and does change occasionally, and when it does so it may be impossible to construct a time series with a consistent definition.

As an example of how this matters, recall that Feenberg and Poterba (1993) used tax-return data from 1951 to 1990 to investigate the rising share of AGI reported on very high-income tax returns. They find that most of this increase is due to a rise in reported income of the top one quarter of 1 percent of taxpayers, and attribute at least part of this increase to the change in tax incentives that resulted from the 1986 Tax Reform Act (TRA86), namely the reduction in the top marginal tax rate from 50 percent to 28 percent. However, Slemrod (1996) noted that the change in how capital gains appears in adjusted gross income increases the apparent concentration of income because capital gains are heavily concentrated among high-income households and because it changes the ranking of taxpayers, moving those households whose income consists of a relatively high share of capital gains towards the top of the income distribution. Indeed, Slemrod

\(^{11}\) More recent changes to the definition of taxable income have included the Mortgage Forgiveness Debt Relief Act of 2007, which excluded debt forgiven on a principal residence from taxable income through 2009, subsequently extended to 2012.
shows that the 3.4 percentage point increase (from 7.4 percent to 10.8 percent) in the share of AGI going to the top 0.5 percent of income earners found by Feenberg and Poterba in fact falls to a 1.9 or 2.7 percentage point increase when a consistent 1990 or 1984 definition of income is used, respectively.

H. Not Much Demographic Data

At least until tax-return data can be linked to other administrative data such as Census or other survey data such as the Health and Retirement Study, income-tax-return data has little demographic information, and none regarding race, educational attainment, labor supply, and employment details including hours worked or the within-year timing of earnings, health status, disaggregated expenses, housing details, or wealth. Information on health insurance coverage is available only beginning in tax year 2015, and information on college tuition and therefore matriculation is available since 1997. Because in the United States married persons generally file a joint income tax return, the data generally contain no person-by-person breakdowns of income and sub-items; this is not an issue in most other countries, where filing is on an individual basis (although intra-family shifting issues may arise). Many of these lacunae could be addressed if survey data could be matched with administrative data. In the meantime, one must be concerned with bias introduced in analysis where relevant demographic determinants are excluded and the difficulty of constructing treatment and control groups that are balanced on a limited set of observable variables.

VI. WHAT IS IT (PARTICULARLY) GOOD FOR?

A. The Elasticity of Taxable Income

Tax-return data is particularly good for studying the elasticity of taxable income (ETI) with respect to the tax rate because, for this question, reported income — be it taxable income or a broader concept such as adjusted gross income — is precisely the variable of interest, warts and all. For this, transcription errors and such remain a problem, but not evasion, as long as the eventual collection of revenue from enforcement activities and shifting across tax bases is properly accounted for. When the definition of taxable income changes, the researcher loses the opportunity to measure the ETI cleanly. Estimating the ETI using data that span periods with different tax bases, or enforcement regimes, is also problematic because the base or enforcement changes will generally affect the ETI itself (narrower bases generate higher elasticities), and standard regression methods will generate a weighted average of the period-specific ETIs over the period, where the correct weights are not straightforward.13

12 For a review of the ETI, see Saez, Slemrod, and Giertz (2012).

B. Behavioral Response of Line Items

Data from income tax returns have been widely analyzed to estimate, inter alia, the behavioral responsiveness of particular line items, including donations to the presidential election campaigns, home mortgage interest, and — perhaps most often — charitable contributions. Most of the caveats already discussed apply. For example, in the case of charitable contributions, the tax return features donations as reported, which may be overstated to reduce tax liability, understated because of the hassle of keeping records, or poorly measured due to sloppiness of record-keeping and the difficulty of valuing noncash donations. Moreover, the data are available only for itemizers, the choice of which is endogenous to a taxpayer’s charitability. Slemrod (1989) explores the role of evasion in charitable contributions in estimating a tax-price elasticity by estimating a then-standard log-linear regression model of charitable contributions on a single year of individual tax returns, once for donations as reported and then separately for donations as corrected by an auditor from the Taxpayer Compliance Measurement Program. The results suggested that the estimated net-of-tax price elasticity of reported contributions was not principally an evasion elasticity; indeed, purging the reported data of overstatements increases the estimated price responsiveness of charitable giving.

VII. WHAT IS IT NOT SO GOOD FOR?

Given the vastly different, and not obviously unchanging, compliance rates between employees and the self-employed, making comparisons across these two groups is extremely problematic. Nor will simply adjusting reported self-employment income by an estimated average compliance rate (i.e., dividing by 0.44, given that the IRS estimates that the noncompliance rate is 56 percent) solve the problem, as there are differences in noncompliance among self-employed occupations. Erard and Ho (2003) analyze post-audit tax-return data to assess the differences in compliance across 34 distinct occupation groups, accounting for both non-filing and misreporting.14 Ranking the occupations in order of the average dollar level of noncompliance, they find that the five least compliant occupations are: (1) vehicle sales, (2) investors, (3) informal suppliers, (4) lawyers and judges, and (5) doctors and dentists.

VIII. WORK IN PROGRESS

A. The Distribution of Income

Tax-return data have been a major source of information for the foundational studies of the extent and trends in income inequality, of which the work of Piketty and Saez (e.g., Piketty and Saez, 2003) stands out. But how well does the distribution of reported income reflect the distribution of actual income? One disconnect is disproportionate tax noncompliance across income groups. Johns and Slemrod (2010) examine the distribu-

14 Recall, however, the possible inaccuracy of self-reported occupation, discussed above.
tional consequences of income tax noncompliance in the U.S. federal income tax for the tax year 2001. They find that, when taxpayers are ordered by their estimated “true income,” defined as reported income adjusted for the underreporting estimated by the IRS tax gap methodology, the ratio of aggregate misreported income to true income generally increases with income, peaking among taxpayers with adjusted gross income in the 99.0 to 99.5 percentile; this implies that income inequality is under-estimated by reported income. In contrast, the ratio of underreported tax to true tax is highest for the lowest-income taxpayers, reflecting the fact that a given percentage reduction in taxable income corresponds to a particularly high percentage reduction in tax liability for taxpayers with taxable income just above the taxpaying threshold. Much of the distributional pattern of noncompliance is associated with the fact that, on average, high-income taxpayers receive their income in forms that have higher noncompliance rates — from self-employment and small businesses.

Another issue is that the tax concepts of income do not accord to the standard Haig-Simons definition of income. Taxable income contains realized nominal capital gains rather than real accrued gains, it does not include the rental value of owner-occupied housing, and excludes employee compensation in the form of medical insurance, to name just a few differences. Capital gains are especially important when assessing the income and trends in income of the very rich, for whom capital gains comprises a large fraction of reported income.

Finally, one must keep in mind that many, especially low-income, people do not file tax returns. Of 161 million filing units in 2012, only 145 million file a tax return.\(^{15}\)

### B. Capital versus Labor Income

Using tax-return data to decompose the level and trend of income inequality by the type of income — labor versus business versus capital income — is hampered by the somewhat arbitrary categories set out by the tax code, including the realization-based taxation of capital gains. For example, before 1986 some individuals used Subchapter C corporations to allow them to pay a lower rate of tax than the individual income tax. This incentive was virtually eliminated by TRA86, so many individuals converted their Subchapter C corporations to Subchapter S, meaning that previously excluded corporate income appeared on personal tax returns.

Income sources such as carried interest and the sale of founders’ stock arguably represent a return to labor but are taxed as, and therefore reported on tax returns as, capital gains. What appears on tax returns as dividends includes not just portfolio dividends, but also “dividend recapitalizations” — when a private equity fund borrows in order to pay a special dividend to private investors or shareholders. Retained earnings of corporations do not appear explicitly anywhere on the personal tax returns of shareholders. Indeed, Fleischer (forthcoming) argues that gains from the sale of founders’ stock and partner-

\(^{15}\) These figures are from Emmanuel Saez’s website, in the regularly updated file http://eml.berkeley.edu/~saez/TabFig2014prel.xls.
ship allocations of capital gains to fund managers account for a very large share of the recent top-end increase in income inequality in the United States and that a significant portion of reported capital gains includes carried interest and income from the sale of founders’ stock. To be sure, the tax characterization of forms of income into distinct, and not conceptually coherent, categories has important implications for the effective tax rate on such income and for the incentives to shift income across these categories, but it limits how much we can learn from tax-return data about the economic forces underlying the trends in income concentration.

C. The Distribution of Wealth

Tax-return data figure prominently in the two major non-survey methods to assess wealth and wealth inequality. The first is the estate multiplier method, under which each estate observation (drawn from estate-tax-return data) is inflated by a multiplier equal to the inverse probability of death estimated from mortality tables by age and gender for each year, often then multiplied by a social differential mortality factor to reflect the fact that the wealthy (those who file estate tax returns) have lower mortality rates than average. In this case, the issue is the accuracy of the estate tax returns rather than income tax returns, and evidence on estate tax noncompliance and accuracy is sparse, as discussed in the literature (Eller, Erard, and Ho, 2001).

The other method capitalizes flows of capital income to derive estimates of wealth. In a prominent recent example, Saez and Zucman (forthcoming) track the change in wealth inequality in the United States since 1913 using the capitalization method based on reported income from individual tax returns. The authors recognize that tax evasion can affect the capitalization method, causing an overestimate of wealth if individuals report labor income as capital income (e.g., through “carried interest” or certain stock options) or an underestimate in the presence of shifting wealth out of the tax base entirely, such as by moving it to tax-haven accounts. The comments above about the difficulties in distinguishing capital from labor income apply in this case as well.

Bricker et al. (2016) note that the levels and trends in income and wealth inequality observed in household survey data are less dramatic than those derived from administrative income tax data, and attempt to reconcile the different findings. They note that, compared to Saez and Zucman, the Survey of Consumer Finances (SCF) shows less than half the increase in the top 1 percent wealth share between 1992 and 2013, rising from 30 to 36 percent instead of from 29 to 42 percent, and argue that the SCF uses administrative data to select the sample and rigorous targeting to ensure that wealthy families are properly represented in the survey data.

One aspect of this ongoing debate reveals a limitation of tax-return data to this point. Saez and Zucman use a uniform capitalization rate to translate the observed flow of interest receipts to an estimated stock of interest-bearing assets. This method will

---

16 See Kopczuk and Saez (2004) for more details on how this method is implemented.
17 These issues are discussed in Kopczuk (2015).
overstate wealth concentration if lower-income people tend to hold low-yielding bank accounts while high-income people hold higher-yielding assets. Bricker et al. (2016) suggest that a market-based rate of return method generates a capitalization rate of 25, while the Saez-Zucman method based on estate-tax filings generates a factor of 100, and the difference between these two capitalization rates explains most of the difference in the recent estimated growth in the top wealth shares between the Saez-Zucman (forthcoming) methods and the SCF-based calculation. Alas, tax-return administrative data, even the 1099-INT forms filed annually by payers of interest income such as banks and savings institutions, do not allow a researcher to identify asset type.

IX. NEW FRONTIERS

Given that researchers have only just begun to exploit the potential of population tax-return administrative data, it may be a bit premature to speculate about even newer frontiers involving tax-related data. But if not here, where? What follows is an unordered list of potential data sources the analysis of which could be very insightful: (1) data from other taxes, linked among taxes, (2) data on payments of undisputed tax debts, (3) data on withholding taxes remitted by, e.g., employers, linked to individual tax returns, (4) data on searches of tax returns in (Nordic) countries where tax-return data are publicly available, and (5) data on enforcement efforts and consequences. In some cases, analysis of such sources has already begun.

X. CONCERNS: ANONYMITY AND REPLICAIBILITY

The potential benefits of private researchers’ access to tax-return administrative data must be weighed against its potential costs. Two such costs are worth mentioning. The first is the threat to the privacy of tax returns. The returns in the public-use sample of personal tax returns are stripped of direct identifying information, and the data are “blurred” to make it very difficult to identify the taxpayer even if the data are matched to other publicly available information. The “raw” data are anonymized but not blurred. All users explicitly agree not to attempt to identify any individuals on the data file, and are subject to the same kind of sanctions for violation of this as are the IRS employees who can access the data. Safeguarding the privacy of returns becomes more difficult as population data become more widely available, and so doing is critical to ensure the continuation of improved data accessibility.

Finally, the research use of tax-return (and other) administrative data raises important questions of replicability. Replicability is central to scientific credibility and progress, and now most economics journals require that the authors of published papers post their programs and data,18 but the users of the tax-return data are prohibited from such public

18 For the American Economic Association policy, see “Data Availability Policy,” American Economic Association, https://www.aeaweb.org/journals/policies/data-availability-policy. Note that, while some proprietary data has been accessed by only one or two groups of researchers, as mentioned above, by now nearly 100 researchers have done research using the tax-return data, so the feasibility of replication is considerably higher.
XI. CONCLUSIONS

For over half a century, researchers outside the U.S. Treasury and IRS have had access to anonymized samples of income tax returns, and have learned much about the tax system by analyzing them. Recently, however, nearly all the income-tax-return information the IRS receives has become available on a limited basis, vastly increasing the sample size and the range of information that can be analyzed. Continued analysis of these data both within and outside of the government promises to offer profound insights into the tax system and its role in important economic phenomena such as inequality and growth. I offer this essay as a user’s guide, both to researchers and consumers of the research, who, while working to unlock the insights the data reveal, should keep in mind the caveats that apply.

ACKNOWLEDGMENTS AND DISCLAIMERS

This paper was prepared for the 2016 National Tax Association Spring Symposium held in Washington, DC on May 12–13, 2016. The paper has benefited greatly from the research assistance of Jennifer Mayo and from comments on an earlier draft provided by Anne Brockmeyer, Charles Brown, Raj Chetty, Adam Cole, John Guyton, Barry Johnson, Aparna Mathur, Daniel Reck, Emmanuel Saez, John Sabelhaus, and Michael Weber; any errors are solely my own responsibility. The views expressed here are my own, and are not necessarily held by the IRS or the U.S. Treasury.

DISCLOSURES

The author has an unpaid contractual relationship with the Internal Revenue Service to make research use of anonymized tax-return data. The author has no financial arrangement that might give rise to conflicts of interest with respect to the research reported in this paper.

REFERENCES


