The Administrative Income Statistics (AIS) Project: Research on the Use of Administrative Records to Improve Income and Resource Estimates*

Adam Bee
U.S. Census Bureau

Jonathan Rothbaum
U.S. Census Bureau

SEHSD Working Paper Number 2019-36

October 2, 2019

Abstract

There has been considerable interest within the U.S. Census Bureau and in the research and policy communities in measurement error and mis-reporting of income on surveys. One possible avenue to correct for these issues is to use administrative data to replace or complement survey responses. However, combining survey and administrative data is not without its challenges. In this paper, we review the literature on how household survey estimates of various income statistics may be biased by issues such as mis-reporting and increasing non-response. We describe the administrative data currently available for research use at the U.S. Census Bureau and the data available to other federal agencies that we could potentially use in the future to improve income statistics. We detail the implementation opportunities and challenges to using administrative data. Finally, we outline our research agenda to address these challenges.

Keywords: Income statistics; administrative records.
JEL Codes: I3, D31

*This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau. This paper has been cleared by the Census Bureau's Disclosure Review Board release authorization number CBDRB-FY19-POP001-001.

1 Introduction

Historically, survey data have been the primary source of information on the income distribution and a major input into academic research on individuals and households in the social sciences. Recently, however, administrative data have also become a major source of information for income statistics and economic research (Chetty 2012).

This increasing attention to administrative data has changed our understanding of income survey data. There has been a considerable amount of research at the Census Bureau and in the broader academic community over the last several decades comparing survey and administrative income reports. At Census this work goes as far back as links between Internal Revenue Service (IRS) data and the 1964 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and, with data from the Social Security Administration (SSA), the 1973 CPS-IRS-SSA Exact Match Study (Kilss and Scheuren 1978).[1]

This research has found several issues with survey income reports. First, survey reports can suffer from mis-reporting, with many income sources subject to severe under-reporting, although administrative data can also contain measurement error. Second, growing non-response and shortcomings in income imputation models may result in increasing bias in income estimates. Third, some work on inequality, using only administrative data, has produced estimates that validate concerns that surveys may have limited coverage at the top of the distribution. As a result of these issues, researchers have developed various methods to use administrative data to improve estimates of income statistics. This includes work on administrative data linked to survey data, work using administrative records alone,

[1]For recent examples of work using linked survey and administrative data at Census, see Abowd and Stinson (2013), Bee (2013), Benedetto et al. (2013), Harris (2014), Bee et al. (2015), Giefer et al. (2015), Hokayem et al. (2015), Bhaskar et al. (2016), Chenevert et al. (2016), Noon et al. (2016), Bee and Mitchell (2017), Fox et al. (2017), O'Hara et al. (2017), Abowd et al. (2018), Abraham et al. (2018), Benedetto et al. (2018), Bhaskar et al. (2018), Bollinger et al. (2018), Brummet et al. (2018), Eggleston and Reeder (2018), Meyer and Wu (2018), Murray-Close and Heggeness (2018), Rothbaum (2018), Shantz and Fox (2018), Bee et al. (2019), Eggleston and Westra (2019), Fox et al. (2019), Hokayem et al. (2019), Imboden et al. (2019), Jones and Ziliak (2019), Meyer et al. (2019), and others. Recently, Census has embarked on a project titled the Comprehensive Income Database (CID, see Medalia et al. 2019) to extend and complement this work. Meyer and Wu (2018) and Meyer et al. (2019) are results of the CID project.

and work using summary statistics from unlinked administrative data to adjust survey-based estimates.

In this paper, we briefly summarize that work and discuss the opportunities and challenges faced by the Census Bureau in combining administrative and household survey data to improve our estimates of income, poverty, and well-being, and to better understand the economic circumstances of the nation's people and households. We believe that we can use the strengths of both types of data to produce estimates and statistics that are superior to what could be done with either separately. For example, administrative data generally do not include information on other demographic and socioeconomic characteristics, such as race, ethnicity, education, occupation, industry, etc., that are available in survey data. Also, some sources of administrative data have potential issues with income under-reporting, which surveys may help address. These include unreported wage and salary earnings (tips, "under-the-table" income) and self-employment income, inter-household transfers, etc. Furthermore, survey data are constructed from random samples of a target population, whereas in most cases administrative data are drawn from a non-random sample, such as tax filers or program participants.

We envision two primary ways administrative data can be used to improve survey data. First, they can be used to replace survey information on particular topics or to augment currently collected surveys. This may allow Census to reduce respondent burden by removing specific questions from surveys or increase the value of surveys without the need to add additional questions. For example, with data from SSA, we can observe whether a given individual received social security income and may not need to ask about it on a survey. Second, administrative records can be used to supplement survey response for topics in which we believe both add information. Given our knowledge that some earnings are unreported to the IRS, survey earnings information can be combined with administrative reports to provide more accurate income estimates than either could separately.

However, there are also many open questions to resolve and considerable research needed

to understand how best to use administrative data to improve income statistics. In this paper, we discuss these issues as well as the existing and pending research that can help us understand how to address them. The primary issues to blending survey and administrative data, which we will discuss in more detail later in the paper, are:

- Administrative records can have measurement error, which can bias income estimates just as mis-reporting and measurement error in surveys can.
- Incomplete linkages - 10 percent or more of individuals cannot be linked to administrative records in a given survey and year.
- Linkage errors - Linkage to administrative records is based on a probabilistic match, and can include errors.
- Universe mismatch - The populations intended to be represented in surveys do not necessarily match those contained in administrative records.
- Geographic coverage - For some income and resource items, administrative data are available only for some locations in some years.
- Income coverage or conceptual misalignment - For some income and resource items, the available administrative data do not capture all the possible income in that category or do not match the concept being asked in the survey.
- Timeliness and continued availability - Not all administrative data are available in time to be used for estimating income statistics on Census's current production schedule.
- Administrative records can change for reasons unrelated to income changes.

The discussion in this paper generally proceeds without elaborating on the status of agreements and approvals between administrative data providers and the Census Bureau on how their data are allowed to be used for the production of official income statistics. In this paper, we try to lay out the technical challenges and possibilities of using administrative and survey data together. We do so with the awareness that additional policy, legal, and regulatory issues would need to be resolved before official income statistics could be estimated

in the manner we describe here. Regardless, this research is necessary to get the best possible estimates of income statistics. Only then can we understand the magnitudes of biases or uncertainties present in any alternative methods, including our current approach using only survey data.

The paper proceeds as follows. In Section 2, we discuss the research on survey and administrative income data in more detail. In Section 3, we detail the administrative data currently available for research at the Census Bureau and some of the limitations of those data. In Section 4, we discuss challenges to using administrative data with survey data to estimate income statistics. In Section 5, we propose a simple model for survey and administrative income reporting which serves as a basic framework for thinking about estimating "true" underlying income in the presence of survey and administrative reports that may each contain reporting error. In Section 6, we lay out the research agenda for the AIS project. Section 7 concludes.

2 Background

2.1 Mis-Reporting and Measurement Error

Earnings, which constitutes about 80 percent of household income, has been the most heavily studied component of income in linked data.[2] (Alvey and Cobleigh 1975, Duncan and Hill 1985, Bound and Krueger 1991, Bound et al. 1994, Pischke 1995, Bollinger 1998, Bound et al. 2001, Roemer 2002, Kapteyn and Ypma 2007, Gottschalk and Huynh 2010, Meijer et al. 2012, Abowd and Stinson 2013, Murray-Close and Heggeness 2018, Bee et al. 2019, Imboden et al. 2019, and many others have studied wage and salary earnings.) Although survey earnings are relatively well reported when compared to external benchmark aggregates (Rothbaum 2015), work with linked microdata has found systematic differences between administrative

[2]Earnings make up about 80 percent of all personal income in the Bureau of Economic Analysis's National Income and Product Account tables as well as in the CPS ASEC (Rothbaum 2015).

records and survey responses.

The aforementioned papers have generally found survey wage and salary earnings are "mean-reverting" relative to administrative reports; i.e., low earners in the administrative data tend to report higher earnings on surveys, and high earners in the administrative data tend to report lower earnings in surveys (see Figure 1). There is also extensive margin disagreement between survey and administrative records – about 10 percent of working-age individuals have earnings in only one data source but not the other (Bee et al. 2019).

When discussing measurement error, we cannot assume that it is present in surveys only. Under-the-table earnings are, by definition, not reported to the IRS, which can bias income estimates for particular subgroups of the population (such as by occupation). Some papers in the survey mis-reporting literature assumed the administrative records were free of error (Bound and Krueger 1991, Bound et al. 1994, Pischke 1995, for example).[3] However, more recent work considers the possibility that administrative data also contain measurement error, such as unreported earnings. Abowd and Stinson (2013) consider a model where both survey and administrative reports for a given job may contain error. Under their approach, "true" earnings is a weighted average of the two reports, but they leave the selection of the proper weight to future work. Using Danish administrative data, Bingley and Martinello (2017) cannot rule out that survey income reports have only classical measurement error given the presence of measurement error in administrative records. In the absence of a "truth set" of data, it is an open question how much of this disagreement is due to mis-reporting on surveys or in the administrative data.

Compounding the challenge, it is not even always the case that different sources of administrative data agree. Bee et al. (2019) find a 0.4 percentage point difference in the estimated poverty rate if survey earnings are replaced using administrative earnings data from SSA compared to data from IRS, both of which are based on the same W-2s.

[3]In some cases, the authors restrict their analysis to a subset of workers where the assumption is more likely to be valid. For example, Pischke (1995) compares surveys of employees of a particular firm against firm reports of the same workers' earnings. Bound and Krueger (1991) specifically remove occupations they suspect may have under-the-table earnings.

The likelihood of measurement error in administrative data is particularly high for self-employment earnings. The IRS publishes estimates of self-employment under-reporting using data from random audits (Internal Revenue Service, Research, Analysis & Statistics. 2016). As a result of under-reporting to the IRS, the Bureau of Economic Analysis (BEA) adjustment for self-employment income averaged 86 percent of the net profit reported in tax returns from 2010 to 2015 (U.S. Bureau of Economic Analysis). In other words, nearly half of the BEA's National Income and Product Account (NIPA) estimate for self-employment earnings is not reported to the IRS. In contrast to wage and salary earnings, Abraham et al. (2018) find that those that report self-employment earnings in the CPS ASEC are unlikely to report any self-employment earnings (on the 1040-SE) to the IRS. They also find the reverse: those with positive self-employment income in a 1040-SE are likely to report no self-employment income in the CPS ASEC. By looking at the relationship between earnings and consumption, Hurst et al. (2014) estimate that self-employment earnings are under-reported by 25 percent.

Surveys capture a smaller share of the NIPA aggregates for income sources other than earnings, especially for property income (interest, dividends, and rental income) and retirement income (Rothbaum 2015). For those 65 and older, Bee and Mitchell (2017) find that under-reporting of interest, dividends, and, especially, retirement income in the CPS ASEC results in an underestimate of median household income by about 30 percent and an overestimate of poverty by about 25 percent, with substantial differences across the distribution (see Figure 2).

Harris (2014), Meyer et al. (2018), Meyer and Mittag (2019), Shantz and Fox (2018), and others have found substantial under-reporting of program benefit receipt in various Census income surveys. For example, Shantz and Fox (2018) compare administrative records for the Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) to CPS ASEC survey responses. They find that 43 percent of SNAP recipients and 62 percent of TANF recipients in the administrative data do not report those

benefits in the survey.[4]

2.2 Non-Response Bias

Administrative data can also shed light on another source of bias in surveys, non-response bias. Non-response to a given survey is often partitioned into two types: 1) unit non-response – where no respondent in a given housing unit answered the survey, and 2) item non-response – where a respondent answered the survey but did not respond to a particular question. Both types of non-response have been increasing for income surveys, including the CPS ASEC (Hokayem et al. 2019), the Survey of Income and Program Participation (SIPP) (Czajka and Beyler 2016), and the American Community Survey (ACS) (Rothbaum 2018). Figure 3 plots unit non-response rates over time, and Figure 4 shows income item non-response trends in the ACS and CPS ASEC. As non-response increases, the nature of selection into non-response becomes an increasingly important potential source of bias.

Bee et al. (2015) compare the 1040 income distribution of CPS ASEC unit non-respondents and respondents and do not find strong evidence of non-response bias, although sample size and lack of statistical power limits their ability to evaluate non-response at the top of the income distribution. Brummet et al. (2018) apply the same method to the Consumer Expenditure Survey and find evidence of higher unit non-response rates for high-income households, which Eggleston and Westra (2019) also find in the SIPP.

Hokayem et al. (2015) and Bollinger et al. (2018) find strong evidence that income and poverty statistics are biased due to item non-response to earnings questions that is not properly accounted for in the CPS ASEC imputation model. They show that low and high earners in administrative records are the least likely to respond to survey earnings questions, shown in Figure 5. Using an improved imputation model, Hokayem et al. (2019) find that the

[4]They calculate how much mis-reporting biases estimates of the Supplemental Poverty Measure (SPM), which includes cash welfare (such as TANF) and also in-kind benefits (such as SNAP, which is not included in official poverty estimates). Despite the high false negative rate, they do not find a statistically significant effect on the SPM rate from TANF under-reporting. However, SNAP under-reporting biases the SPM estimate upward by 0.5 percentage points or about 4 percent.

non-response and imputation bias causes current estimates to overstate median household income by 1 to 5 percent and to understate poverty by 0.5 percentage points, with the impact of improved imputation on the income distribution shown in Figure 6.

2.3 Income Estimates using Administrative Data

There has been a lot of work, both in the U.S. and internationally, to estimate income statistics that correct for survey under-reporting. In some cases, such as inequality statistics, where large samples and better coverage of extremely high-income individuals may be necessary, researchers have used administrative data exclusively (Piketty and Saez 2003, Piketty et al. 2017, Larrimore et al. 2019, and Auten and Splinter 2018, to cite a few). Burkhauser et al. (2012) find that surveys and administrative records generally agree on the trends for inequality over the period where both data sets are available (1967 forward). Larrimore et al. (2019) create households from tax-filing units to better match the unit of analysis in survey data. They find that inequality in the adjusted tax data is higher than in the CPS ASEC, but also less than when calculated using tax-filing units, as has generally been done in the literature.[5]

Other researchers have used survey data linked to administrative records. For example, using linked data for the United Kingdom, Burkhauser et al. (2017) find that estimates that incorporate administrative data reveal greater increases in inequality over the period studied than the survey data alone. Meyer et al. (2019) find that, due to survey under-reporting, many fewer households have <$2 per day in resources than surveys indicate. Meyer and Mittag (2019) find that under-reporting causes surveys to understate the anti-poverty impact of safety net programs. Meyer et al. (2018) find that survey estimates of SNAP participation understate participation by certain groups, such as single parents, non-whites, and low-income households. They also find that surveys understate the share of households that participate in multiple means-tested benefit programs.

[5]They find Gini coefficients of 0.483, 0.524, and 0.570 in the CPS ASEC, household-based tax data, and filing-unit-based tax data, respectively.

Others have attempted to adjust survey data for under-reporting without direct links to linked administrative data, including teams at the BEA for Distributional National Accounts (Fixler and Johnson 2014), the Urban Institute's Transfer Income Model (TRIM) (Zedlewski and Giannarelli 2015), and the Congressional Budget Office's (CBO) regression-based adjustment framework (Habib 2018). Jenkins (2017), Hyslop and Townsend (2018), and Fixler et al. (2019) use statistics from administrative data to estimate parameters of the Pareto distribution to correct survey estimates of income at the top of the distribution. Fixler and Johnson (2014) make a series of adjustments from administrative and other sources to correct for survey under-reporting. The Small Area Estimates Program at the Census Bureau uses aggregate administrative statistics to generate more precise estimates of income, poverty, and health insurance for small geographies than is possible using existing surveys alone (U.S. Census Bureau a and U.S. Census Bureau b).

2.4 Administrative Records in Production at Census

There are several programs at the Census Bureau where linked administrative income data are already being used to edit survey responses and estimates. The SIPP uses information on program receipt of Supplemental Security Income (SSI) and social security benefits to address program confusion, in which respondents report benefits actually received from one program in the other (Giefer et al. 2015). The SIPP Synthetic Beta (SSB) and Gold Standard files include administrative information on earnings and social security and SSI benefit receipt linked to SIPP respondent information (Benedetto et al. 2013 and Benedetto et al. 2018). The SSB project makes synthetic data available to the public, with validation on the original data possible, subject to disclosure review (Vilhuber and Abowd 2016). The SIPP uses model-based imputation with administrative data for topic flags, which are indicators for the presence of various income items (U.S. Census Bureau 2019). Administrative data is also used for survey sampling, for example in targeting households with children in the National Survey of Children's Health (U.S. Census Bureau 2018).

3 Administrative Data

Title 13, Chapter 1, Section 6 of the U.S. Code states, "To the maximum extent possible and consistent with the kind, timeliness, quality and scope of the statistics required, the (Secretary of Commerce) shall acquire and use information available from any source referred to in subsection (a) or (b) of this section instead of conducting direct inquiries." Under this authority, the Census Bureau has entered into agreements with various federal and state agencies and private organizations for access to administrative data to be used for statistical purposes.

In this section, we describe the data currently available under these agreements for research purposes only, including their strengths and shortcomings. We also identify data not currently available to the Census Bureau, but available to other federal or state agencies, which could help improve our income statistics. The available administrative records are shown in Table 1.

First, we will briefly list the available data by agency, and then discuss the data by income type. From the IRS, the Census Bureau receives several files with a subset of variables from the universe of filings and information returns. In the 1040 extract, this includes taxable wage and salary earnings, taxable and tax-exempt interest, taxable dividends, social security income, gross rental income, total money income, and adjusted gross income. The 1040 extract also includes flags to indicate whether various 1040 schedules were filed, but not the income reported on those returns. These include flags for Schedules A, C, D, E, F, and SE. The caveat for any income source only available on the 1040 is that the information is only present for tax filers.[6] Furthermore, income data from 1040s is for the tax unit and does not identify the individual in the unit who received the income.

The IRS also provides income information in universe W-2 and 1099-R files. Each of these files contains a subset of the fields reported on the respective forms. The Census also receives a file containing flags that indicate whether certain other information returns were

[6]Roughly 20 percent of households in the ACS do not file tax returns in an average year.

filed, including only Forms 1099-MISC, 1099-S, 1099-G, SSA-1099, 1099-INT, and 1099-DIV. Census does not receive income amounts from this latter set of forms.

From the SSA, the Census receives several files with income information. Unlike the IRS data, these are not universe files. Instead SSA provides extracts for those individuals in the CPS ASEC and SIPP that have been linked to their Social Security Number using the personal identifying information (such as name and date of birth) in the survey.[7] The most important files that SSA provides are the Detailed Earnings Record (DER), Payment History Update System (PHUS), and Supplemental Security Record (SSR) files. The DER contains longitudinal wage and salary earnings from W-2s. The PHUS contains social security Old-age, Survivor, and Disability Insurance (OASDI) payments. The SSR contains SSI payments.[8]

From the Department of Health and Human Services (HHS), Census receives information on Medicare and Medicaid participation as well as data from a subset of states' TANF programs. From the Department of Housing and Urban Development (HUD), Census receives data on housing assistance payments and public housing provision.

As a part of the Longitudinal Employer-Household Dynamics (LEHD) program, states provide data on wage and salary earnings as reported by firms for the administration of the unemployment insurance program. Some states also provide SNAP and Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) benefit receipt. Energy Assistance program data from the Low Income Home Energy Assistance Program (LIHEAP) is currently available to Census for one state.

3.1 Earnings

As noted above, four main sources of administrative earnings data are available to the Census Bureau: 1) universe W-2 records from the IRS, 2) the DER file from SSA, 3) the LEHD

[7]For details on the linkage process, see, e.g., Wagner and Layne (2014).

[8]SSA also provides a few other files with income information, such as the Summary Earnings Record (SER) file, but we have focused on the ones mentioned above for brevity and clarity.

earnings files, and 4) earnings reported on 1040s. The available W-2 data include select boxes from the form, including wages and salary net of pre-tax deductions for health insurance premiums and deferred compensation, as well as the total amount of deferred compensation. This means that employee and employer pre-tax contributions to health insurance premiums are not available in the W-2 data. The DER data include the same information from the W-2, but only for linked CPS ASEC and SIPP respondents. The DER also includes social security covered self-employment income from the 1040 SE form. LEHD data come from employer reports of earnings to states for the administration of the unemployment insurance program, as well as data available from the Office of Personnel Management for some federal employees. The LEHD contains the universe of all wage and salary workers, as reported to the states. However, many government employees, such as postal workers and Department of Defense employees, are not present in the LEHD. Additionally, some private sector workers, including those employed by religious organizations, are not covered by unemployment insurance, and are therefore not present in the LEHD data.

Because it contains gross earnings, the LEHD may better capture total wage and salary earnings for a given job than the currently available W-2 or DER files. This comes with the downside that many jobs are missing in the LEHD that are present in the W-2 or DER files. The DER file is the only source available to Census that contains administrative reports of self-employment (Schedule SE) earnings. However, the DER is not a universe file as it only contains earnings information for linked CPS ASEC and SIPP respondents.

For tax filers, earnings information is available on the 1040. Collins et al. (2019) find that a small, but non-trivial, share of tax filers report higher earnings on their 1040 than is reported in their W-2s. This could be the result of mis-reporting on W-2s or proper reporting of earnings on the 1040 that are not required to be reported on a W-2, such as tips not reported to one's employer, scholarship and fellowship grants, employer-provided adoption benefits, etc. Although it may be a challenge to assign this income to a particular individual, this can be a form of "under-reporting" of earnings in W-2 forms.

Some self-employed workers receive 1099-MISC forms as independent contractors. Because we do not have access to information on income amounts from the 1099-MISC, we can only validate whether the presence of a 1099-MISC corresponds to survey reports of self-employment. Others (such as "gig" workers) receive 1099-K forms reporting income received from "Payment Card and Third Party Network Transactions." Census does not currently receive any information from 1099-K filings. Census also has flags for whether various 1040 self-employment schedules were filed (such as C, F, and SE) for the universe of filers, but these also do not contain reported income amounts. Given the currently available data for these forms, Census can only evaluate whether the presence of the form corresponds to survey reports of self-employment.

3.2 Transfer Income

Census surveys ask about several sources of transfer income provided by federal, state, and local governments, including unemployment insurance benefits, social security benefits, SSI, public assistance and cash welfare, and veterans benefits.

Unemployment insurance benefits are reported to the IRS on Form 1099-G. As with most other information returns, the Census Bureau receives a flag indicating the presence of a 1099-G return for each individual, but not the amount of income received.

For social security benefits, Census receives payment information from SSA for CPS ASEC and SIPP respondents. The 1040 extract provided by the IRS includes taxable social security income for the universe of tax filers. SSI is not taxable income and is therefore not reported to the IRS. The SSR file contains SSI payment information for CPS ASEC and SIPP respondents.

For public assistance and cash welfare data, TANF program data are available for some states. However, TANF is not the only possible source of cash welfare, as other state and local programs can provide cash assistance to needy individuals and households.

Veterans benefits data from the U.S. Department of Veterans Affairs (VA) are available

for some benefit programs.

3.3 Retirement, Survivor, and Disability Income

IRS provides an extract of the universe of 1099-R information returns including information on amounts of defined-benefit pension payments (including for survivor and disability pensions) and withdrawals from defined-contribution retirement plans. The DER also includes, through 1990, amounts from IRS Form W-2P "Statements for Recipients of Annuities, Pensions, Retired Pay, or IRA Payments," the predecessor to Form 1099-R.

3.4 Property Income

Interest, dividend, and rental income are available in the 1040 file from IRS. However, the 1040 extract received by Census contains only the gross rental income amount. The CPS ASEC and ACS ask for rental income net of expenses, and the SIPP asks for both net and gross rental income.[9]

3.5 Other Income

Various other income items are reported to the IRS, but are not currently available to Census. These include educational assistance (Pell grants, scholarships, etc.), alimony, capital gains, and gambling winnings.

Other income items are either not reported to the IRS or not clearly distinguishable from other sources of income in the IRS data. These include child support received and workers' compensation.

[9]Net rental income on the 1040 deducts depreciation, which is not likely to be accounted for in net rental income on surveys.

3.6 Taxes and Credits

The Census Bureau does not receive information on tax obligations or tax credits (such as the Earned Income Tax Credit [EITC]) from the IRS. Instead, taxes are modeled based on responses to family relationship and income questions.[10] Comparisons to administrative EITC data (Jones and Ziliak 2019) and comparisons to tax aggregates (Wheaton and Stevens 2016) reveal biases in Census tax model estimates, resulting from some combination of measurement error in the survey data, the absence of survey questions to estimate some deductions and taxable income items, missing information on dependent claiming, and imperfect assumptions about the construction of tax-filing units from individuals and households.[11]

3.7 In-Kind Benefits

For in-kind health insurance benefits, such as Medicare and Medicaid, beneficiary information is available from HHS. SNAP and WIC data are available for some states, but not all, as Census must negotiate data sharing agreements with states individually. Housing assistance information is available from HUD. However, like TANF and cash welfare, the federal data only cover a portion of all housing assistance programs, as state and local programs are not part of the HUD data. For energy assistance, Census has LIHEAP data from one state. LIHEAP also does not cover all possible sources of energy assistance.

[10] See the tax model documentation at: https://www.census.gov/content/dam/Census/library/working-papers/2005/demo/cpsasec2005taxmodeldoc.pdf.

[11] Wheaton and Stevens (2016) also evaluate the TAXSIM model produced by the National Bureau of Economic Research (Feenberg and Coutts 1993, the Bakija Model (Bakija 2009), and Urban Institute's TRIM model (Zedlewski and Giannarelli 2015) and do not find large differences in the quality of the model estimates from CPS ASEC data (as compared to each other or the Census model) relative to administrative benchmarks.

4 Challenges to Incorporating Administrative Records into Income Estimates

4.1 Administrative Records Include Measurement Error

Even the best administrative data contain some error. The nature and severity of that error relative to survey responses can determine how we should use the data. For example, some people are paid under the table. That income will not show up in a W-2 and is unlikely to be reported on a 1040 return.

Collins et al. (2019) find that many recipients of 1099-MISC and 1099-K forms do not report that income in a 1040 Schedule SE. That income information is therefore not available to Census in the currently available administrative data, which is in effect a form of "under-reporting" in the available data. This missing income information limits our ability to assess survey mis-reporting and to research how we would use administrative data in these cases.

The nature of these errors and other potential sources of error in administrative records is difficult to assess (even if only qualitatively) or to communicate to end users, because the provenance and custody of most administrative data sources are poorly documented. In theory, survey data are originally collected under a controlled protocol using interviewers that have been trained under similarly consistent conditions. Nearly every aspect of the survey design, weighting, collection, and processing is controlled and documented by the survey organization. Administrative data, on the other hand, usually have a murkier provenance. While administrative data analysts can usually find the forms from which administrative record fields are populated, it can be difficult to triangulate exactly what set of records is included in each file and what processes have been applied for data capture, processing, and cleaning.

For example, wage and salary values on the SSA DER files and the IRS W-2 files are theoretically identical, since the DER is built from the same Box 1 values from the W-2. However, Bee et al. 2019 find significant differences between the two data sources for reasons

that remain unclear. Presumably both SSA and IRS undertake some cleaning to flag, change, or remove duplicate or mistaken entries, but the details of those processes remain an object of ongoing inquiry.

4.2 Incomplete Linkages

The Census Bureau developed the Person Identification Validation System (PVS) to probabilistically match survey respondents to their Social Security Number (SSN) using personally identifying information (PII) from the survey, such as name, date of birth, and residential address (Wagner and Layne 2014). Then the record is assigned a Protected Identification Key (PIK) and the SSN and PII are removed from the file. SSNs in administrative records are also replaced with PIKs. The PIK is used as the anonymized linkage key to match respondents to their administrative records.

However, this means that if PVS is unable to assign a PIK to a given survey respondent, no administrative data are available for that person. From 2006 to 2013, the linkage rate in the ACS varied between 90.7 and 93.3 percent (Rothbaum 2018). The rates are slightly lower for other income surveys such as the CPS ASEC and SIPP.[12] Observable characteristics, such as race, ethnicity, citizenship status, etc., are correlated with PIK assignment (Bond et al. 2014).

This creates a missing data problem in linked records. If we assume that administrative data are "missing at random" conditional on observables due to incomplete linkage, we can use inverse probability weights to adjust our PIKed sample to make it representative of the entire survey sample. However, if that assumption is violated, then we may have to consider alternative approaches to handling incomplete linkage.

[12]Monti and Reeder (2016) report linkage rates by SIPP panel, which vary substantially over time. The linkage rate in the 2008 and 2014 panel were 82 and 89 percent, respectively. Bollinger et al. (2018) find a linkage rate in their CPS ASEC sample (2006-2011) of 86 percent.

4.3 Linkage Errors

Because the linkage process is probabilistic, there will be cases where we have linked an individual to the incorrect SSN. In those cases, survey responses and administrative records will disagree even if both were measured without error as we would be comparing the income of two different individuals. Layne et al. (2014) estimate the rate of linkage error at less than 1 percent.

Kapteyn and Ypma (2007) consider the possibility of mis-linkage in evaluating earnings mis-reporting. They find that mis-linkage can explain the mean-reversion found in survey earnings compared to administrative data. Meijer et al. (2012) find that with even a small probability of mis-linkage, survey reports may provide more reliable measures of income. An important caveat for both of these studies is that they consider mis-linkage as the only reason administrative data would contain error. Allowing for mis-reporting in administrative data would likely reduce their estimates of mis-linkage and therefore of the impact of mis-linkage on data quality.

4.4 Universe Mismatch

Not all tax filers are in-universe to be sampled for income surveys, which typically aim to describe the civilian, non-institutionalized population. For example, foreign nationals with income in the U.S. may be required to file taxes. They may be among the universe of administrative records, but are not intended to be included in income statistics.

Very high income households may also be differentially likely to be included in survey samples for a variety of reasons, including unit non-response or frame coverage challenges. This might result in differential likelihoods for different tax units to be in the survey frame.

Conversely, not all households are "in universe" for all administrative records. For example, some households are not required to file a tax return if their income is below a specified threshold. There are similar thresholds for the filing of information returns as well. For example, if the employer did not deduct withholdings, they are not required to file a W-2

on earnings less than $600. The limit is $20,000 for a 1099-K.[13]

These types of issues can result in universe mismatches between surveys and administrative records, such that individuals that should be included may be absent from one or both of the data sources.

4.5 Geographic Coverage

In many cases, Census has access to administrative data for only a subset of geographic locations covered by the survey. For example, SNAP, WIC, and TANF data are currently available for some states in some years. Even if we assume that data are relatively free of error and can be used to directly replace survey responses, we still must address the survey reporting error in the locations for which we do not have administrative records.

One possible method to do so is the Transfer Income Model (TRIM) developed by the Urban Institute under contract with HHS (Zedlewski and Giannarelli 2015). For SNAP mis-reporting, TRIM uses aggregate statistics on state program spending and quality control data on the relationship between reported income and SNAP receipt and the amount received to distribute the missing SNAP spending to CPS ASEC households. Another approach used by the CBO to adjust for under-reporting in the absence of administrative microdata (Habib 2018) is to assign an expected probability to each household that did not report benefit receipt using a probit regression on survey reports of receipt. They then assign program receipt and benefits to households with the highest expected probability of receipt until they match program aggregates. Mittag (2019) proposes an approach that uses linked survey and administrative data to produce summary statistics that can be used to adjust for survey under-reporting. Another possible approach being explored by Census is to use the available administrative data to impute across state lines (Fox et al. 2019).

Each approach makes assumptions about the relationship between benefit mis-reporting and household and individual characteristics. Because of data constraints, the TRIM and

[13] A 1099-K is also required if there are 200 payment transactions to the given payee.

CBO approaches require assumptions that are difficult to verify. When imputing across state lines, the assumptions can be tested by leaving a state with administrative data out of the model and comparing estimates and regression results using the administrative data against the imputed data.

4.6 Income Coverage or Conceptual Misalignment

In some cases, the income concept being measured in a survey is broader or not identical to the concept being measured in administrative data. For example, non-taxable income may not appear on a 1040, but would be intended to be included in the survey responses.

For some program benefits, Census has access to administrative records for a subset of all possible programs that provide a benefit. For example, there are questions about cash public assistance or welfare payments on both the ACS and CPS ASEC. Census has data from some states' TANF programs on benefits paid. However, TANF is not the only possible source of cash assistance. Similarly, for housing assistance, Census has data from HUD that does not necessarily cover all possible sources. In these cases, there is the potential for "under-reporting" of program benefits in the administrative data because the data are not available for other sources of benefits. This presents a challenge in evaluating survey responses. If survey benefits exceed administrative benefits, is it because of over-reporting (either random or systematic) or because some of the benefits were from a program not covered by the administrative source? For benefits provided at the state and, especially, the local level, it is a bigger challenge to get good data on program benefits either at the individual or aggregate levels.

4.7 Timeliness and Continued Availability

A major challenge to using administrative data for income statistics is the timeliness of data availability to the Census Bureau. The CPS ASEC is the source for the Annual Income and Poverty Report (Fontenot et al. 2018) and associated tables. These are released each

September, with details on income and poverty in the previous calendar year. The CPS ASEC is also the income survey with the shortest period between data collection (February through April) to data release (in September).[14]

For the CPS ASEC, after data collection ends at the end of April, processing and review begin in May and proceed through data release in September. For administrative data to be incorporated into production, it must be available for use at the beginning of this processing period. Some administrative data sources may not be available to Census in a timely enough fashion. Others may be available in incomplete form. For example, in 2018, 10.3 percent of all returns were filed between May 23rd and November 21, covering 21.2 percent of Adjusted Gross Income (AGI). For those with AGI of $1,000,000 or more, 54.1 percent of returns were filed in this period, which would be too late for use in CPS ASEC processing.[15]

Each administrative data set is available through an agreement between the Census Bureau and other federal or state agencies or private sector organizations. However, existing agreements are not always renewed. In the production processes, Census must be prepared for the contingency that data could cease to be available. Doing so without causing a break in data series will be challenging.

4.8 Administrative Records Can Change for Reasons Unrelated to Income Changes

By definition, administrative records are collected for program administration, not for producing income statistics. They can change in type or quality for reasons that are unrelated

[14]The ACS generally releases data tables and briefs in September using survey data from the previous calendar year. The period covered by the ACS income data is more complicated. Each ACS respondent is asked about income in the previous 12 months. As the ACS is fielded continuously throughout the year, the design results in a rolling reference period. For the 2017 ACS, January respondents were asked to report income primarily from 2016 while December respondents were asked to report income primarily from 2017. Therefore, the 2017 ACS data, released in September of 2018, included income from 2016 and 2017. This also means that ACS reporting periods (prior 12 months at the time of the survey) and administrative data reporting periods (calendar years) do not align. The upcoming ACS content test will explore changing the reference period to prior calendar year to better align the survey with administrative data (U.S. Census Bureau 2017).

[15]Calculated from 2018 Filing Season Statistics data available online (Internal Revenue Service).

to changes in the underlying income or benefits. For example, the Economic Stimulus Act of 2008 provided a tax credit for low-income tax filers, which increased the tax filing rate and affected the income distribution of 1040 filers (see Figure 7). Auten and Splinter (2018) argue that tax changes in the Tax Reform Act of 1986 changed filing behavior causing an overstatement of the increase in inequality when measured in tax data, for example through increased use of pass through entities (S-Corporations) where income is visible in individual tax returns relative to that from C-Corporations, where it is not.

Changes to program administration, data collection, extract creation, etc. can all affect the data shared with the Census Bureau (and our ability to clean and process it in a timely fashion). Any of these changes can affect income or resource estimates in a way that is similar to a survey redesign, causing a potential break in series.

5 Model

In this section, we present a simple model of income or resource mis-reporting. This model is intentionally simplified relative to some used in the literature which specify different types of errors, such as reporting error, error due to mis-linkage, etc., and specific relationships and functional forms of the error terms.

Suppose a given individual i has true income or resources in a given category c (such as wage and salary earnings or SNAP benefits) of $y^*_{i,c}$, which is not observed. Instead, we observe given report r, where r may be the survey (s) and/or one or more administrative report (a). For any given set of individual characteristics and true income, there can be systematic reporting error, $e^r_{i,c} = f^r(y^*_{i,c}, X_i)$, where there is a distribution that describes the expected reporting error that is not just random variation and can be a function of true income ($y^*_{i,c}$ and a vector of individual characteristics X_i. For example, with some probability, an income report may indicate no income ($y^r_{i,c} = 0$) when true income is not zero ($y^*_{i,c} \neq 0$). It is also possible for the expected reported income (given a set of characteristics)

to under- or overstate true income. Finally, any income report can contain random noise, $v^r_{i,c} = g^r(y^*_{i,c}, X_i)$, and the variance of that noise may depend on the individual characteristics and true income, as well. For a single survey and administrative report, we observe the reported incomes $y^r_{i,c}, r = a, s$ as in:

$$\begin{aligned} y^s_{i,c} &= y^*_{i,c} + e^s_{i,c} + v^s_{i,c} \\ y^a_{i,c} &= y^*_{i,c} + e^a_{i,c} + v^a_{i,c}. \end{aligned} \tag{1}$$

An important characteristic of the model is that reporting error, whether systematic ($e^r_{i,c}$) or random ($v^r_{i,c}$) is possible for both survey and administrative reports. In fact, there are cases where we have strong evidence that administrative records have systematic error. For example, the absence of pre-tax employee contributions to health insurance premiums in currently available W-2 data is akin to systematic under-reporting. So are under-the-table earnings, as they are unlikely to be present in any administrative data on wage and salary earnings. IRS audit studies also show that the self-employed have substantial earnings that are not present in tax data (Internal Revenue Service, Research, Analysis & Statistics. 2016).

There may exist income sources where we believe there is little or no systematic reporting error or random noise in the administrative records. Absent error, $y^a_{i,c} = y^*_{i,c}$, and there is little or no additional information in a survey response when the administrative record is available. In those cases, the most accurate way to estimate income is with direct replacement, in which we replace survey responses with administrative income.

On the other hand, there are income sources that are not available in any administrative data source available to the Census Bureau. In those cases, the survey responses may be the most accurate possible source of income estimates.

A third possibility is that we can combine multiple sources of information that each contain error using a model to more accurately estimate income. For example, if we believe some individuals systematically under-report their earnings in surveys (such as high earners) whereas others systematically under-report in the administrative data (such as those in fields

that are more likely to be paid under-the-table), we can combine the two reports, giving more weight to survey reports in some cases and administrative reports in others. Modeling may also be necessary in cases where administrative data are not available in all locations, does not cover all possible income in a category, or is not available in a timely manner.

In the framework of equation 1, we only observe $y^r_{i,c}$. In the absence of a "truth" set that we know is free of error, we will have to make some assumptions about the nature of the errors in the administrative and survey reports and their covariance relationships. However, we must be transparent about these assumptions and evaluate them, to the extent possible.

For example, one method that has been used for wage and salary earnings is to take the maximum of administrative and survey reports.[16] In terms of the model, this would be akin to asserting that $y^*_{i,c} = \max(y^s_{i,c}, y^a_{i,c})$. For this to be unbiased, one must assume that there is no group for which earnings are systematically over-reported ($e^r_{i,c} \leq 0$) and that there is no random noise in the reports ($v^r_{ic} = 0$). By comparing survey to administrative reports, we can check whether these appear to be reasonable assumptions. Figure 1 shows a scatter plot of survey and administrative reports of wage and salary earnings in the ACS and W-2s. While most of the points are clustered around the 45-degree line, there is evidence of dispersion in the survey responses around the administrative values. This does not support the assumption that there is little or no random noise in survey responses unless we assume that all points above the 45-degree line are due to under-reporting to the IRS and all points below the 45-degree line are due to under-reporting to the ACS. Therefore, taking the maximum of the two reports is likely to bias income estimates up.

Table 2 shows our best, current assessment of how administrative records could be used to best estimate various income and resource items.[17] The assessment takes as given the set of currently available administrative data and the assumption that Census had access to broader state and federal tax and program data. For earnings, because of survey reporting

[16] As in Meyer et al. (2019).

[17] In this section, we are abstracting away from confidentiality concerns that would likely result in additional noise added to the estimates for disclosure protection.

error and under-the-table earnings missing from administrative records, we believe some form of modeling is necessary, as discussed above. For many income or resource items, such as some veterans benefits, capital gains, employer contributions to health insurance, etc., we currently do not have access to administrative records which we could use for direct replacement of survey responses. Instead, the estimates would be based on survey responses for those items.[18]

Even with all possible federal and state data, we believe modeling will be necessary for some resource categories, either because of possible under-reporting in the administrative data (earnings and maybe rental income) or because the federal and state administrative records do not cover all possible income.[19]

6 Administrative Income Statistics Research Agenda

We are organizing much of the ongoing and planned research at Census into the Administrative Income Statistics (AIS) project. Under the umbrella of AIS, Census will continue to research how survey and administrative data can be used to create more accurate data sets and income and poverty estimates and to improve the ability of Census and outside researchers to evaluate the well-being of households, families, and individuals in the United States. We outline the main tasks of this project as follows.

[18] "Survey Only" estimates could include modeling as well, but the modeling would not be based on linked administrative microdata for those items. For example, without state administrative data, SNAP under-reporting can be modeled using state spending aggregates and other data, as the Urban Institute does in the Transfer Income Model and the Congressional Budget Office does using a regression-based correction.

[19] As noted previously, possible cases include TANF for cash welfare income, HUD data for housing assistance, and LIHEAP data for energy assistance. In each case, other state and local programs may exist that are not covered by the administrative data, but that we would like to account for in our income and resource estimates.

6.1 Process Administrative Data and Compare to Survey Responses

The first step in the research agenda is to clean the income administrative data and compare the administrative reports to survey responses. As noted before, substantial research, both within Census and in the academic community, has been done along these lines. However, work remains in cleaning and evaluating various items, including state program benefits and less studied income items, such as rental income, interest, and dividends. This cleaning, processing, and benchmarking will also result in more complete, standardized documentation that describes the provenance of these records in as much detail as possible.

6.2 Resolve Practical Issues in Estimation using Multiple Data Sources

6.2.1 Error in Administrative Data

For any given item, once we understand how surveys and administrative reports differ, we must decide how to combine them to create our estimate of "true" income ($y^*_{i,c}$). For some items, we may decide that direct replacement is not appropriate because of reporting error in administrative records. In these cases, research is necessary to determine how best to combine the information in the survey and administrative reports. This involves testing assumptions about the relationship between reporting error and individual characteristics, estimating reporting error parameters, and potentially imputing values for $y^*_{i,c}$ given those assumptions and estimated parameters.

Earnings is likely the most important income type for this work, as earnings compose about 80 percent of household income and administrative earnings have known sources of measurement error. For wage and salary earnings, unreported and under-the-table earnings may cause administrative data to understate true earnings. For self-employment, both survey and administrative records understate estimates of true earnings from audit studies

and consumption data. It is very much an open question how to distribute the missing self-employment earnings to individuals and households given the available survey and administrative data.

When various data sources disagree, auxiliary survey and administrative data can help provide resolution. Just as tax officials might audit income by collecting non-income measurements, income analysts can use data on consumption, health, assets, mortality, residential location, occupation, etc., to triangulate true income. While Hurst et al. (2014) provide the clearest example of this approach, as they compare reported food consumption between self-employed and wage earners at given levels of income, a few other literatures also point in this direction. For example, Bollinger et al. (2013) and Pradhan and Ravallion (2000) use subjective well-being reports to calibrate poverty measures. Evans and Garthwaite (2014) show that biomarkers measured in the National Health and Nutrition Examination Survey, including those from collected blood specimens, reflect income received from the EITC, which as we note above is itself difficult to measure or model in household surveys. Meyer and Sullivan (2012) apply a suite of correlated well-being measures to judge among various alternative poverty measures. One caveat to this line of inquiry, however, is that usual overfitting concerns apply; information that becomes incorporated into modeling choices in the production of income statistics will lose its utility for independent checks of those statistics.

6.2.2 Incomplete Geographic Coverage

We can treat the problem of incomplete geographic coverage of administrative data as a standard case of missing information. It is as if all individuals in particular locations refused to answer the question, "What do administrative records in your state SNAP office say about your participation in the program?" We can then impute responses as we would any other piece of missing information by modeling administrative program data given the observable information.[20] As with any imputation model, we must assume the data are "missing at

[20]The observable information can include survey response, other administrative income and program data, state program information, other state level characteristics, etc.

random," which means that for a given estimate or parameter of interest, we can control for selection into response (state sharing of the data with Census) given the observable characteristics in the data available.

The advantage of this approach over alternative approaches that lack access to administrative program data linked to survey responses is that the underlying assumption can be partially tested. We can compare estimates using the administrative data to estimates produced using the imputed data. For example, if we have program data from New York, we can treat that data as missing and impute SNAP participation and amount received using the other state data we have. We can then compare summary statistics and regression coefficients for New York using the imputed data and the actual administrative records. To the extent they differ, we have evidence either that the data are not missing at random or that our imputation models are mis-specified.[21] This research is ongoing for SNAP benefits, with encouraging results (Fox et al. 2019). We plan to expand the work to include other programs, such as WIC, TANF, and housing assistance. By incorporating many characteristics into the imputation model, we also make the data useful to as many possible analyses given potential restrictions on data access.

6.2.3 Incomplete Income/Resource Coverage

A more complicated issue arises if the administrative income data do not include all possible sources of income that are in the survey report. For example, TANF is one possible source of general assistance offered by state and local governments. The housing assistance data cover HUD-administered programs, but not other sources of assistance. In both cases, the administrative reports could be completely free of error for the payments they cover while also understating the assistance received. In these cases, survey reports could exceed administrative records because of error in the survey report or error in the administrative record due to incomplete program coverage.

[21]The usual concerns regarding overfitting still apply, however, and one can never be completely certain that a model that works well for observable states will continue to fit beyond that sample.

To address under-reporting of benefits from other sources, we could use additional information on the programs available to households at the state and local levels, such as eligibility rules, number of participating households, and aggregate benefits. We could also use information about the nature of survey reporting error (under-reporting and random noise) for other programs, such as SNAP, where incomplete benefit coverage is not an issue. With this information we could attempt to "impute" true program benefits given the two benefit reports under our modeling approach from section 5. These imputations would necessarily require assumptions that would be difficult to directly verify in the data.

6.3 Leverage Administrative Records to Improve Survey Operations

6.3.1 Survey Frame Coverage

We would like to evaluate how survey frames and administrative data universes compare to each other. In doing so, we can test whether housing units with different characteristics are being included with equal probability in survey frames. This may be especially relevant for households at the right tail of the income distribution. If we find differences in frame coverage, we can then evaluate mechanisms for correcting surveys, such as adjusting survey weights and including un- or under-represented households from administrative records in the survey data directly or through imputation and/or data synthesis.

6.3.2 Unit Non-Response Bias

Even if all households are surveyed with equal probability (or their intended probability given an over-sampling mechanism), they are not necessarily equally likely to respond to surveys. Unit non-response is when no one in the household can be reached or they refuse to participate in the survey. To address this, surveys include weights that are adjusted to match target population moments, such as state populations, counts of people by race and Hispanic

origin, etc. However, if income is correlated with unit non-response, income statistics may be biased. There is ongoing research on both the CPS ASEC (Bee et al. 2015) and SIPP (Eggleston and Westra 2019) to evaluate the nature of this bias. We plan to continue to study the relationship between income and unit non-response.

6.3.3 Item Non-Response and Imputation

Even if an individual responds to the survey, that person may choose not to respond to particular questions. As noted in Section 2.2, this item non-response to income questions has been increasing in Census surveys. When direct replacement is not possible or advisable for a given income type, we can use imputation to draw values from the distribution of responses conditional on the characteristics of the individual and/or household. For the CPS ASEC, research is underway using administrative data to evaluate the assumptions in our imputation models and improve our imputations (Hokayem et al. 2019). The SIPP is using administrative data in their imputation models, building on the imputation work from the SIPP Synthetic Beta file (Benedetto et al. 2013 and Benedetto et al. 2018).

6.3.4 Incomplete PIKing and Errors in PIKing

Similarly, we can treat the fact that not all individuals can be linked to a PIK as a problem of missing information. Here we can test our assumptions in a variety of ways. If we believe that un-PIKed individuals are missing at random (we can control for PIK assignment with observable characteristics), then reweighting or imputation of missing administrative data would be sufficient to produce unbiased income statistics.

However, that would not be the case if we believe that unobservable characteristics are associated with receiving a PIK and the relationship between survey and true income. For example, PIKs are assigned based on a linkage to an SSN or Individual Taxpayer Identification Number (ITIN). Immigration status may be one such characteristic that is imperfectly observed in the survey data. Although it is more likely that a PIK cannot be assigned

to non-citizens, it is possible that un-PIKed non-citizens differ systematically from PIKed non-citizens in ways that would not be captured in our reweighting or imputation models.

Several studies have used census's Master Address File (MAF) to link addresses in administrative records to survey housing units (for example, Bee et al. 2015 and Brummet et al. 2018). This has the potential to help address linkage errors and incomplete linkage, but address linkage also has potential challenges with incomplete linkage and linkage errors, as well. We would like to continue and expand this work.

6.4 Calculate Income Distribution Statistics

As a part of our ongoing research on administrative data and income statistics, we plan to produce a variety of summary measures, including mean and percentiles of individual and household income, inequality statistics (percentile ratios, Gini, Theil, and Atkinson indices, etc.), various measures of poverty (official, SPM, extreme, alternative resource measures), measures of income volatility, various income and resource concepts (official money income, disposable income, with various in-kind benefits included, etc.) and under different assumptions about household equivalence (official, per capita, square-root, etc.).

6.5 Produce Estimates in the Absence of Contemporaneous Data

Of particular importance in this area is how to estimate income statistics in the absence of some administrative data for a given year. This is essential for two reasons. First, we would like to create a consistent time series for years prior to the availability of any given administrative data set. For example, suppose we show with administrative data that a given statistic was biased in the survey data in a given year. We would also like to know if that bias increased or decreased over time. This is complicated by the fact that not all administrative data are available in every year. We have 1040 tax returns going back as far as 1969, although they are not available for each year in between. The DER provides W-2

earnings histories to 1978.[22] Information returns (W-2, 1099-R) are not generally available prior to 2005. State program data, such as for SNAP or WIC, are also not available for all years, especially the further back one goes in time. Creating estimates that correct for survey reporting error in the absence of contemporaneous administrative data is crucial to allow us to understand how statistics such as poverty, median household income, or inequality have changed over time.

The second reason we need to research how to estimate income statistics in the absence of contemporaneous administrative data is to publish timely estimates each year. For example, under our current production schedule, we release the Annual Income and Poverty Report using CPS ASEC data each September. It is likely that some administrative data will not be available in time for use given that schedule. If we want to produce timely estimates, it would be difficult to delay production to wait for administrative data to become available.

To generate these statistics, we plan to develop and test methods to use the data available to us for a given year (or to simulate the data that would be available during production). This could include administrative data from other years as well as other externally available statistics on income, program participation, employment, etc.[23]

6.6 Research Production and Implementation Issues

6.6.1 Disclosure Protection

Administrative data offer the potential to improve income statistics and data quality. However, disclosure protection, already an important issue with survey data, becomes even more crucial with the use of administrative data. This is an issue for both summary tables and public-use microdata samples. This is part of a larger issue facing the Census Bureau in the age of increased computing power and data availability (Abowd and Schmutte 2019).

[22]However, the DER is only linked to CPS ASEC samples from 1973, 1991, 1994, 1996-present and SIPP samples starting in 1984.

[23]For example, this could include state and county employment information, state program benefit statistics, and NIPA estimates on personal income, among others.

6.6.2 Implementation Timeline

Given the ongoing research and open questions about how to best use administrative data, there are several possible options for how administrative data could be incorporated into the production of official statistics. We consider three general approaches to implementation:

1. Staged Introduction - Use administrative data as soon as it is approved for production use and the research supports its use, on an ongoing basis. For example, if we believe we should use payment data from the PHUS and SSR to replace survey responses and have the relevant approvals in a given year, we would go ahead and implement that direct replacement. If, in the next year, we get approval to do the same with interest and dividends from 1040s, we implement the change that year. This has the advantage of improving the estimates and eliminating potential biases as soon as possible. The downside is that without revisions, it would result in frequent breaks as each new use of administrative data would render the estimates from one year not comparable to estimates from prior years. To avoid these breaks in series, we would need frequent revisions to prior data, which would require considerable resources to implement and which may confuse end users.

2. Research Period(s) to Resolve Major Issues - In this scenario, we continue producing official estimates as we do now while researching the various issues and securing approvals for production use of the data. During that period, we can produce research estimates that get revised as the research progresses. With the approvals secured and research sufficiently advanced (perhaps after a period of public consultation), we implement a large set of changes to our production processes. This has the downside of leaving known biases in the data for a longer period, but the upside of creating fewer breaks in series and less need for data revisions.

3. Jump In - Upon receiving approval, we could incorporate administrative records into production without a long research period for each item. For example, we may believe

biases from some types of survey under-reporting sufficiently severe that we are willing to accept some bias from error in administrative data even in the absence of additional research and public consultation on the particular methods used for incorporation of the administrative data. For example, public assistance under-reporting may be considered sufficiently severe that even using just TANF administrative data that does not incorporate all general assistance, results in less biased estimates of income and poverty statistics.

We are not advocating for one timeline over the others in this section, but only trying to highlight possible approaches and their advantages and disadvantages.

6.6.3 Rethinking How Surveys are Conducted and Statistics Estimated

We could combine information from multiple surveys and administrative records to estimate statistics that cannot be estimated from one survey alone. For example, it may not be feasible for the ACS to include the kinds of detailed questions on wealth asked in the SIPP as in Chenevert et al. (2017).

Over the longer term, it may be possible to consider methods for redesigning surveys that take as given the use of administrative records. Surveys could, over time, adapt to more fully leverage their comparative advantages relative to administrative records. For example, we could assess from administrative records available prior to interview (through address linkage) the likelihood that members of a given housing unit will be PIKed and have particular administrative records. We could then use this information to adjust the probability that a given individual is asked particular income questions. This could reduce respondent burden on those that are more likely to have administrative data, maintain the questions for those likely not to have administrative data, and preserve a sample of each group with survey responses for modeling and imputation.

Another possibility is to use administrative information to over-sample subsets of the population, such as those that more likely have income that is not covered by administrative

data or those that are less likely to be assigned a PIK. Similarly, survey questionnaires could focus on capturing information to improve linkage and representativeness, and to cover topics that are difficult to capture in administrative records, such as subjective well-being and informal employment.

6.7 Update Research Files and Estimates over the Course of the Research

As part of this work, we will be creating and updating research data sets that incorporate the results of AIS research. From these data sets, we can create our best estimates of income and poverty statistics that correct for the issues discussed above. Those data sets and estimates will be periodically updated and revised to reflect our advancing knowledge and improved methods.

7 Conclusion

In this paper, we have described the administrative data available for research at this point in time. We elaborated on the challenges and outstanding research issues that need to be addressed as we plan how administrative records could be used in conjunction with surveys to more accurately estimate income statistics. Finally, we laid out our research agenda to do just that.

In light of the growing evidence on survey mis-reporting and increasing non-response rates across all surveys, and to income questions in particular, we are committed to continuing to explore how administrative data can be used to improve our estimates of income and poverty statistics and the microdata from which they are produced. To that end, the Census Bureau is continuing the extensive research program detailed in this paper.

References

Abowd, J. M., McKinney, K. L., and Zhao, N. L. (2018). Earnings inequality and mobility trends in the United States: Nationally representative estimates from longitudinally linked employer-employee data. *Journal of Labor Economics*, 36(S1):S183–S300.

Abowd, J. M. and Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202.

Abowd, J. M. and Stinson, M. H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, 95(5):1451–1467.

Abraham, K. G., Haltiwanger, J. C., Sandusky, K., and Spletzer, J. R. (2018). Measuring the gig economy: Current knowledge and open issues. *National Bureau of Economic Research Working Paper #24950*.

Alvey, W. and Cobleigh, C. (1975). Exploration of differences between linked Social Security and Current Population Survey earnings data for 1972. *Proceedings of the Social Statistics Section, American Statistical Association*.

Auten, G. and Splinter, D. (2018). Income inequality in the United States: Using tax data to measure long-term trends. *Draft subject to change. http://davidsplinter.com/AutenSplinter-Tax_Data_and_Inequality.pdf*.

Bakija, J. (2009). Documentation for a comprehensive historical US federal and state income tax calculator program. `https://web.williams.edu/Economics/papers/bakijaDocumentation_IncTaxCalc.pdf`. Accessed: 2019-06-19.

Bee, A. and Mitchell, J. (2017). Do older Americans have more income than we think? *U.S. Census Bureau SEHSD Working Paper #2017-39*.

Bee, A., Mitchell, J., and Rothbaum, J. (2019). Not so fast? How the use of administrative earnings data would change poverty estimates. *Unpublished U.S. Census Bureau Working Paper*.

Bee, C. A. (2013). An evaluation of retirement income in the CPS ASEC using Form 1099-R microdata. *Unpublished U.S. Census Bureau Working Paper*.

Bee, C. A., Gathright, G. M., and Meyer, B. D. (2015). Bias from unit non-response in the measurement of income in household surveys. *Unpublished U.S. Census Bureau Working Paper*.

Benedetto, G., Stanley, J. C., and Totty, E. (2018). The creation and use of the SIPP Synthetic Beta v7.0. *U.S. Census Bureau Working Paper*.

Benedetto, G., Stinson, M., and Abowd, J. M. (2013). The creation and use of the SIPP Synthetic Beta. *U.S. Census Bureau Working Paper*.

Bhaskar, R., Noon, J. M., O'Hara, B., and Velkoff, V. (2016). Medicare coverage and reporting: A comparison of the Current Population Survey and administrative records. *U.S. Census Bureau CARRA Working Paper #2016-12*.

Bhaskar, R., Shattuck, R., and Noon, J. (2018). Reporting of Indian Health Service coverage in the American Community Survey. *U.S. Census Bureau CARRA Working Paper #2018-14*.

Bingley, P. and Martinello, A. (2017). Measurement error in income and schooling and the bias of linear estimators. *Journal of Labor Economics*, 35(4):1117–1148.

Bollinger, C. R. (1998). Measurement error in the Current Population Survey: A nonparametric look. *Journal of Labor Economics*, 16(3):576–594.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (2018). Trouble in the tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch. *Journal of Political Economy*.

Bollinger, C. R., Nicoletti, C., and Pudney, S. (2013). Two can live as cheaper as one... But three's a crowd. *Working paper*.

Bond, B., Brown, J. D., Luque, A., O'Hara, A., et al. (2014). The nature of the bias when studying only linkable person records: Evidence from the American Community Survey. *U.S. Census Bureau CARRA Working Paper #2014-08*.

Bound, J., Brown, C., Duncan, G. J., and Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3):345–368.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843.

Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, 9(1):1–24.

Brummet, Q., Flanagan-Doyle, D., Mitchell, J., Voorheis, J., Erhard, L., and McBride, B. (2018). What can administrative tax information tell us about income measurement in household surveys? Evidence from the Consumer Expenditure Surveys. *Statistical Journal of the IAOS*, 34(4):513–520.

Burkhauser, R. V., Feng, S., Jenkins, S. P., and Larrimore, J. (2012). Recent trends in top income shares in the United States: Reconciling estimates from March CPS and IRS tax return data. *Review of Economics and Statistics*, 94(2):371–388.

Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2017). Top incomes and inequality in the UK: Reconciling estimates from household survey and tax return data. *Oxford Economic Papers*, 70(2):301–326.

Chenevert, R., Gottschalck, A., Klee, M., and Zhang, X. (2017). Where the wealth is: The geographic distribution of wealth in the United States. *U.S. Census Bureau SEHSD Working Paper #2016-129*.

Chenevert, R. L., Klee, M. A., and Wilkin, K. R. (2016). Do imputed earnings earn their keep? Evaluating SIPP earnings and nonresponse with administrative records. *U.S. Census Bureau SEHSD Working Paper #2016-18*.

Chetty, R. (2012). Time trends in the use of administrative data for empirical research. *Presented at the NBER Summer Institute, July 2012*.

Collins, B., Garin, A., Jackson, E., Koustas, D., and Payne, M. (2019). Has the gig economy replaced traditional jobs over the last two decades? Evidence from tax returns.

Czajka, J. L. and Beyler, A. (2016). Declining response rates in federal surveys: Trends and implications (background paper). Technical report, Mathematica Policy Research.

Duncan, G. J. and Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(4):508–532.

Eggleston, J. and Reeder, L. (2018). Does encouraging record use for financial assets improve data accuracy? Evidence from administrative data. *Public Opinion Quarterly*, 82(4):686–706.

Eggleston, J. and Westra, A. (2019). Incorporating administrative data in survey weights for the survey of income and program participation. *Unpublished U.S. Census Bureau Working Paper*.

Evans, W. N. and Garthwaite, C. L. (2014). Giving mom a break: The impact of higher EITC payments on maternal health. *American Economic Journal: Economic Policy*, 6(2):258–90.

Feenberg, D. and Coutts, E. (1993). An introduction to the TAXSIM model. *Journal of Policy Analysis and Management*, 12(1):189–194.

Fixler, D., Gindelsky, M., and Johnson, D. (2019). Improving the measure of the distribution of personal income. *BEA Working Paper*.

Fixler, D. and Johnson, D. S. (2014). Accounting for the distribution of income in the US national accounts. In *Measuring Economic Sustainability and Progress*, pages 213–244. University of Chicago Press.

Fontenot, K., Semega, J., and Kollar, M. (2018). Income and poverty in the United States: 2017. *U.S. Census Bureau Current Population Reports*.

Fox, L., Rothbaum, J., and Shantz, K. (2019). Correcting for SNAP underreporting: Using administrative records to develop a model for improving reported SNAP receipt in the CPS ASEC. *Unpublished U.S. Census Bureau Working Paper*.

Fox, L. E., Heggeness, M. L., and Stevens, K. (2017). Precision in measurement: Using SNAP administrative records to evaluate poverty measurement. *U.S. Census Bureau SEHSD Working Paper #2017-49*.

Giefer, K., Williams, A., Benedetto, G., and Motro, J. (2015). Program confusion in the 2014 SIPP: Using administrative records to correct false positive SSI reports. *FCSM 2015 Proceedings*.

Gottschalk, P. and Huynh, M. (2010). Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *Review of Economics and Statistics*, 92(2):302–315.

Habib, B. (2018). How CBO adjusts for survey underreporting of transfer income in its distributional analyses. *Congressional Budget Office Working Paper 2018-07*.

Harris, B. C. (2014). Within and across county variation in SNAP misreporting: Evidence from linked ACS and administrative records. *U.S. Census Bureau CARRA Working Paper #2014-05*.

Hokayem, C., Bollinger, C., and Ziliak, J. P. (2015). The role of CPS nonresponse in the measurement of poverty. *Journal of the American Statistical Association*, 110(511):935–945.

Hokayem, C., Raghunathan, T., and Rothbaum, J. (2019). Match bias or nonignorable nonresponse? Improved imputation and administrative data in the CPS ASEC. *Unpublished U.S. Census Bureau Working Paper*.

Hurst, E., Li, G., and Pugsley, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *Review of Economics and Statistics*, 96(1):19–33.

Hyslop, D. R. and Townsend, W. (2018). Earnings dynamics and measurement error in matched survey and administrative data. *Journal of Business & Economic Statistics*, pages 1–13.

Imboden, C., Voorheis, J., and Weber, C. (2019). Measuring systematic wage misreporting by demographic groups. *Unpublished U.S. Census Bureau Working Paper*.

Internal Revenue Service. Filing season statistics. `https://www.irs.gov/statistics/filing-season-statistics`. Accessed: 2019-06-18.

Internal Revenue Service, Research, Analysis & Statistics. (2016). Federal tax compliance research: Tax gap estimates for tax years 2008–2010. *Publication 1415 (Rev. 5-2016)*.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, 84(334):261–289.

Jones, M. R. and Ziliak, J. P. (2019). The antipoverty impact of the EITC: New estimates from survey and administrative tax records. *U.S. Census Bureau Center for Economic Studies Working Paper*.

Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, 25(3):513–551.

Kilss, B. and Scheuren, F. J. (1978). The 1973 CPS-IRS-SSA exact match study. *Social Security Bulletin*, 41:14.

Larrimore, J., Mortenson, J., and Splinter, D. (2019). Household incomes in tax data: Using addresses to move from tax unit to household income distributions. *Journal of Human Resources*.

Layne, M., Wagner, D., Rothhaas, C., et al. (2014). Estimating record linkage false match rate for the Person Identification Validation System. *U.S. Census Bureau CARRA Report Series #2014-02*.

Medalia, C., Meyer, B. D., O'Hara, A. B., and Wu, D. (2019). Linking survey and administrative data to measure income, inequality, and mobility. *International Journal of Population Data Science*, 4(1).

Meijer, E., Rohwedder, S., and Wansbeek, T. (2012). Measurement error in earnings data: Using a mixture model approach to combine survey and register data. *Journal of Business & Economic Statistics*, 30(2):191–201.

Meyer, B. D. and Mittag, N. (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics*, 11(2):176–204.

Meyer, B. D., Mittag, N., and Goerge, R. M. (2018). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *National Bureau of Economic Research Working Paper #25143*.

Meyer, B. D. and Sullivan, J. X. (2012). Identifying the disadvantaged: Official poverty, consumption poverty, and the new supplemental poverty measure. *Journal of Economic Perspectives*, 26(3):111–36.

Meyer, B. D. and Wu, D. (2018). The poverty reduction of social security and means-tested transfers. *ILR Review*, 71(5):1106–1153.

Meyer, B. D., Wu, D., Mooers, V., and Medalia, C. (2019). The use and misuse of income data and the rarity of extreme poverty in the United States. *National Bureau of Economic Research Working Paper #25907*.

Mittag, N. (2019). Correcting for misreporting of government benefits. *American Economic Journal: Economic Policy*, 11(2):142–64.

Monti, H. and Reeder, L. (2016). An introduction to the SIPP Synthetic Beta. *Presented at the Workshop "Conducting Research using the Survey of Income and Program Participation" at the University of Michigan, August 8, 2016*.

Murray-Close, M. and Heggeness, M. L. (2018). Manning up and womaning down: How husbands and wives report their earnings when she earns more. *U.S. Census Bureau SEHSD Working Paper #2018-20*.

Noon, J., Fernandez, L., and Porter, S. (2016). Response error and the Medicaid undercount in the Current Population Survey. *U.S. Census Bureau CARRA Working Paper #2016-11*.

O'Hara, A., Bee, A., and Mitchell, J. (2017). Preliminary research for replacing or supplementing the income question on the American Community Survey with administrative records. *Center for Administrative Records Research and Applications Memorandum Series #16-7*.

Piketty, T. and Saez, E. (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics*, 118(1):1–41.

Piketty, T., Saez, E., and Zucman, G. (2017). Distributional national accounts: Methods and estimates for the United States. *The Quarterly Journal of Economics*, 133(2):553–609.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business & Economic Statistics*, 13(3):305–314.

Pradhan, M. and Ravallion, M. (2000). Measuring poverty using qualitative perceptions of consumption adequacy. *The Review of Economics and Statistics*, 82(3):462–471.

Roemer, M. (2002). Using administrative earnings records to assess wage data quality in the March Current Population Survey and the Survey of Income and Program Participation. *U.S. Census Bureau Center for Economic Studies Working Paper*.

Rothbaum, J. (2015). Comparing income aggregates: How do the CPS and ACS match the National Income and Product Accounts, 2007–2012. *U.S. Census Bureau SEHSD Working Paper #2015-01*.

Rothbaum, J. (2018). Evaluating the use of administrative data to reduce respondent burden in the income section of the American Community Survey. *Unpublished U.S. Census Bureau Working Paper*.

Shantz, K. and Fox, L. E. (2018). Precision in measurement: Using state-level Supplemental Nutrition Assistance Program and Temporary Assistance for Needy Families administrative records and the Transfer Income Model (TRIM3) to evaluate poverty measurement. *U.S. Census Bureau SEHSD Working Paper #2018-30*.

U.S. Bureau of Economic Analysis. Table 7.14 National Income and Product Accounts. `https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=2#reqid=19&step=2&isuri=1&1921=survey`. Accessed: 2019-06-20.

U.S. Census Bureau. Small Area Health Insurance Estimates (SAHIE) Program. `https://www.census.gov/programs-surveys/sahie/about.html`. Accessed: 2019-06-20.

U.S. Census Bureau. Small Area Income and Poverty Estimates (SAIPE) Program. https://www.census.gov/programs-surveys/saipe/about.html. Accessed: 2019-06-20.

U.S. Census Bureau (2017). Proposed information collection; comment request; American Community Survey methods panel tests. https://www.federalregister.gov/documents/2017/11/17/2017-24943/proposed-information-collection-comment-request-american-community-survey-methods-pan Accessed: 2019-07-09.

U.S. Census Bureau (2018). 2017 national survey of children's health, source and accuracy statement. https://www.census.gov/content/dam/Census/programs-surveys/nsch/tech-documentation/source-accuracy-statement/2017-NSCH-Source-and-Accuracy-Statement.pdf. Accessed: 2019-09-10.

U.S. Census Bureau (2019). Survey of income and program participation 2014 panel users' guide. https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2014-SIPP-Panel-Users-Guide.pdf. Accessed: 2019-09-10.

Vilhuber, L. and Abowd, J. (2016). Usage and outcomes of the synthetic data server. *Presented at the Society of Labor Economists Annual Meetings, May 2016*.

Wagner, D. and Layne, M. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records and Research and Applications' record linkage software. *U.S. Census Bureau CARRA Report Series #2014-01*.

Wheaton, L. and Stevens, K. (2016). The effect of different tax calculators on the Supplemental Poverty Measure. *Urban Institute Research Report*.

Zedlewski, S. and Giannarelli, L. (2015). TRIM: A tool for social policy analysis. *Urban Institute Research Report*.

Table 1: Available and Potential Administrative Data for Income and Resource Statistics

Income Item	Administrative Item	Data Source	Notes
Wage and Salaries	Limited W-2 information	IRS, Current Regs	Do not measure gross earnings. (Ex: employee deductions for health insurance) Unreported earnings (tips, etc.) not in tax data
	More Complete W-2 Information	IRS, Not available to Census Bureau	Unreported earnings (tips, etc.) not in tax data
	LEHD (State UI administrative data)	States	Unreported earnings (tips, etc.) not in UI data. Not available for all states in all years. Missing many government employees.
	Detailed Earnings Record (DER)	SSA	Do not measure gross earnings. (Ex: employee deductions for health insurance) Unreported earnings (tips, etc.) not in tax data
Self-Employment (Sole Proprietor/Independent Contractor)	Detailed Earnings Record (DER)	SSA	SSA covered earnings on 1040, Schedule SE (if positive, only). Under-reported income not in tax data (as shown in IRS audit studies)
	1040, Schedule C and SE 1099-MISC	IRS, Not available to Census Bureau	Under-reported income not in tax data (as shown in IRS audit studies)
	1099-K	Not available	Form only exists in recent years
	K-1s	IRS, Not available to Census Bureau	
Self-Employment (Pass-Through)	1040, Schedule E and F, K1s	IRS, Not available to Census Bureau	Income from owners of C-corps not reported unless dividends are taken
Unemployment Compensation	1099-G 1040	IRS, Not available to Census Bureau	
Worker's Compensation		Not available	Mostly administered by private insurance companies. There is some trail of this income in W-2s and 1040s, but it is not generally identifiable as WC
Social Security	Payment History Update System (PHUS)	SSA	Note: CPS ASEC and SIPP only
	SSA-1099	IRS, Not available to Census Bureau	
SSI	Supplemental Security Record	SSA	Note: CPS ASEC and SIPP only This is nontaxable and not on any IRS form
Public Assistance	TANF	States	Not available for all states. Not all cash assistance covered
	TANF	HHS	Not available for all states. Not all cash assistance covered
Veteran's Benefits (Pension, Survivor, Disability Comp)	VA Admin Data	VA	Some benefit data available for limited uses.
Disability Compensation (Except Social Security, VA)	1099-R	IRS, Current Regs	
	1099-R	IRS, Not available to Census Bureau	More detailed information than current regs version
Survivor Income (Except Social Security, VA)	1099-R	IRS, Current Regs	
	1099-R	IRS, Not available to Census Bureau	More detailed information than current regs version
Retirement Income (Except Social Security, VA)	1099-R	IRS, Current Regs	
	1099-R	IRS, Not available to Census Bureau	More detailed information than current regs version
Interest (Excludes tax-preferred accounts)	1040	IRS, Current Regs	Taxable and non-taxable covered
	1099-INT	IRS, Not available to Census Bureau	Taxable and non-taxable covered
Dividends (Excludes tax-preferred accounts)	1040	IRS, Current Regs	
	1099-DIV	IRS, Not available to Census Bureau	
Rent and Royalty Income	1040	IRS, Current Regs	Only gross rent
	1040 Schedule E K1s	IRS, Not available to Census Bureau	Excludes depreciation
Educational Assistance	1098-T 1099-Q	IRS, Not available to Census Bureau	1098-T covers financial aid 1099-Q covers spending from tax-preferred education accounts (529, Coverdell)
Financial Assistance		Not available	Inter-household transfers only available in surveys
Other Income			
Capital Gains	1040	IRS, Not available to Census Bureau	
	1099-B K1s		
Alimony	1040	IRS, Not available to Census Bureau	Both paid and received for filing households only
Gambling Winnings	W-2G	IRS, Not available to Census Bureau	
Alaska Dividend	1099-MISC 1040	Not available	Not identified in Plumley 1099-MISC, likely can be inferred in tax data with reasonable accuracy
Non-Cash/Deferred Compensation			
Employer Contributions to Retirement Plans	Form 5500	Public	Data only available at the aggregate level by firm. We are unaware of any information return or counterparty filing that lists employer contributions to retirement plans at the individual level.
Employer Contributions to health insurance	W-2	Not available	Not present in tax year 2010. Not available in any administrative data currently at census.
Other benefits (moving expenses, etc.)	W-2	IRS, Not available to Census Bureau	
Government Taxes and Transfers			
Federal tax obligations	1040	IRS, Not available to Census Bureau	
State, local, property tax obligations	1040	IRS, Not available to Census Bureau	Depends on itemized deductions, imperfectly observable
EITC	1040	IRS, Not available to Census Bureau	
Various other credits (child tax credit, dependent care credit, education expense credit, etc.)	1040	IRS, Not available to Census Bureau	
Near-Income Items			
SNAP		States	Not available for all states. Availability varies by year
WIC		States	Not available for all states. Availability varies by year
Medicare		HHS	
Medicaid		HHS	
School Lunch		Not available	Survey
Housing Assistance		HUD	Does not cover all possible sources of housing assistance
Energy Assistance/LIHEAP		States	Available for 1 state in some years.

Notes: This table shows the potential sources of administrative data by income/resource item.

Table 2: Possible Uses of Administrative Data for Improving Income Statistics

	Category: Direct Replacement, Survey Only, or Model (Source Data)		
Income Source	Current Data	All Possible Federal and State Data	Notes
Wage and Salary Earnings	Model (W-2, DER, LEHD)	Model (W-2, DER, LEHD)	Unreported "under-the-table" earnings. Current W-2s and DER do not include pre-tax employee contributions to health insurance premiums.
Self-Employment Earnings	Model (DER, Flags for 1040 schedules and 1099s)	Model (1040 Schedules, 1099 forms, K1s)	Under-reported substantially on surveys and in administrative records.
Worker's Compensation	Survey Only	Survey Only	Not available federal administrative data.
Social Security	Direct Replacement (PHUS)	Direct Replacement (PHUS, SSA-1099)	PHUS only available for CPS ASEC and SIPP respondents.
SSI	Direct Replacement (SSR)	Direct Replacement (SSR)	SSR only available for CPS ASEC and SIPP respondents. Not taxable, SSA is only possible source.
Public Assistance	Model (TANF from HHS and state data)	Model (TANF from HHS and state data)	Current data only covers some states. TANF data does not cover all possible cash assistance programs.
Veteran's Benefits	Survey Only or Model	Direct Replacement (VA data)	Some administrative data currently available
Disability, Survivor, and Retirement Income	Direct Replacement (1099-R)	Direct Replacement (1099-R)	
Interest	Direct Replacement (1040)	Direct Replacement (1099-INT)	1040 only covers filers.
Dividends	Direct Replacement (1040)	Direct Replacement (1099-DIV)	1040 only covers filers.
Rent and Royalty Income	Model (1040)	Model or Direct Replacement (1040, Schedule E, K1s)	Currently available 1040 data includes gross, not net, rental income. Unclear how accurately rental income is reported to IRS.
Educational Assistance	Survey Only	Direct Replacement (1098-T, 1099-Q)	No administrative data currently available
Financial Assistance	Survey Only	Survey Only	
Capital Gains	Survey Only	Direct Replacement (1040, 1099-B, K1s)	
Alimony	Survey Only	Direct Replacement (1040)	1040 only covers filers. Perhaps modeling needed for non-filers.
Gambling Winnings	Survey Only	Direct Replacement (1099-G and 1040)	1040 potentially needed for gambling losses.
Alaska Dividend	Survey Only	Direct Replacement (1099-MISC, 1040)	
Child Support	Survey Only	Survey Only	Administrative data potentially available at the state level.
Employer Contributions to Retirement Plans	Survey Only	Survey Only	Could potentially be modeled using aggregate information from Form 5500 filings at the firm level.
Employer Contributions to Health Insurance	Survey Only	Direct Replacement (W-2)	Only present in W-2 from 2011 on.
Other Employee Non-cash Benefits	Survey Only	Direct Replacement (W-2)	
Taxes and Credits	Survey Only	Direct Replacement (1040)	
SNAP	Direct Replacement and Model (States)	Direct Replacement (States)	Modeling necessary for states where data is currently unavailable.
WIC	Direct Replacement and Model (States)	Direct Replacement (States)	Modeling necessary for states where data is currently unavailable.
Medicare	Direct Replacement (HHS)	Direct Replacement (HHS)	
Medicaid	Direct Replacement (HHS)	Direct Replacement (HHS)	
Housing Assistance	Model (HUD)	Model (HUD)	Current data only covers some states. HUD data does not cover all possible housing assistance programs.
Energy Assistance	Model (States)	Model (States)	Current data only covers some states. LIHEAP data does not cover all possible energy assistance programs.

Notes: This table shows the likely use of administrative data given approval for the currently available data ("Current Data") and all data available to federal and state agencies, whether they are currently available to Census for any use ("All Possible Federal and State Data"). "Survey Only" could also require modeling, but that modeling would be primarily from survey responses on that income/resource item and other information (such as program expenditures, eligibility rules, etc.), without the use of administrative data for that item. The current data are not necessarily approved for production (as opposed to research) use.

Table 3: Administrative Income Statistics (AIS) Project Research Agenda

			Year				
		Already in progress	**1**	**2**	**3**	**4**	**5**
1	**Process Administrative Data and Compare to Survey Responses**						
	Subtask	**Description**					
	Document administrative data and standardize (when necessary, especially state data)	Future work in this task will include all income and resource data, but with a particular focus on in-kind benefit data	**X**	**X**	**X**		
	Compare administrative data to survey sources	Evaluate nature of disagreements and observable characteristics (in survey or administrative records) that are associated with the disagreement	**X**	**X**	**X**		
2	**Resolve Practical Issues in Estimation using Multiple Data Sources**						
	Subtask	**Description**					
	Addressing discrepancies among survey and adrec income reports (when we don't believe either are without error)	Assumptions about error structure and appropriate models Estimation of error and model parameters "Imputation" of true underlying income Validating income measures/concepts and variables using external outcomes such as consumption, health, mortality, residential location, etc.	**X**	**X**	**X**	**X**	**X**
	Incomplete geographic coverage	Address locations without available administrative data	**X**	**X**	**X**		
	Incomplete income/resource coverage	Address misalignment of income concepts where they exist		**X**	**X**	**X**	**X**

Notes: This table shows the proposed research agenda of the AIS project along with a prospective timeline for the research.

Table 3 Administrative Income Statistics (AIS) Project Research Agenda, continued

			Year				
		Already in progress	1	2	3	4	5
3	**Leverage Administrative Records to Improve Survey Operations**						
	Subtask	**Description**					
	Survey frame coverage	Evaluate how survey frame coverage compares to administrative universe of individuals and households		X	X	X	X
		Address issues with survey frame					
	Unit non-response bias	Given the survey frame, compare characteristics of non-responding units (and units recorded as ineligible)	X	X	X		
		Address issues with unit non-response bias					
	Item non-response and imputation	Address non-response bias and match bias in survey data	X	X	X		
		Improve imputation quality					
	Incomplete PIKing and errors in PIKing	Evaluate different assumptions about selection into PIK-ing and their impacts on income statistics	X	X	X	X	X
		Evaluate how assumptions on the nature of mistaken linkage affects estimates					
4	**Calculate income distribution statistics**						
	Subtask	**Description**					
	Statistics	Household income (mean, percentiles, etc.)	X	X	X	X	X
		Individual income (mean, percentiles, etc.)					
		Inequality (Gini, P-tile ratios, permanent/transitory, etc.)					
		Poverty (official, SPM, extreme, multiples of poverty line)					
		Income volatility, trajectory					
	Income concepts	Money income (official)	X	X	X	X	X
		Other more expansive measures that include in-kind benefits or value of benefit flows from durable goods					
	Other variations	Household/per capita vs. equivalence-adjusted	X	X	X	X	X
		Absolute vs. relative poverty measures					
		International standards and comparisons					
5	**Produce estimates in the absence of contemporaneous data**						
	Production-like environment	Imputation of administrative records given data in adjacent or nearly adjacent years of information to mimic production schedule		X	X	X	X
	Consistent time-series given limited administrative data in the past	Imputation into years prior to administrative data availability to create consistent time series		X	X	X	X

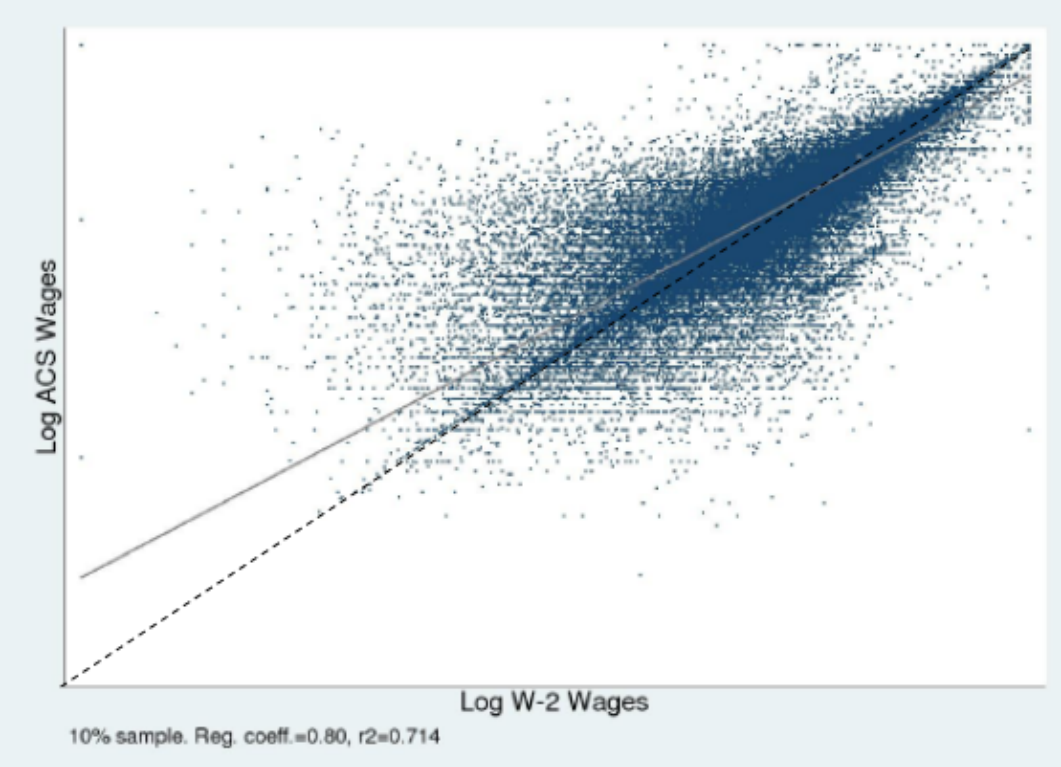
Notes: This table shows the proposed research agenda of the AIS project along with a prospective timeline for the research.

Table 3 Administrative Income Statistics (AIS) Project Research Agenda, continued

			Already in progress	Year 1	2	3	4	5
6	**Research production and implementation issues**							
	Subtask	**Description**						
	Disclosure protection	Official statistics			**X**	**X**	**X**	**X**
		Public-use microdata files						
		Staged Introduction?						
	Implementation timeline	Research Period(s) to Resolve Major Issues?						
		Jump In?						
	Rethinking how surveys are conducted and statistics are estimated	Combine information from multiple surveys and adrecs for more accurate and precise estimates						
		Redesign surveys given adrec use						
7	**Update research files and estimates over the course of the research**							
	Subtask	**Description**						
	Updated files given research results at regular milestones	Create version 1.0, 2.0, etc. files at regular intervals (annually or upon completion of major project milestones)		**X**	**X**	**X**	**X**	**X**
		Publish working papers detailing the updates to the files at each milestone						
		Within those papers, publish research estimates of a set of income and poverty statistics to allow comparisons to official statistics and previous versions of the file						

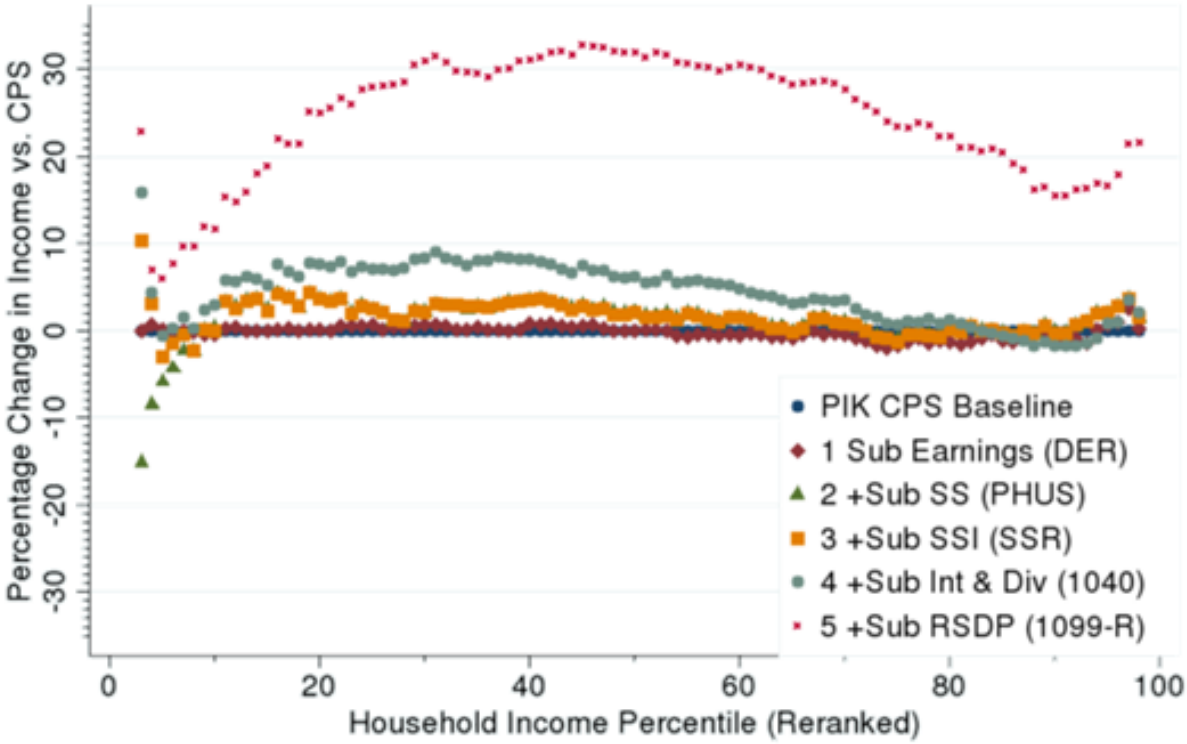
Notes: This table shows the proposed research agenda of the AIS project along with a prospective timeline for the research.

Figure 1: Scatter Plot of Survey and W-2 Earnings



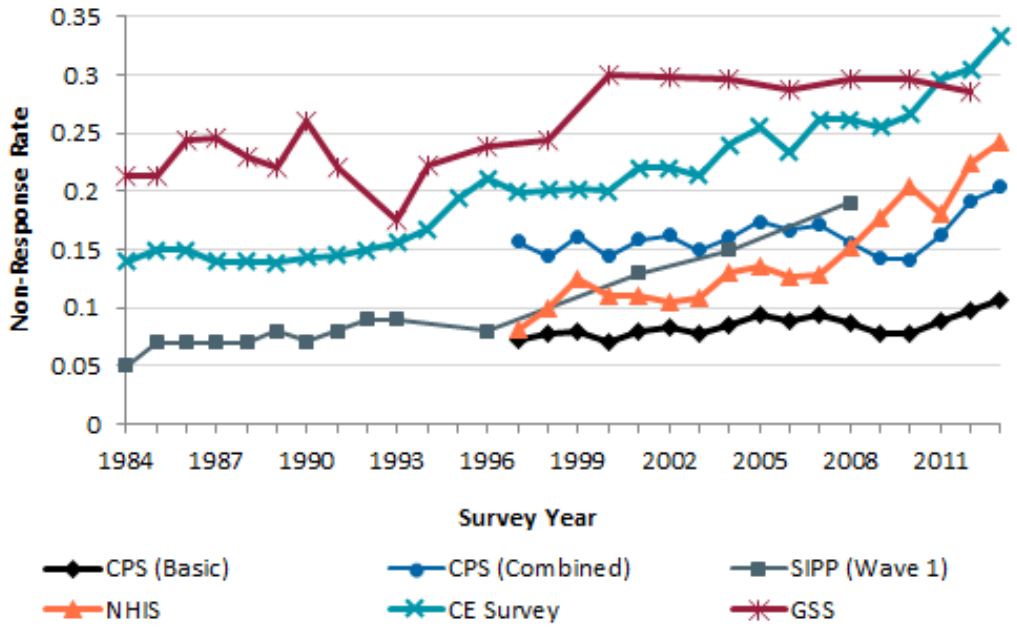
Notes: This figure shows a scatter plot of log earnings in the ACS and W-2s. The 45 degree line is dotted and the regression slope is the solid line.
Source: O'Hara et al. (2017) using the 2011 ACS linked to 2010 W-2 records.

Figure 2: Impact of Direct Replacement on Household Income for those 65 and Older



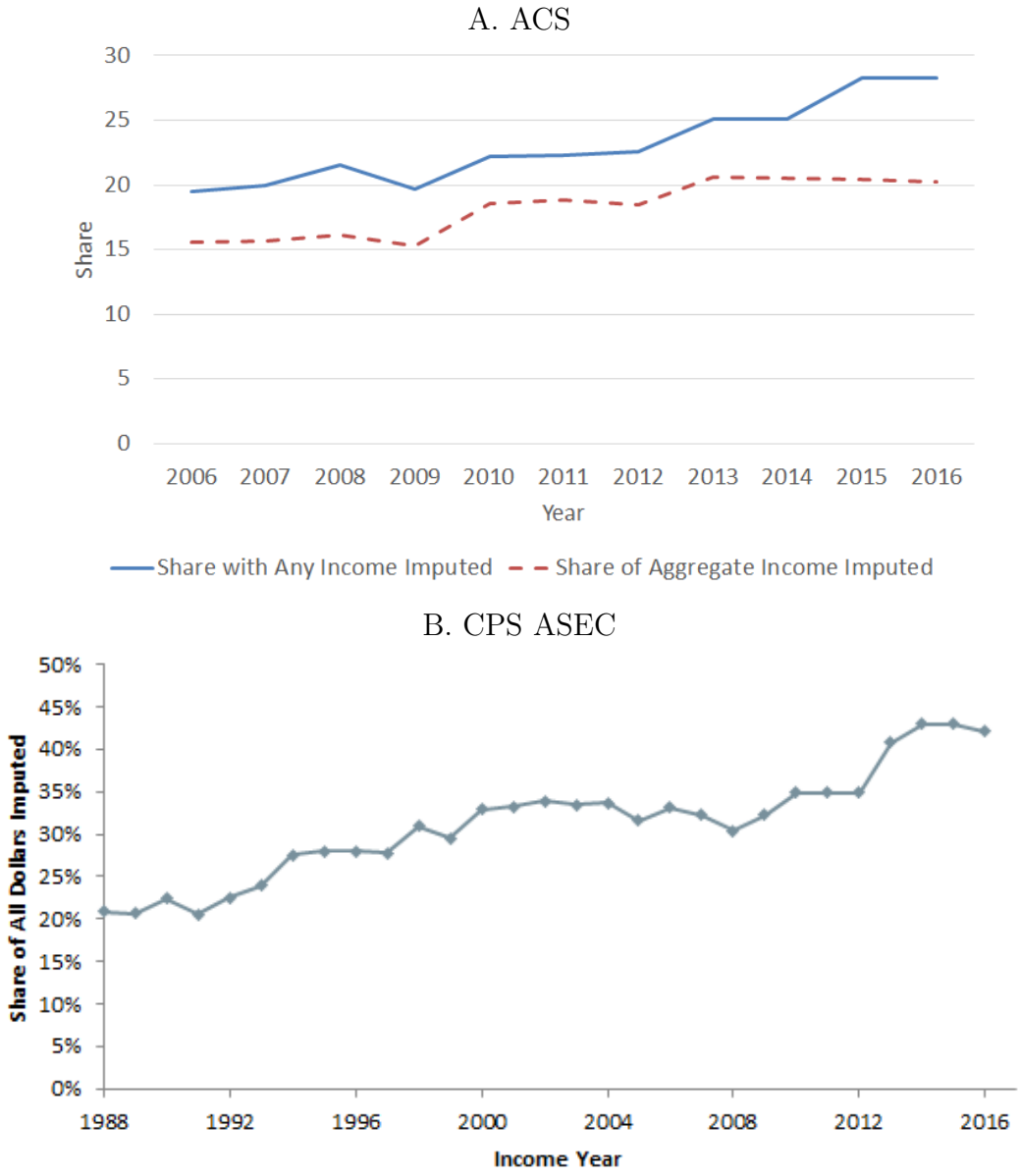
Notes: This figure shows the impact on household income of those 65 and older relative to the survey of using direct replacement for each income item sequentially. Each dot shows the percent difference at a given percentile between the marginal household income distribution with survey items replaced with administrative reports compared to the marginal distribution given the survey responses. The replacements are: 1) earnings with DER earnings (both wage and salary and self-employment), 2) social security with PHUS payment records from SSA, 3) SSI with SSR payment records from SSA, 4) interest and dividends replaced with 1040 interest and dividends (for filers only), and 5) retirement, survivor, and disability income with 1099-R reports from IRS. The replacements are done one at a time and are cumulative, so first earnings is replaced, then earnings and social security are replaced, then earnings, social security and SSI income are replaced, etc.
Source: Bee and Mitchell (2017) using the 2013 CPS ASEC linked to DER, 1040, and 1099-R Information.

Figure 3: Increasing Unit Non-response in Income Surveys



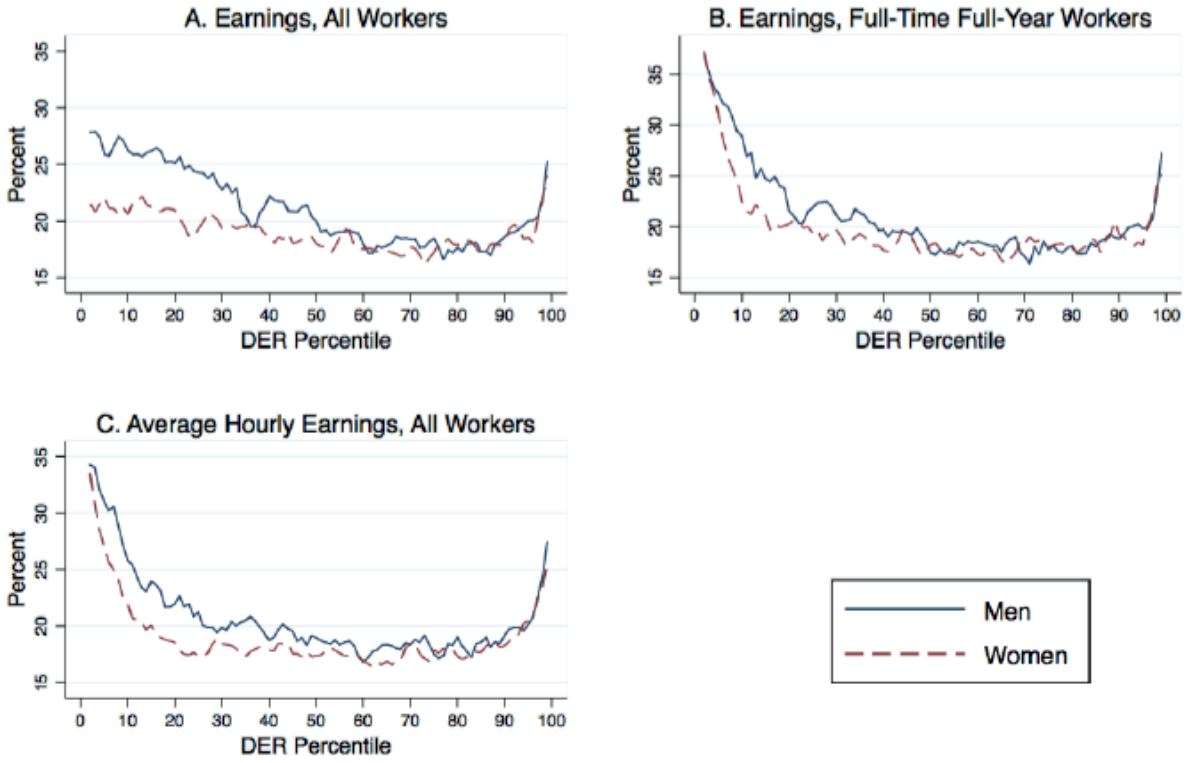
Notes: This figures shows the increasing unit non-response in various income surveys.
Source: Bee et al. (2015).

Figure 4: Increasing Income Imputation in the ACS and CPS ASEC



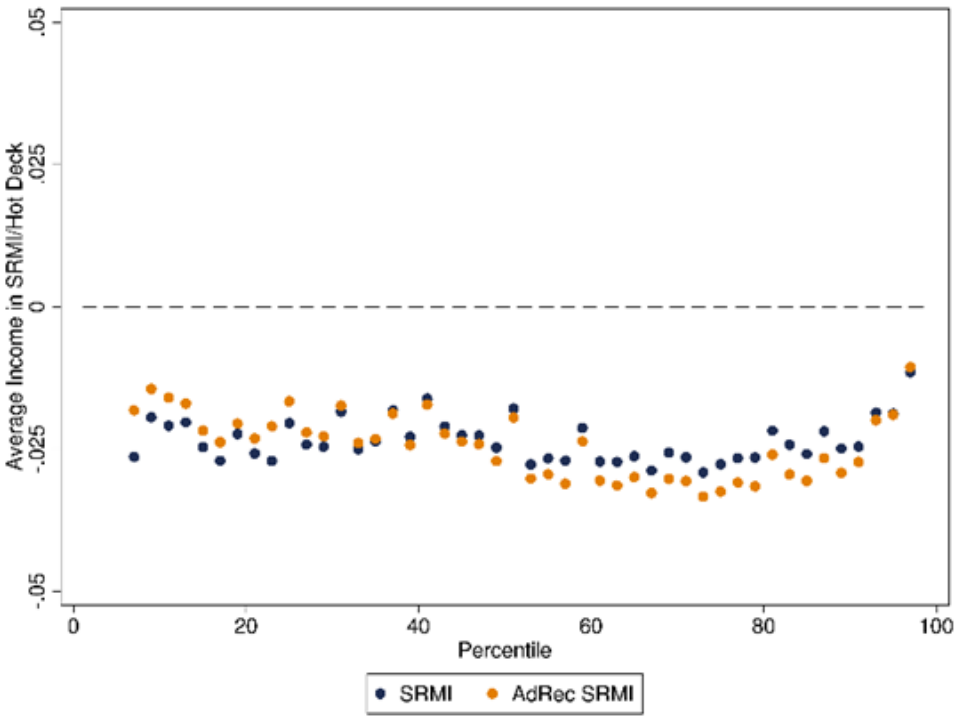
Notes: Panel A shows item non-response rates and share of income imputed in the ACS. Panel B shows the share of all income imputed due to item or supplement non-response in the CPS ASEC since 1987.
Source: For panel A, Rothbaum (2018) using the 2006-2016 ACS. For panel B, Hokayem et al. (2019) using the 1988-2017 CPS ASEC.

Figure 5: Probability of Earnings Item Non-Response by Administrative Earnings



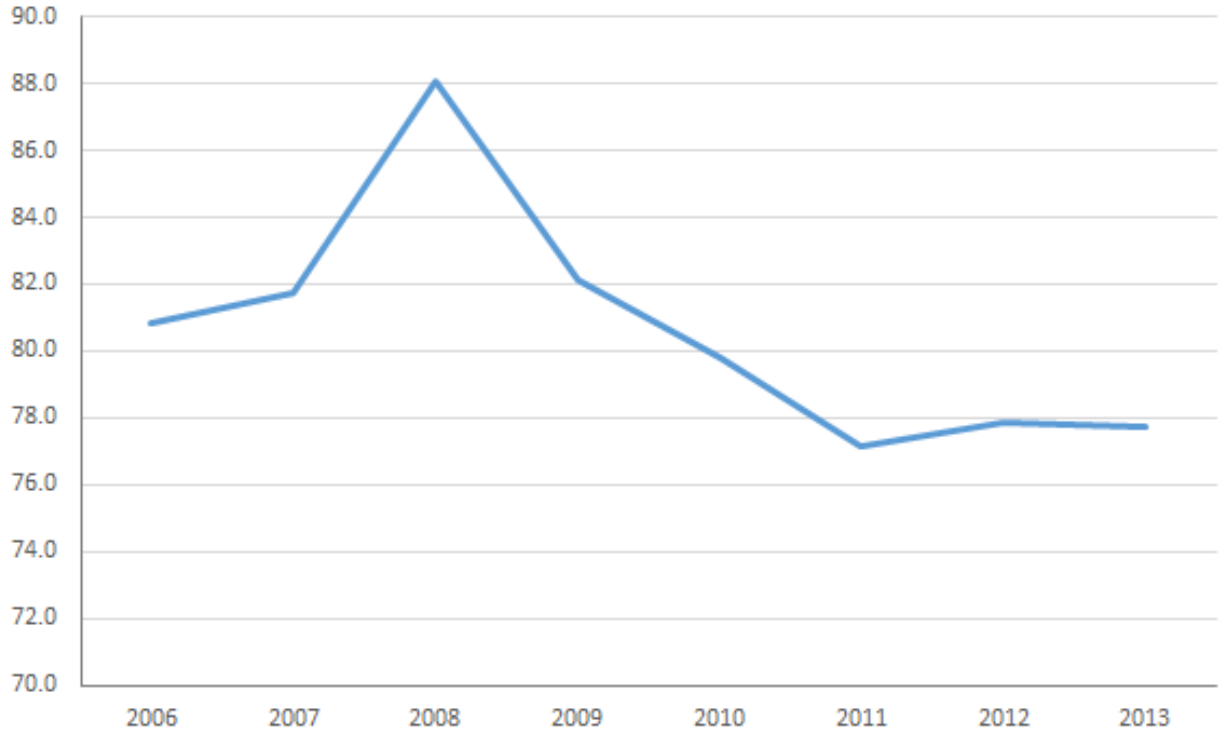
Notes: These panels show 3-point moving averages across percentiles using the overall DER earnings distribution.
Source: Bollinger et al. (2018) using the 2006-2011 CPS ASEC linked to the DER.

Figure 6: Impact of Improved Imputation on Household Income



Notes: This figure shows the impact on an improved imputation model with administrative data (Adrec SRMI) and without administrative data (SRMI). In the figure the dots show the average income by percentile with the improved models relative to the existing imputation model that does not properly account for selection into non-response.
Source: Hokayem et al. (2019) using the 2009 CPS ASEC linked to administrative data, including the DER, 1040s, and 1099-Rs.

Figure 7: Filing Rate over Time for ACS Households



Notes: This figure shows the rate at which at least one individual in an ACS household file taxes (where the householder was linked to a PIK). The Economic Stimulus Act was responsible for the increase in filers in 2008.
Source: Rothbaum (2018) using the 2006 to 2013 ACS linked to 1040s.