National Experimental Wellbeing Statistics

Version 1, released February 14, 2023

SEHSD Working Paper Number 2023-02
CES Working Paper Number 23-04

Adam Bee, Joshua Mitchell, Nikolas Mittag, Jonathan Rothbaum, Carl Sanders, Lawrence Schmidt, and Matthew Unrath

Abstract

This is the U.S. Census Bureau's first release of the National Experimental Wellbeing Statistics (NEWS) project. The NEWS project aims to produce the best possible estimates of income and poverty given all available survey and administrative data. We link survey, decennial census, administrative, and commercial data to address measurement error in income and poverty statistics. We estimate improved (pre-tax money) income and poverty statistics for 2018 by addressing several possible sources of bias documented in prior research. We address biases from (1) unit nonresponse through improved weights, (2) missing income information in both survey and administrative data through improved imputation, and (3) misreporting by combining or replacing survey responses with administrative information. Reducing survey error substantially affects key measures of wellbeing: We estimate median household income is 6.3 percent higher than in the survey estimate, and poverty is 1.1 percentage points lower. These changes are driven by subpopulations for which survey error is particularly relevant. For householders aged 65 and over, median household income is 27.3 percent higher than in the survey estimate and for people aged 65 and over, poverty is 3.3 percentage points lower than the survey estimate. We do not find a significant impact on median household income for householders under 65 or on child poverty. Finally, we discuss plans for future releases: addressing other potential sources of bias, releasing additional years of statistics, extending the income concepts measured, and including smaller geographies such as state and county.

*Send correspondence to census.newsproject@census.gov
Bee: U.S. Census Bureau, charles.adam.bee@census.gov; Mitchell: U.S. Census Bureau, joshua.w.mitchell@census.gov; Mittag: CERGE-EI, nikolas.mittag@cerge-ei.cz; Rothbaum: U.S. Census Bureau, jonathan.l.rothbaum@census.gov; Sanders: U.S. Census Bureau, carl.e.sanders@census.gov; Schmidt: MIT Sloan School of Management, ldws@mit.edu; and Unrath: U.S. Census Bureau, matthew.unrath@census.gov

1 Introduction

Accurately measuring household income and poverty is essential to understanding the nation's overall economic wellbeing. Previous studies suggest that measurement error stemming from unit nonresponse, item nonresponse, and misreporting biases key official statistics such as mean or median income and the official poverty rate. The direction of bias differs among these sources of measurement error. Unit and item nonresponse have been found to bias income up and poverty down (Rothbaum et al., 2021; Rothbaum and Bee, 2022; Bollinger et al., 2019; Hokayem, Raghunathan and Rothbaum, 2022), while misreporting can bias income down and poverty up (Bee and Mitchell, 2017; Meyer et al., 2021*b*; Larrimore, Mortenson and Splinter, 2020). These previous papers document aspects of the overall problem of survey error in isolation, so the overall impact of these sources of error on the accuracy of survey estimates remains unclear.[1] Important next steps are to study the joint impact of these error sources, and to develop a comprehensive solution that addresses all partial problems simultaneously. Doing so would provide survey users with the best possible measure of income.

This paper summarizes the National Experimental Wellbeing Statistics (NEWS) Project, a project to create the most accurate estimates of household income and poverty. The NEWS project makes three unique contributions towards a more comprehensive solution to the problem of measuring income accurately. First, we address as many sources of bias as we can simultaneously, including unit and item nonresponse and underreporting in surveys as well as the various challenges in administrative data such as measurement error, conceptual misalignment, and incomplete coverage. Simultaneously addressing these error sources is crucial, since they have been found to bias key statistics in different directions. Second, we bring together all of the available survey and administrative data in order to overcome the shortcomings of individual data sources. For example, we use 5 different sources of wage and

[1] We discuss these existing approaches and how our methodology compares with them in section 2.4.

salary earnings, each of which capture earnings and jobs not on reported on others. Third, we propose a model to combine survey and administrative earnings data given measurement error in both sources, replacing ad hoc assumptions that have been used in prior work.[2]

To demonstrate the importance of more accurate data, we estimate pre-tax money income and poverty statistics for 2018, mirroring the Census Bureau's annual income and poverty report (Semega et al., 2019). Under our approach, median household income is 6.3 percent higher than the survey-only estimate. The official poverty rate is 1.1 percentage points lower than the survey-only estimate, with 9.4 percent fewer people in poverty.[3] However, these differences vary considerably across groups. Median household income is 27.3 percent higher for householders aged 65 and older, 5.0 percent higher for those aged 55-64, and not statistically different or lower for all other householder ages. Likewise, poverty is 3.3 percentage points lower for persons aged 65 and over (34.2 percent fewer people in poverty), compared to 0.7 percentage points lower for those aged 18-64 (6.7 percent fewer people in poverty), and not statistically different for children 17 and under.[4]

We find that combining survey responses and administrative records matters for the measured income distribution, with different roles played by non-response and misreporting. At the bottom of the income distribution, we find that weighting and imputation augmented with administrative records decreases income at the lowest percentiles of the survey-response only income distribution. This negative shift of the income distribution is more than offset, however, by the additional income that administrative records report relative to surveys. We compare the household income distribution with and without the administrative data and find large effects across the distribution, from 17.1 percent more income at the 10th percentile, to 10.3 percent more at the 25th, 6.8 percent more at the median, and 3.6 percent more at the 75th. As a result, while the survey estimate of household income at the 90^{th}

[2]More detail on the earnings measurement error model will be provided in a forthcoming companion paper, Bee et al. (2023).

[3]All comparisons are statistically significant at the 5 percent level unless otherwise noted.

[4]Estimates are shown in for median household income by subgroup in Table 1 and Figure 1, for poverty by subgroup in Table 2 and Figure 2, and for inequality in Table 3.

percentile is 12.5 times as large as at the 10^{th} percentile, with the NEWS estimates, the ratio is 11.5.

In addition to the substantive differences summarized above, our analyses yield three key methodological takeaways. First, to obtain an improved income measure, it is indeed necessary to simultaneously address error sources such as nonresponse and misreporting. Our combined nonresponse bias corrections (weighting and improved income imputation) generally adjust the point estimates of income down and poverty up.[5] Including administrative wage and salary earnings to address underreporting, particularly when survey-reported earnings are zero, shifts income up and poverty down. Addressing retirement income underreporting (defined benefit pensions and defined contribution withdrawals) has the biggest impact on household income across much of the distribution, echoing findings from Bee and Mitchell (2017). For householders under 55 whose income comes predominantly from wage and salary earnings, which is one of the best reported income sources in surveys, we find limited differences in income and poverty estimates. However, for those 55 and over and particularly for those 65 and over, who have more income in underreported sources (retirement, interest, dividends, etc.), the increase in income due to the underreporting adjustment is greater than the decline in income from the nonresponse bias correction.

A second key takeaway is that each data source has its own strengths and shortcomings, making it difficult to produce accurate estimates of income and poverty when relying only on a single source. As is well-established in the literature, survey data have a number of limitations. For example, over 40 percent of all income is imputed in the CPS ASEC (Hokayem, Raghunathan and Rothbaum, 2022), including 46 percent of wage and salary earnings from a primary job. However, analyses which use administrative data alone are not a panacea. Administrative sources can miss income as well – 5 percent of adults report wage and salary earnings in the CPS ASEC but do not receive a W-2 (Bee, Mitchell and

[5]The differences in this paper are not generally statistically significant, however, as shown in Figure 3 Panel A.

Rothbaum, 2019). Likewise, 7 percent of occupied addresses in the 2018 CPS ASEC cannot be linked to any available source of administrative data. Program eligibility requirements often imply that certain regions or jobs are excluded from the administrative data.

A third key takeaway is that it is critical to incorporate multiple survey and administrative data sources. Using multiple data sources allows us to combine their strengths and thereby reduce the shortcomings we point out above. On the positive side, we find that for some populations, a single data source can yield quite accurate estimates. Yet each single data source also misses or contains substantial error for categories of gross income that are of crucial importance to other subpopulations. Thus, improving measures of income for a wider population requires combining multiple data sources. Overall, we find that a comprehensive approach that leverages the strengths of each data source is required to construct the most accurate estimates of poverty and inequality.

2 Income Measurement Challenges

The major challenge to estimating income is that we do not observe all the information that we would like for all individuals.

2.1 Survey Income

With the survey data, there are several potential sources of missing data and measurement error, such as:

1. **Survey unit nonresponse** - not all individuals or households respond to the survey, which has been found to bias income up and poverty down (Rothbaum et al., 2021; Rothbaum and Bee, 2022).

2. **Survey item nonresponse** - individuals who do respond may choose not to respond to specific questions (a particular problem for income questions), which has been found

to bias income up and poverty down (Bollinger et al., 2019; Hokayem, Raghunathan and Rothbaum, 2022).

3. **Survey mis- and underreporting** - income is not always reported accurately on surveys and can be severely underreported for many income types, which has been found to bias income down and poverty up (Bee and Mitchell, 2017; Rothbaum, 2015). We refer to this as misreporting in the rest of the paper.

As Meyer and Mittag (2021) showed in decomposing bias in estimates of means-tested program benefits, the various sources of measurement error can have biases of different signs and magnitudes across different programs and surveys. Also, correcting for one source of bias without addressing others does not necessarily reduce the overall bias in the estimates.

We address all of these sources of measurement error simultaneously, building on prior work at the Census Bureau that addressed them separately. First, we create improved weights to address survey unit nonresponse (extending Rothbaum et al. 2021 and Rothbaum and Bee 2022). We use imputation to address survey item nonresponse (extending Hokayem, Raghunathan and Rothbaum 2022). We combine survey and administrative data (including replacing survey responses), which also helps address survey item nonresponse as well as survey misreporting (extending Bee and Mitchell 2017).

2.2 Administrative Income

Replacing survey responses with administrative records does not fully address measurement error concerns. Many of the same types of issues in survey data are also present in administrative data, including:

1. **Selection into administrative data** - not all individuals, households, or firms may be present in the administrative data due to how and why the administrative data is collected. For example, many low-income individuals are not required to file a tax return, meaning they may be not represented in tax data. And certain jobs are not

covered by unemployment insurance, meaning those jobholders are not included in commonly used earnings data.

2. **Administrative data "nonresponse"** - some records may be absent from the administrative data that should have been present. For example, although firms are required to file a W-2 for nearly all workers, some may not for a variety of reasons such as firm closure, or paying workers "under the table", etc.

3. **Administrative misreporting** - even when an administrative record exists, it may not be accurate. For example, "under-the-table" earnings, such as unreported tips or underreported self-employment earnings, would result in underreporting in administrative earnings.

4. **Conceptual misalignment** - in some cases the income concept measured by administrative data does not match the concept we would like to measure. For example, the W-2s information received by the Census Bureau do not include information on employee pre-tax earnings used to pay health insurance premiums.[6] For these workers, W-2 earnings are effectively an "underreport" of gross earnings.

5. **Incomplete data coverage** - we may not have access to the data for specific individuals. For example, state-provided data on earnings and means-tested program participation are not available for all states.

These make it inappropriate to rely on administrative data alone. For example, selection into administrative data can exclude subpopulations of interest, such as low-income households which may be underrepresented in tax data. Larrimore, Mortenson and Splinter (2020) created households using addresses from tax filings and information returns to estimate poverty over time, addressing income underreporting in surveys. However, they could not observe individuals and households that did not receive any information return or file taxes.

[6] Starting in 2012, Box DD on the W-2 reports the total cost of the employee's health insurance premium, including the employer and employee contribution. Box DD is not currently available for this work.

Instead, they had to impute the presence and poverty status of an unknown number of individuals per year, which they estimated at 4 to 6 million. Through random sampling from the universe of residential addresses, surveys do not have the coverage gaps we see in administrative data.[7] For example, in the 2019 Current Population Survey Annual Social and Economic Supplement (CPS ASEC), 7 percent of occupied housing units cannot be linked to any administrative or commercial data. But thanks to information from survey responses, we can generate improved weights, imputations, and income measures to better approximate our target universe of individuals and households, even in the absence of administrative data for some.

2.3 Addressing These Challenges

The best estimates of income and poverty would rely on both survey and administrative data. Having different sources of information allows us to address shortcomings in each source. For example, we use 5 separate sources of wage and salary earnings. These are (1) W-2s, (2) the Detailed Earnings Record (DER) file from the Social Security Administration (SSA), (3) Longitudinal Employer-Household Dynamics (LEHD) data reported by firms to state unemployment insurance offices, (4) 1040 tax filings, and (5) survey responses. Survey earnings can help with “nonresponse” in administrative data, as 5 percent of adults report wage and salary earnings in the CPS ASEC but do not receive a W-2 (Bee, Mitchell and Rothbaum, 2019). Some individuals with no W-2s also report wage and salary earnings on their tax returns.

Given the possibility of misreporting in administrative data, we develop a measurement error model for survey and administrative reports of wage and salary earnings. We use that model to replace ad hoc assumptions about when to use survey or administrative earnings given measurement error in both. We discuss that model in Section 4.3.1 and in more detail in a forthcoming companion paper (Bee et al., 2023).

[7] The Master Address File, from which housing units are sampled, is discussed in Section 3.

Likewise, conceptual misalignment in one source can be addressed using information from other sources. For example, while available W-2 data do not include employee pre-tax contributions to health insurance premiums, LEHD earnings for the same job should. Workers with survey-reported private health insurance coverage are 3 to 5 times more likely to have LEHD earnings that exceed the W-2 amounts by 1-3 percent, 3-5 percent, 5-10 percent, and 10+ percent, shown in Table A1.[8]

However, incomplete data coverage makes it more difficult to measure gross earnings in the administrative data. Many jobs are not covered by unemployment insurance, and are excluded from the LEHD – for 2018, there are nearly 20 million more W-2 jobs than LEHD jobs, shown in Table A2.[9]

LEHD and state-provided means-tested program data are also not available for some states. We use imputation to address this source of incomplete data coverage, to correct for underreporting of means-tested program receipt in surveys, and to estimate missing gross earnings (given incomplete LEHD data), extending the work in Fox et al. (2022) and Hokayem, Raghunathan and Rothbaum (2022).

An additional challenge in using linked survey and administrative data is selection into linkage. Linkage rates vary by group, which can bias income estimates that include only linked individuals (Bond et al., 2014), but if unlinked individuals are are also subject to survey measurement challenges above, then income estimates are biased if we measure unlinked

[8]Note that the CPS ASEC variable indicates private coverage, but not necessarily whether that job was the source of that coverage, rather than another job or another individual's job, such as from a spouse, partner, or other family member.

[9]Workers not covered by unemployment insurance include federal employees and those in various private sector occupations For example, Maryland's Department of Labor lists the following jobs as exempt from unemployment insurance: barbers and beauticians, taxicab drivers, owner-operated tractor drivers in certain E and F classifications, maritime employment, election workers, church employees, clergy, certain governmental employees, railroad employment, newspaper delivery, insurance sales, real estate sales, messenger service, direct sellers, foreign employment, other state unemployment insurance programs, work-relief and work-training, family members, hospital patients, student nurses or interns, yacht salespersons who work for a licensed trader on solely a commission basis, services of aliens who are students, scholars, trainees, teachers, etc., who enter the U.S. solely to pursue a full course of study at certain vocational and other non-academic institutions, recreational sports officials, home workers, and casual labor. Refer to `https://www.dllr.state.md.us/employment/empfaq.shtml` accessed 11/1/2022.

individuals' incomes using survey responses only. We use weighting of households with all adults linked conditional on their survey responses to create a representative sample of linked individuals, extending Rothbaum et al. (2021) and Rothbaum and Bee (2022).

2.4 Relationship to Prior Research

This is not the first project to attempt to address shortcomings in survey data to estimate improved income and poverty statistics.[10] There have been several efforts to adjust survey data for underreporting in the absence of linked administrative data, include from the Congressional Budget Office (CBO), Bureau of Economic Analysis (BEA), and the Transfer Income Model (TRIM) from the Urban Institute. In each case, researchers had to make assumptions about underreporting that could not be verified without linked data, such as whether underreporting is on the extensive or intensive margin, which households are more likely to misreport, etc. If those assumptions are not correct, which is likely in the absence of linked data, they risk imputing income and benefits to the wrong individuals and households, introducing biases of unclear direction and magnitude.[11]

[10] There has been considerable work on measurement error in income data, as well as comparing survey income to administrative data. As far back as the 1970's, Kilss and Scheuren (1978) used CPS data linked to data from the Internal Revenue Service (IRS) and Social Security Administration (SSA) to evaluate survey income data. More recent examples include Abowd and Stinson (2013), Bee (2013), Benedetto, Stinson and Abowd (2013), Harris (2014), Bee, Gathright and Meyer (2015), Giefer et al. (2015), Hokayem, Bollinger and Ziliak (2015), Bhaskar et al. (2016), Chenevert, Klee and Wilkin (2016), Noon, Fernandez and Porter (2016), Bee and Mitchell (2017), Fox, Heggeness and Stevens (2017), O'Hara, Bee and Mitchell (2017), Abowd, McKinney and Zhao (2018), Benedetto, Stanley and Totty (2018), Bhaskar, Shattuck and Noon (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, Mitchell and Rothbaum (2019), Bollinger et al. (2019), Imboden, Voorheis and Weber (2019), Jones and Ziliak (2019), Eggleston and Westra (2020), Larrimore, Mortenson and Splinter (2020), Abraham et al. (2021), Eggleston (2021), Larrimore, Mortenson and Splinter (2021), Meyer and Mittag (2021), Rothbaum et al. (2021), Carr, Moffitt and Wiemers (2022), Fox et al. (2022), Hokayem, Raghunathan and Rothbaum (2022), Larrimore, Mortenson and Splinter (2022), McKinney and Abowd (2022), Moffitt et al. (2022), Moffitt and Zhang (2022), Rothbaum and Bee (2022), and others. For a more complete discussion of nonsampling error in income and poverty statistics, refer to Bee and Rothbaum (2019), which also discusses the challenges in addressing these issues and discussed the research agenda that led to this project.

[11] For example, BEA's approach scales up income on the intensive margin in same cases, risking imputing income to accurate reporters rather than for extensive margin misreporting, which is common for retirement income (Bee and Mitchell, 2017) and means-tested program benefits (Shantz and Fox, 2018; Meyer and Mittag, 2019). The CBO model imputes missing income and benefits on the extensive margin conditional on survey characteristics, but underreporting is often not well captured by the observable survey information

Similar work has been pursued under a separate project at the Census Bureau, the Comprehensive Income Database (CID, refer to Medalia et al. 2019), including Meyer and Wu (2018), Meyer et al. (2021*b*), Meyer et al. (2021*a*), and Corinth, Meyer and Wu (2022). A main focus of the CID project has been on addressing misreporting in income and means-tested program benefits. We additionally address nonresponse bias, missing administrative data, and model measurement error in survey and administrative earnings.

3 Data

We would like to use any available data that can help inform estimates of income, resources, or wellbeing, broadly defined. This includes survey and decennial census data collected by the Census Bureau, administrative data, and commercial data. The data could be useful to directly measure resources, to model estimates of resources, to validate measures, to address nonresponse, etc. In this section, we discuss each source of data, also shown in Table 4. Figures 4 and 5 show how we put these data sources together to create the files we use to generate the income and poverty estimates, which we discuss in Section 3.7.

3.1 Survey Data

Surveys collect information on many characteristics of individuals and households that are not available or well-measured in administrative data for all or subsets of the population. These include race, Hispanic origin, tenure (homeownership vs. renting), educational attainment, household composition, and much more. Surveys also include information on income, although we have considerable evidence on misreporting of income on surveys.

(Mittag 2019 and Fox et al. 2022). The TRIM model uses unlinked auxiliary data and program rules to impute missing benefits on the extensive margin. However, Shantz and Fox (2018) and Mittag (2019) show that the underreported program benefits may not be missing from households that appear to qualify for them either through the rules-based imputations or from matching to auxiliary data, with the caveat that income item nonresponse means that household income and program receipt may be less correlated as the regular survey imputations do not condition on administrative program data.

Survey operations also provide information that can be crucial for these estimates. First, major surveys conducted by the Census Bureau are stratified random samples of addresses, in which the occupancy status of housing units (vacant/occupied) is assessed as part of the survey. This provides a sample of units in our target universe, occupied housing units, and their sampling probability. In administrative records, it can often be unclear or even impossible to identify the set of occupied units with no available data – i.e., households and individuals that received no W-2 or other information return and did not file taxes or units with no linked information because they are not the primary residence for a high-income or wealth household. The unobserved units in administrative data may be more likely to be at one end of the income distribution than the other – making their absence particularly problematic when measuring inequality or hardship, such as poverty.

We use data from two household surveys. First, we use the **Current Population Survey's (CPS) Annual Social and Economic Supplement (ASEC)**. The CPS ASEC is an annual survey conducted from February to April each year as a supplement to the monthly CPS. Respondents are asked social and demographic questions, as well as questions about their income and resources in the prior calendar year. CPS ASEC data are available at the Census Bureau from 1967 to the present. In 2019, approximately 95,000 addresses were sampled for the CPS ASEC.[12] It is the source of the official poverty measure produced by the Census Bureau as well as widely cited measures of the household income distribution (Semega et al., 2019). In Version 1, we estimate income and poverty statistics on the 2019 CPS ASEC sample for income in year 2018.

Second, we use the **American Community Survey (ACS)**, which is available from 2005 to the present. The ACS is an ongoing survey of more than 2 million respondent households each year. Respondents are asked similar (although generally less detailed) questions than the CPS ASEC, particularly for income. Additionally, ACS respondents are asked about income

[12]Refer to the CPS ASEC technical documentation at `https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar19.pdf`.

in the prior 12 months, rather than the prior calendar year as in the CPS ASEC.[13] For Version 1 of this project, the ACS provides summary information by geography and occupation that are used in our weighting model and earnings measurement error model.

Both the CPS and ACS use field representatives to assess the occupancy status of housing units, the CPS as part of the Housing Vacancy Survey and ACS for estimates of vacancy rates.[14]

3.2 Other Census Bureau Data

The Census Bureau has other data available on the nation's people and households that we use. First, we use data from the **decennial census**. This includes information on each individual's race, Hispanic origin, and age.

We also use information from the **Master Address File File (MAF)**.[15] The MAF contains continuously updated information of all known living quarters in the United States. The MAF is used to select housing units for inclusion in household surveys, including the CPS and ACS, as well as for decennial census operations. The MAF also includes housing unit characteristics, such as whether addresses are in single-family or multi-family units.

We also use the **Master Address File Auxiliary Reference File (MAF-ARF)** which links addresses in the MAF to individuals who reside there in each year. The MAF-ARF is constructed from multiple administrative data sources, including from the IRS, Department of Housing and Urban Development (HUD), and the U.S. Postal Service, among others.

Each of these other Census Bureau data sources provide information that can help us address nonresponse bias and better estimate income and poverty statistics on representative samples

[13] ACS technical documentation is available at `https://www.census.gov/programs-surveys/acs/technical-documentation.html` and `https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_general_handbook_2020.pdf`.

[14] Refer to `https://www.census.gov/topics/housing/guidance/vacancy-fact-sheet.html` for a discussion of housing vacancy estimates in the Housing Vacancy Survey (from the CPS), ACS, and American Housing Survey.

[15] The specific file we use is the MAF extract file, or MAFx.

of individuals, families, and households.

3.3 Federal Administrative Data

The federal government data we use are provided primarily by the IRS and Social Security Administration (SSA). The Census Bureau also has an agreement with the Department of Health and Human Services (HHS) for data on the Temporary Assistance for Needy Families (TANF) program from some states. That data will be discussed in Section 3.4, as TANF data are also shared with the Census Bureau by individual partner state agencies.

3.3.1 IRS Data

From the IRS, we have the following data:

1. the **Information Return Master File (IRMF)** from 2005 to the present,
2. the universe of **Form 1099-R** returns on "Distributions From Pensions, Annuities, Retirement or Profit-Sharing Plans, IRAs, Insurance Contracts, etc." from 1995 to the present,
3. the universe of **Form W-2** returns on "Wage and Tax Statement" for all W-2 covered jobs from 2005 to the present, and
4. the universe of **Form 1040** tax filings every five years from 1969 to 1994, 1995, and then each year from 1998 to the present.

The IRMF includes an indicator for each individual that received one of several information returns in a given year as well as their address, including for Forms 1098, 1099-DIV, 1099-G, 1099-INT, 1099-MISC, 1099-R, 1099-S, SSA-1099, and W-2. The IRMF allows us to link individuals to their addresses and is used in constructing the MAF-ARF. The IRMF does not include any information on income amounts.

The 1099-R extracts provided by the IRS include information on amounts of defined-benefit pension payments (including survivor and disability pensions) and withdrawals from defined-contribution retirement plans. These extracts exclude 1099-R records corresponding to direct rollovers between accounts.

The W-2 extracts provided by the IRS include select W-2 boxes, 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 1040 extracts provided by the IRS include information on tax-unit wage and salary income, gross rental income, gross Social Security income, taxable and tax-exempt interest income, dividends, Adjusted Gross Income, and a constructed measure of Total Money Income (TMI). TMI is the sum of taxable wage and salary income, interest (taxable and tax-exempt), dividends, gross Social Security income, unemployment compensation, alimony received, business income or losses (including for partnerships and S-corps), farm income or losses, and net rent, royalty, and estate and trust income.[16] The 1040 also includes information on marital status through filing status and filer information and identifies up to four dependents.

We use IRS data to address nonresponse bias and measurement error.

3.3.2 Social Security Administration (SSA) Data

From the SSA, we use the following data:

1. the **Numerical Identification System (Numident)** file,

2. extracts from the **Detailed Earnings Records (DER)**.

3. several files from the **Payment History Update System (PHUS)**, and

[16]Prior to tax year 2018, TMI also included total pensions and annuities. However, this was removed from TMI due to a change to income reporting on the Form 1040 and the regulations regarding data sharing between IRS and the Census Bureau.

4. several files from the **Supplemental Security Records (SSR)**.

The Numident contains information on any individual to ever receive a Social Security Number (SSN), including their sex, date of birth, date of death, information on their citizenship status, and their location of birth.

The DER contains job-level W-2 information that generally corresponds to the data provided by IRS, but with the potential for additional cleaning and error correction from SSA as part of their administration of the Social Security system. The DER also includes Social Security covered self-employment earnings reported on the Form 1040 SE (if at least $400). Like many SSA data sets, including some PHUS and SSR files, the DER is only available for linked respondents from specific surveys and years.[17]

The PHUS contains monthly Old Age, Survivors, and Disability Insurance (OASDI) program payment information from 1984 to the present. There are several PHUS files available to the Census Bureau. One set of PHUS files includes OASDI recipients in 2020 and 2021, with one record per address. There are also PHUS files for linked respondents from specific surveys and years.

The SSR contains monthly Supplemental Security Income (SSI) payments for both federal SSI payments and state payments administered by the SSA, from 1984 to the present. One set of PHUS files includes SSI recipients in 2020 and 2021, with one record per address. There are also SSR files for linked respondents from specific surveys and years.

We use the survey-linked SSA data (DER, PHUS, and SSR) to address item nonresponse bias and measurement error. The Numident and address-level SSA data (PHUS and SSR) are useful for weighting to address nonresponse bias.

[17] Specifically, the DER includes respondents with an assigned Protected Identification Key (discussed in Appendix A) who can be linked to the Numident from the CPS ASEC in 1973, 1979, 1981-1991, 1994, and 1996-present, the Survey of Income and Program Participation (SIPP) in 1984, 1990-83, 1996, 2001, 2004, 2008, 2014, and 2018-present, and the ACS in 2019.

3.4 State Administrative Data

We use several data sets shared with the Census Bureau by state government agencies:

1. the **Longitudinal Employer-Household Dynamics (LEHD)** files,
2. data on **Supplemental Nutrition Assistance Program (SNAP)** participation, and
3. data on **Temporary Assistance for Needy Families (TANF)** program participation.

3.4.1 LEHD

Under the LEHD program, states provide data on wage and salary earnings reported by firms for the administration of the unemployment insurance (UI) program. Firms report gross earnings to UI offices, so the LEHD should include non-taxable earnings that are not reported on a Form W-2 for the same job such as pre-tax employee contributions for health insurance premiums. However, coverage in the LEHD data we use is not complete, as many government employees (such as federal civilian employees, postal workers, and Department of Defense employees) are not covered by state UI benefits. Furthermore, some private-sector employees, including those employed by religious organizations, are not covered by UI, and are therefore not present in the LEHD data. Finally, data sharing agreements between a state and the Census Bureau are not always available, resulting in LEHD earnings missing for all jobs in specific states and years.[18]

LEHD data are useful for addressing nonresponse bias and misreporting.

[18]More information on the LEHD program and data is available at `http://lehd.ces.census.gov/data/lehd-snapshot-doc/latest/`, accessed 12/16/2022. While the LEHD program does receive data from the Office of Personnel Management (OPM) for many federal employees, those data are not part of the more recent years of data in the LEHD Interleave file used in this project.

3.4.2 SNAP

The Census Bureau has agreements with many states to receive data on SNAP participation, although the available states vary by year.[19] The SNAP data includes benefits received for each case as well as the individual members recorded in that SNAP case.

SNAP data are useful for addressing misreporting of other income items. SNAP is not included in money income, but these data will be used to address misreporting of in-kind benefits in future releases.

3.4.3 TANF

The Census Bureau also has agreements with many states to receive data on TANF participation. In addition to the state agency data, the Census Bureau also has data on TANF cash assistance receipt from HHS. As with SNAP, the available states vary by year.[20] TANF data are also available by case (benefit amounts) with individuals in each TANF case recorded as well.

TANF data are useful for addressing misreporting.

3.5 Commercial Data

We use information on home values from Black Knight, a third party aggregator of property tax records, which can be useful in correcting for selection into nonresponse on surveys.[21]

These data are useful for weighting to address nonresponse bias.

[19] For example, SNAP data are available for 17 states in 2018, 20 states in 2014, 16 states in 2010, and 6 states in 2006. In 2018, the states with available SNAP data are Arizona, Connecticut, Florida, Hawaii, Idaho, Indiana, Kentucky, Maryland, Mississippi, Montana, Nevada, New Jersey, New York, North Dakota, Tennessee, Utah, and Wyoming.

[20] TANF data are available for 36 states in 2018, 37 states in 2014, 36 states in 2010, and one state in 2006.

[21] Chapin et al. (2018) evaluated the use of similar data from CoreLogic in ACS production and discuss some strengths and limitations of this kind of data. One limitation is that the coverage varies by location.

3.6 Firm Data

We also use data on firm characteristics from the Longitudinal Business Database (LBD), which is described in Chow et al. (2021). The LBD contains establishment-level information on firm employment and payroll. The LBD is constructed from other data sources at the Census Bureau, including the Business Register (BR), that are constructed using data from the IRS and surveys of businesses, including the Economic Census.

Firm data are useful for addressing nonresponse bias, because they help predict survey responses. They can also be used to address misreporting when there is measurement error in both survey and administrative data, since firm information might help us diagnose error in both data sources.

3.7 Linkage and File Construction

To make use of all of this data, we link them to create two main files: (1) the Address File and (2) the Person File, with linkages made at the following levels:

- Individual - using Protected Identification Keys (PIKs),
- Address - using Master Address File identifiers (MAFIDs),
- Job - using PIKs and Employer Identification Numbers (EINs) and by the job matching procedure described below,
- Firm - using the LBD firm identifiers (LBDFID) and Employer Identification Numbers (EINs), and
- Geography - by state, county, and census tract.

The data linkage process for the individuals and addresses is straightforward. We match observations using unique identifiers attached to each person (PIK) and address (MAFID) in each file. The assignment of these identifiers is discussed in Appendix A. To link a survey respondent to any administrative data, we must assign that respondent a PIK using the personally identifiable information (PII) on the survey. If a survey respondent is not assigned a PIK, they cannot be linked to *any* administrative data.

As discussed in Section 2, we have many sources of wage and salary earnings information. Three of them are available at the job level – W-2s, the DER, and LEHD. However, linking LEHD and W-2 jobs is not trivial.[22] In the simplest case, a firm files a W-2 and reports the job to the UI office with the same EIN. We can link these "direct matches" by PIK and EIN. However, some firms do not file their W-2s and UI reports under the same EIN. We use individual and job-level information from the universe of W-2 and LEHD jobs to create indirect matches of firm identifiers across datasets. We discuss this process in detail in Appendix A.3 with an example in Figure A1.

After direct and indirect linkage, of the 264 million jobs, we find 82 percent of jobs matched directly by PIK-EIN, 6 percent matched indirectly, 10 percent unmatched from W-2s, and 3 percent unmatched from the LEHD (shown in Table A2). We use this linked job information to better estimate gross earnings at the job and person level for use in our income estimates.

Because firms do not necessarily correspond to unique EINs, we use information from the redesigned Longitudinal Business Database (LBD) to link workers (through EINs in the job data) to unique firms (Joint Committee on Taxation, 2022; Chow et al., 2021), which we discuss in Appendix A.4.

We create the Address File by linking the sample of occupied (non-vacant) housing units in the survey to the aforementioned sources of administrative, survey, census, and commercial data, as shown in Figure 4. By starting with addresses, we have information from all occupied units, including respondents *and* nonrespondents. In the address file, we do not use any information from survey responses other than whether the unit responded. This file is used

[22] As the DER is sourced from W-2s, linking DER and W-2 jobs is generally simple.

with the Person File to construct the weights that address selection into our sample and selection into linkage, issues discussed in Section 2.

We then create the Person File by linking survey respondents to administrative data, as shown in Figure 5. In combination with the weights created using this and the Address File, the Person File is used for all of the subsequent steps in generating the income and poverty estimates.

The Address and Person Files are discussed in more detail in Appendix B.

4 Methodology

In this section, we describe the steps needed to take the data described in Section 3 through to estimating income and poverty statistics, shown in Table 5. We have categorized the steps into three groups: (1) weighting, (2) imputation, and (3) estimation.

4.1 Weighting

Our analysis sample is the set of households that respond to the CPS ASEC with all survey-adults assigned a PIK.[23] We use weighting to address several measurement challenges discussed in Section 2, particularly survey unit nonresponse and selection into linkage. Weighting is particularly useful when all of the information is missing for a subset of units – in our case we have no survey information for nonrespondents and no administrative information for individuals that cannot be assigned a PIK.

To address survey unit nonresponse, we use information from the linked administrative and decennial census data which is not observed in the survey. This information is available for all linkable households regardless of whether they responded, as is the geographic summary information. We weight respondent households so that the weighted estimates for these

[23] We define survey-adults as those 15 and over as the survey income questions are asked for all individuals 15 and over in a household.

linked characteristics match the estimates obtained using all occupied households given their sampling probability in the CPS while the person-level weights also match to external population controls by state. This should address survey unit nonresponse, following prior work in the ACS (Rothbaum et al., 2021) and the CPS ASEC (Rothbaum and Bee, 2022).

To address selection into linkage, we extend that work by estimating statistics from survey responses in the respondent sample and reweight households with all adults linked (our analysis sample) so that the weighted estimates from analysis sample simultaneously match: (1) the linked administrative characteristics from the sample of occupied units, (2) the survey-response estimates from the respondent sample, and (3) the external population controls by state. This step should address selection into linkage, extending the prior work that was focused only on survey estimates and survey unit nonresponse.

Weighting also helps address selection into administrative data and administrative data nonresponse. The survey frame contains geographic summary information at the address level for each occupied household and survey responses for respondent households that we cannot link to administrative data, whether at the individual or address level.

For a more complete discussion of weighting, including the underlying assumptions, implementation details, and statistics validating the model, refer to Appendix C.

4.2 Imputation

Many of our measurement challenges are not the result of blocks of information missing completely for defined subsets of observations. For example, an individual that does not respond to the survey earnings question (46 percent of all workers) or has a missing LEHD job may have all the other information (e.g., other survey responses, W-2 job earnings, etc.) that we need to estimate income and poverty. For these measurement challenges, imputation is a better approach to fully utilize the information that is available (Raghunathan et al., 2001).

There are four sets of variables that we impute:

1. Survey earnings,
2. LEHD job-level gross earnings,
3. Means-tested program benefits (TANF and SNAP), and
4. Administrative income for tax nonfilers in certain categories (unemployment compensation, interest, and dividends)

In the 2019 CPS ASEC, 46 percent of individuals with earnings in the survey had their primary job earnings imputed.[24] We impute earnings for these individuals (and the individuals with missing earnings from other jobs/employers) conditional on the survey and linked administrative data. These imputed values reflect the distribution of differences between survey and administrative earnings, conditional on the observed information. This allows us to address potential measurement error in administrative earnings for survey nonrespondents.

Likewise, we are missing LEHD job-level gross earnings for 8 percent of individuals' highest earning job.[25] There are additional jobs where W-2 earnings exceed LEHD earnings or the disagreement between them is sufficiently large that we impute gross earnings out of concerns about data quality. As discussed in Section, 2, we would like gross earnings from all jobs because of the conceptual misalignment between available W-2 earnings and the gross earnings we would like to measure. However, gross earnings is not available because of incomplete data coverage (some states missing from the LEHD), selection into administrative data (some jobs not covered by unemployment insurance and thus missing from the LEHD), administrative data "nonresponse" (missing jobs in the LEHD that should be present), and administrative data misreporting.

Following Fox et al. (2022), we also use imputation for missing means-tested program benefits

[24]Refer to Table 6 for rates of missing data for imputed income items.
[25]If we order jobs from highest to lowest earning in the job-level administrative data.

due to incomplete data coverage.

Finally, we impute specific administrative income items for individuals that do not file taxes using parameters estimated on more detailed data by Rothbaum (2023). 85 percent of survey-adults can be linked to a 1040 tax filing (refer to Table A4). For those individuals, the Total Money Income measure includes many income items that are underreported on surveys such as unemployment insurance compensation, interest, and dividends, even if not all items are available separately. However, we observe only whether non-filers received several information returns, including Forms 1099-G, 1099-INT, and 1099-DIV in the IRMF. From these we have information on whether they received UI compensation, interest income, and dividends, respectively. Each of these income sources are significantly underreported on surveys (Rothbaum, 2015). Rothbaum (2023) worked with more detailed data available under a separate agreement between the Census Bureau and IRS, for limited use. In that data, the 1099-G, 1099-INT, and 1099-DIV data are available, including income amounts. Rothbaum (2023) released coefficients that can be used to impute these amounts for nonfilers conditional on survey responses and the administrative data used in this project. We use that information to impute these underreported income items for nonfilers. This imputation addresses selection into administrative data (tax filing) and survey misreporting of these specific income types.

For a more complete discussion of imputation, including the underlying assumptions, implementation details, and statistics on the imputed values, refer to Appendix D.

4.3 Estimation

With the Person File, weights, and imputations, we have complete data for all the inputs used in the NEWS estimates. The final step in processing is putting that data together to estimate income and poverty.

4.3.1 Earnings Measurement Error Model

Earnings represent 80 percent of all income (Rothbaum, 2015). Measurement error in survey and administrative earnings, therefore, merits particular attention.[26]

Although survey wage and salary earnings are relatively well reported when compared to external benchmark aggregates (Rothbaum, 2015), work with linked microdata has identified systematic differences between administrative records and survey responses.[27] This work has 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. There is also extensive margin disagreement between survey and administrative records – about 10 percent of working-age individuals have earnings in one data source but not the other (Bee, Mitchell and Rothbaum 2019).

Some papers in the survey misreporting literature assumed the administrative records were free of error (Bound and Krueger 1991, Bound et al. 1994, Pischke 1995, for example).[28] 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 in which both survey and administrative reports for a given job may contain error. Under their approach, "true" earnings are 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. We

[26] Some of the discussion in this section follows Bee and Rothbaum (2019) closely.

[27] Alvey and Cobleigh (1975), Duncan and Hill (1985), Bound and Krueger (1991), Bound et al. (1994), Pischke (1995), Bollinger (1998), Bound, Brown and Mathiowetz (2001), Roemer (2002), Kapteyn and Ypma (2007), Gottschalk and Huynh (2010), Meijer, Rohwedder and Wansbeek (2012), Abowd and Stinson (2013), Murray-Close and Heggeness (2018), Bee, Mitchell and Rothbaum (2019), Imboden, Voorheis and Weber (2019), Jenkins and Rios Avila (Forthcoming), and many others have studied wage and salary earnings.

[28] In some cases, the authors restrict their analysis to a subset of workers for which 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.

do not assume that measurement error is only present in surveys. 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). In the absence of a "truth set" of data, it is an open question how much of this disagreement is due to misreporting on surveys or measurement error in the administrative data.[29]

We have several separate reports of administrative earnings. In Table 7, we show summary statistics on the number of individuals assigned a PIK with any wage and salary earnings reported from all possible combinations of W-2s, the DER, and the LEHD. We also show the probability that survey respondents report non-zero survey earnings for each combination of administrative wage and salary sources. The vast majority of individuals with earnings in one source have earnings in all three.[30]

From the three separate administrative job-level wage and salary earnings sources (including gross earnings imputed as discussed in Section 4.2), we construct our job-level estimate of gross earnings. We aggregate these job-level earnings to estimate total administrative wage and salary earnings for each individual. This gives a measure of total administrative wage and salary earnings (y_a), which we then use in the model with our final post-imputation total survey wage and salary earnings (y_s) discussed in Appendix D.

[29] Compounding the challenge, it is not always the case that different sources of administrative data agree. Bee, Mitchell and Rothbaum (2019) found 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.

[30] Table 7 also has information on how the W-2 earnings information available in the DER differs from the IRS W-2 information. In Panel B, we focus on individuals we can and cannot link to the Numident (a proxy for having a valid SSN). If individuals have W-2 and DER earnings, they are basically always present in the Numident and are very likely to report wage and salary earnings in the survey (87 percent). However, if individuals are in the Numident and have W-2 earnings, but no DER earnings, then they are very likely *not* to report wage and salary earnings in the survey. This suggests that there is measurement error in the W-2 file for these cases that is not in the cleaned, SSA-provided DER data. We therefore default to the DER information in these cases of no job-level administrative earnings. However, if individuals are not in the Numident and have W-2 earnings, but no DER earnings, they are very likely to report wage and salary earnings on the survey (85 percent). In these cases, we conclude the DER is missing earnings for those without SSNs that are correctly present in W-2s. For these individuals, we default to the W-2 information of positive job-level earnings. This is a clear example of how administrative data are not necessarily free of error and different sources of administrative data covering the same concept (wage and salary earnings) from the same tax information do not necessarily agree.

The survey and administrative earnings can differ on the extensive or intensive margin. With extensive margin disagreement, where earnings are present in one but not both sources, we default to the earnings report that is non-zero. In other words, we assume that any survey report in the absence of administrative earnings reflects under-the-table income or a reporting or linkage issue in the administrative data. We also assume that any administrative earnings without a corresponding survey earnings report reflect under-/misreporting on the survey. These are both assumptions that we plan to examine in future work.

The other difference we observe is intensive margin differences in reporting, where the reported values are not equal. Figure 6 shows a scatterplot of survey versus administrative reports of wage and salary earnings.[31] Several important features of the data are visible in the figure. First, survey and administrative earnings generally agree, reflected in the clustering around the 45° line. However, regressing survey on W-2 wage and salary earnings (in logs) yields a slope of 0.8, which is consistent with mean reversion in survey earnings reports.[32]

In our forthcoming companion paper, Bee et al. (2023) define a model that parameterizes the measurement error in y_a and y_s relative to the unobserved true earnings (y) for intensive margin disagreement. We provide a concise summary of the model here.

Since there can be measurement error in both survey and administrative earnings reports and we do not have data on "true" earnings for anyone, we must impose assumptions on the data that are untestable or can only be tested indirectly. For example, we believe that administrative earnings could be underreported either because some income is missing (such as some portion of tips) or some jobs may be missing. Likewise, we do not assume that administrative earnings are free of classical measurement error, or noise, even if we believe that noise may be of lower variance than the noise in survey earnings reports.

[31] The figure is reproduced from O'Hara, Bee and Mitchell (2017) as more recent disclosure rules limit the possibility of releasing such detailed information of individual survey and administrative earnings values.

[32] For example, if we assumed no measurement error in W-2 earnings, then a slope that is less than one could indicate mean-reverting error non-classical measurement error in survey responses.

These assumptions provide some structure to our earnings measurement error model. The model setup consists of two earnings measures: (a) survey earnings, which are conditionally unbiased but have potentially downward-biased conditional variances, and (b) administrative earnings records, which can be conditionally biased but have accurate conditional variances.[33]

While these assumptions on survey versus administrative records are not directly testable, they were chosen to be both consistent with prior literature on measurement error in earnings and to be consistent with previous measurements of average income. Under our assumptions, the survey would be unbiased for average income measures but may have trouble accurately assessing income in the tails of the distributions. On the other hand, relying only on administrative records may generate significant biases in the estimation of income for populations with income typically not captured by those data. Combining these two sources allows us to mitigate both these problems simultaneously.

With our assumptions on survey and administrative earnings from above, Bee et al. (2023) define a model in a Mean Squared Error (MSE) framework with a set of parameters on the random noise and relative mean reversion in survey report, y_s, and administrative record, y_a, conditional on other observed characteritsics, x. The model also defines a "survey confidence" (SC) measure that is a function of two sets of terms. The first is a measure of the estimated bias in the administrative data by comparing $E(y_s|x)$ to $E(y_a|x)$. The second set of terms compares the relative variance of the random noise in the two reports conditional on x. We

[33]To further motivate the relevance of these assumptions, consider estimating earnings for auto mechanics as a group. Assumption (a) would imply that if you asked auto mechanics to report what they earned on a survey, some would over-report and some would under-report, but you would still recover an unbiased estimate of average earnings. On the other hand, at the individual-level these mechanics might not remember their exact earnings and so might report their earnings from an average of prior years, such that variation across survey reports would not reflect true variation in earnings for that year. On the other hand, assumption (b) implies that administrative records would fail to generate a correct average for auto mechanic earnings, presumably due to the prevalence of under-the-table payments. Under assumption (b), administrative data better capture variation across individual-level earnings, such that a mechanic whose W2 earnings were twice as large as another mechanic would be expected to have actually earned twice as much in that year. This would be satisfied if, for example, all auto mechanics reported 50 percent (or any fixed percent) of their income to the IRS.

select the survey report if the squared bias term exceeds the difference in the variance terms, or if in the MSE framework, the estimated administrative bias is exceeded by its relatively lower noise.

The model is only identified and possible to estimate with an assumption about the degree of mean reversion in survey reports relative to administrative reports. This mean reversion parameter, κ (or "kappa" in tables and figures in this paper), cannot be estimated, and must be assumed because true earnings, y, are never observed. If $\kappa = 1$, there is no mean reversion in the survey relative to the administrative data. We assume greater mean reversion as κ decreases from 1. With a given κ, we can estimate the SC measure for each individual conditional on his or her x characteristics, which would reflect the model's "confidence" by comparing the bias and variance terms in an MSE framework. We use this SC measure in our decision rule to select the survey or administrative wage and salary earnings report — if SC > 0, we select the survey report.[34]

We select the "best" wage and salary earnings report for individuals based on their observable characteristics x, but *not* conditional on their actual survey or administrative reports. This is in contrast with Meyer et al. (2021*b*), which takes the maximum of survey-reported and administrative earnings in at least some cases. In other words, we take survey reports for people whose characteristics suggest that their survey reports are better according to the SC measure than their administrative reports. Bee et al. (2023) discuss potential limitations and extensions of this approach to incorporate the actual earnings reports and additional information, such as longitudinal earnings histories, to improve our estimates of earnings given survey and administrative reports.

Misclassification of wages versus self-employment earnings further complicates efforts to reconcile multiple earnings reports. If individuals report wage and salary earnings on the

[34]Bee et al. (2023) discuss the implementation details of the estimation and additional features of our decision rule in the case when we determine that $E(y_s|x) < E(y_a|x)$ with some confidence for a given individual.

survey but self-employment earnings on their tax returns, it's not clear whether those represent two separate sources of income or the same income reported in different categories. Misclassification appears to be a common issue. Only 35 percent of individuals with positive administrative self-employment earnings report any self-employment earnings on the survey and less than 50 percent of the survey self-employed have positive self-employment earnings in the administrative data (Abraham et al., 2021). At this time, we generally defer to the administrative data when there is disagreement about the source of earnings (wage and salary vs. self-employment) or if self-employment is reported in both survey and administrative data. In the future, addressing misclassification of earnings and self-employment earnings misreporting is an important avenue of research and improvement of our income estimates.

In Table A8, we summarize the possible combinations of survey and administrative reports of wage and salary and self-employment earnings and show which we use in our income estimates. The measurement error model discussed in this section is used for 53 percent of adults[35] and for 74 percent of individuals with any reported earnings in either source. Another 39 percent of adults had no survey or administrative earnings or reported earnings in one source, but not the other. Given that we default to the source with reported earnings under extensive margin disagreement, that leaves above 8 percent of adults or 12 percent of individuals with earnings in either source for whom we ignore survey reported wage and salary earnings and use only administrative data due to potential misclassification or other data issues.

In Table A9, we show the share of individuals whose survey earnings would be used for various κ mean-reversion parameter values (from the set of people listed as using the measurement error model in Table A8). The share varies from 6 percent ($\kappa = 0.7$) to 31 percent ($\kappa = 1$, no survey-report mean reversion). For the NEWS estimates, we select $\kappa = 0.9$ as it implies

[35]In this context, we define adult as people aged 15 and above who are asked the CPS ASEC earnings questions.

a relatively modest level of mean reversion and selects the survey wage and salary earnings report 21 percent of the time. However, we assess robustness to alternative values of κ in Section 5.2.

Given our chosen survey mean reversion parameter, Table 8 reports the share of individuals whose survey earnings were used as part of our measurement error model (as a share of workers from Table A8 for whom the measurement error model was used). Overall, we use survey earnings for 21 percent of workers. The rate at which survey earnings are used varies by age, race, occupation, and industry. For example, survey earnings are used less often for Black workers and younger (18-24) and older (55+) workers. However, survey earnings are used for 59 percent of workers in the construction industry.

4.3.2 Income Replacement

In this section, we discuss the final step – combining the survey and administrative data and replacing particular survey income components with their counterparts in the administrative data in order to estimate each survey respondent's money income. We use separate processes for filers and nonfilers. There is more income information available for tax filers, but some of it is only available at the tax unit, but not the individual, level. Table A10 summarizes the income information available for filers and nonfilers.

For tax filers, we start with Total Money Income (TMI) constructed from their 1040s, which is the sum of taxable wage and salary income, interest (taxable and tax-exempt), dividends, alimony received, business income or losses (including from partnerships and S-corps), farm income or losses, net rent, royalty, and estate and trust income, unemployment compensation and gross Social Security benefits (as noted in Section 3.3.1).

For wage and salary earnings, TMI includes taxable wage and salary earnings reported on the 1040. This amount will understate true earnings if gross earnings are greater than taxable earnings, for example, if individuals have deferred compensation or use pre-tax earnings to pay health insurance premiums. It will also understate earnings if filers underreport their true earnings to the IRS. Therefore, we replace the wage and salary earnings component of TMI with our survey or job-level administrative earnings according to the rules shown in Table A8 and discussed in Section 4.3.1. We also replace 1040-reported Social Security income, as we are more confident in the data quality of the SSA data than in the gross 1040 amounts, which may not be well-reported in tax returns (particularly for non-taxable Social Security income).

For retirement income, we cannot distinguish defined contribution (DC) plan withdrawals from defined benefit (DB) pensions in the 1099-R data.[36] In the CPS ASEC, DC withdrawals are only counted as income for people aged 59 and above. We therefore follow that convention and include 1099-R retirement income for all individuals aged 59 and older. For those under 59, we include the 1099-R income if they reported pension or annuity income on the survey. We add this retirement income to TMI.

Finally, we add several income components that are not taxable. From administrative sources, we add SSI and TANF and from the survey, we add educational assistance, financial assistance, workers' compensation, and veterans benefit payments. For filers, that gives us our adjusted TMI, which we use in the income and poverty estimates.

For nonfilers, we must add up the available components individually, since we do not have a 1040 TMI amount. To get the nonfiler equivalent of adjusted TMI, we start with wage and salary and self-employment earnings as indicated in Table A8. From administrative data sources, we add Social Security income (PHUS), retirement income (from the 1099-R following the same rules for filers as noted above by age), SSI (SSR), and TANF (state data). We add UI compensation, interest, and dividends imputed using the parameters estimated on the complete 1099-G, 1099-DIV, and 1099-INT data (Rothbaum, 2023). From the survey, we add rent and royalty income, educational assistance, financial assistance,

[36] We will apply and extend the work in Bee and Mitchell (2017) to characterize individual withdrawals as defined benefit or defined contribution in future work.

workers' compensation, and veterans benefit payments. The sum of these amounts represents our best estimate of adjusted TMI for nonfilers, which we use in the income and poverty estimates in the next section.

5 Results

5.1 NEWS Estimates

Table 1 and Figure 1 compare the NEWS estimates for median household income in 2018 to the survey estimates released in Semega et al. (2019).[37] Across all households, the NEWS estimate for median household income was 6.3 percent higher ($67,170 vs. $63,180). Median household incomes were also higher for nearly all subgroups shown. The main exceptions were by age of householder. Pooled together, median household income for households under age 65 was not statistically different (-0.1 percent lower point estimate) whereas households 65 and older had 27.3 percent greater median household income ($55,610 vs. $43,700). For households aged 55-64, the difference was 5.0 percent ($72,430 vs. $68,950). For all age groups below 55, the point estimates were not statistically different from zero or negative.

Figure 7 shows estimates from the 10^{th} to 95^{th} percentiles of the household income distribution overall and by race and Hispanic origin, age of householder, and educational attainment. Overall, income increased more in proportional terms at the bottom of the distribution than at the top. This is particularly true for age 65 and over households, for which NEWS household income was 31 percent higher at the 25^{th} percentile, 20 percent higher at the 75^{th} percentile, and 15 percent higher at the 90^{th} percentile.

Comparisons between NEWS and survey estimates for poverty are shown in Table 2 and Figure 2. Overall, poverty was 1.1 percentage points lower than in the survey estimate, equivalent to 9.4 percent fewer people in poverty. As with income, poverty was much lower

[37] All estimates are in 2018 dollars. To adjust to 2021 dollars using the R-CPI-U-RS as in official Census Bureau publications, multiply each income estimate by 399.0/369.8 = 1.079.

for the 65 and older population. We estimate a 3.3 percentage-point lower poverty rate and 34.1 percent fewer people in poverty. There were no groups for which poverty was statistically higher with the NEWS estimates. However, we did not find a statistically significant difference in poverty for Black individuals, children, residents of the Midwest, those outside of Metropolitan Statistical Areas, those with a disability, and those with some college education.

Finally, in Table 3, we compare NEWS estimates for inequality statistics to the survey estimates, including for income shares, the Gini index, and various percentile ratios.[38] For shares of income, we find a decrease in the share of income in the 2^{nd} to 4^{th} quintile and an increase in the share of income in the top quintile and particularly the top 5 percent. We estimate an increase in the Gini coefficient from 0.459 to 0.476. This is likely coming from no top coding and higher extreme income values in the administrative data relative to the survey, despite the larger increase in income at lower percentiles of the income distribution shown in Figure 7, Panel A.[39] However, consistent with that figure, we find declines in the percentile ratio estimates (90/10, 90/50, and 50/10). For example, in the survey responses, household income at the 90^{th} percentile is 12.5 times as large as at the 10^{th} percentile. With the NEWS estimates, the ratio is 11.5.

5.2 Robustness to Alternative Uses of Earnings Data

Figure A5 compares NEWS estimates of household income to estimates using alternative combinations of survey and administrative wage and salary earnings. In Panel A, we show how income varies under different rules for using earnings when the survey and administrative

[38] One important area of future research is how to address potential data issues that affect inequality, including how well our sample captures income at the far right tail of the distribution and how to address administrative data issues (like implausible extreme values) that might bias inequality statistics. We note this when discussing our future plans in Section 6. This will affect statistics such as income shares and the Gini coefficient that condition on the entire income distribution, but have less of an impact on statistics such as percentile ratios.

[39] Survey income top codes vary by income item, but generally do not exceed $1.1 million dollars for a given income source.

data disagree at the extensive margin, whether any earnings are present. We compare four scenarios to the NEWS estimates (with y_a for administrative earnings and y_s for survey earnings: (1) use y_a unless $y_a = 0$ and $y_s \neq 0$, (2) use y_a, even if $y_a = 0$ and $y_s \neq 0$, (3) use y_s unless $y_s = 0$ and $y_a \neq 0$, and (4) use y_s, even if $y_s = 0$ and $y_a \neq 0$. Scenarios (1) and (2) give priority to administrative earnings and (3) and (4) give priority to survey earnings. If we use either source of earnings when the other is zero, income declines substantially ((2) and (4)), particularly at lower income levels. If we use administrative earnings if $\neq 0$, scenario (1), the household income point estimates are generally lower than the NEWS estimates, although most of the differences are not statistically significant. If we use survey earnings if $\neq 0$, scenario (3), the household income point estimates are lower everywhere, but the differences are only statistically significant in the tails of the distribution.

To summarize, how we handle extensive margin disagreement substantially affects our income estimates, as does whether we prioritize survey or administrative earnings. Compared to just using administrative earnings (if $\neq 0$), the measurement error earnings model does not have a substantial impact on household income overall, despite using survey earnings for 21 percent of the individuals the model was used on. In Figure A5 Panel B, we estimate the household income distribution for alternative κ/survey mean-reversion parameters in the earnings measurement error model. As κ varies from 1 to 0.7, the share of individuals whose survey earnings are used changes from 6 to 31 percent. Despite this, and while there are statistically significant differences between the NEWS estimates ($\kappa = 0.9$) and estimates with other κ, there are few economically meaningful differences in the household income estimates. For example, none of the alternative κs estimates a statistically significant difference in median household income and the range on the point estimates is from -0.05 percent to 0.03 percent different from NEWS estimate. At the 95^{th} percentile, the estimates range from -0.46 percent to 0.89 percent different from the NEWS estimate (with only 0.89 percent different for $\kappa = 0.7$ statistically different from the NEWS estimate).

However, the choice of how to combine survey and administrative earnings *could* matter considerably more, shown in Panel C of Figure A5. We add another possible decision rule, which is to take the maximum of the two reports. This approach might be reasonable if one thinks all misreporting in both survey and administrative data is underreporting, although that does not seem consistent with the noise in survey reports around administrative wage and salary earnings we observe in Figure 6.[40] Taking the maximum of reported wage and salary earnings would vastly increase measured household income across the distribution. Across the percentiles plotted in Figure A5, the income estimate using the maximum rule would be 13.5 percent greater than the NEWS estimate, on average.

5.3 Impact of Different Processing Steps on Income and Poverty Estimates

The NEWS estimates reflect several bias correction steps, including reweighting for nonresponse, reweighting for linkage to administrative data, imputing to address nonrandom nonresponse, replacement of survey responses with administrative income information (including observed and imputed TANF and gross earnings), and the earnings measurement error model to select survey or administrative earnings. In Figure 3, we decompose the adjustments to show the impact of each of these steps on the distribution of household income. In Panel A, we show the weighting and survey imputation steps compared to the survey estimates, as these steps use administrative data to adjust for bias in survey-only information (the weights and imputed earnings). In Panel B, we show the impact of using administrative data (as discussed in Section 4.3.2) and the earnings measurement error model compared to the adjusted survey estimates from Panel A. In other words, Panel A illustrates the effect the survey-only adjustments and Panel B shows the effect of the final two steps after accounting

[40]Meyer et al. (2021*b*) take the maximum of survey and administrative earnings (total earnings, not just wage and salary) at least in some cases. However, they argue their estimates of extreme poverty are not affected by this because in most cases both the survey and the administrative earnings measure exceeds their extreme poverty thresholds when they disagree on the intensive margin.

for the survey-only adjustments.

The weighting steps lower income across most of the distribution by 1 to 2 percent.[41] Replacing the survey earnings imputations (and accounting for uncertainty through multiple imputation) lowers the point estimates at the bottom of the distribution, consistent with the selection into response observed by Bollinger et al. (2019) in the tails and results in confidence intervals that are wider on average.

In Figure 3 Panel B, we show the impact of the final two steps, income replacement and the earnings measurement error model, compared to the estimate after survey earnings imputation from Panel A. We compare the household income distribution with and without the administrative data and find large effects across the distribution, from 17.1 percent at the 10^{th} percentile, to 10.3 percent at the 25^{th}, 6.8 percent at the median, and 3.6 percent at the 75^{th}. Panel B also shows the impact of the earnings measurement error model and the use of survey earnings, which has a minimal impact on household income.[42] Panel C shows the overall comparison between the NEWS and survey estimates.[43]

Figure A6 shows the same decomposition by survey adjustments (Panel A) and administrative income replacement and measurement error model (Panel B) for the subgroups in Table 1. Figure A7 does the same for poverty. In both, it is generally the case that the survey adjustments move point estimates for median household income down and poverty up, but generally the differences are not statistically significant. The administrative income replacements move income up and poverty down for most subgroups as well.

[41] This is slightly different than Rothbaum and Bee (2022), which found no statistically significant differences across the distribution with an average point estimate of -0.23. However, we use more data, particularly contemporaneous rather than lagged 1040 income in the NEWS project, which may reflect selection into response that was not captured in that paper using data available during the regular CPS ASEC production schedule.

[42] We discuss how alternative uses of survey earnings could have had a large impact in the next section.

[43] The same information by age of householder (under 65 and 65 and over) is available in the Appendix in Figure A2.

5.4 Impact of Different Income Types on Income and Poverty Estimates

Finally, we assess how specific administrative income components affect the household income distribution and poverty. To do so, we start with the NEWS income estimates and replace each administrative income item one by one (not sequentially or cumulatively) with its survey counterpart and compare each statistic after the replacement to the NEWS estimate. The results are shown for income in Figure 8 and poverty in Figure 9.

For income, we make several replacements: (1) interest and dividends, (2) retirement income, including DC withdrawals and retirement, survivor, and disability pensions, (3) Social Security and SSI, and (4) wage and salary earnings.

For interest and dividends, we make three replacements: 1) replace administrative interest income with survey interest income, including the survey measure of interest (and other returns) on retirement accounts, 2) replace administrative income with survey interest income, excluding the retirement account interest, and 3) replace administrative dividends with survey dividends, with detail shown in Figure A3 Panel A. If we include interest on retirement accounts (as is the case in the survey income estimate), we get more income across the distribution than using administrative income (which does not include this interest). Because we already count withdrawals from these same retirement accounts as income, this risks double counting the same income, which is why we exclude it from the NEWS estimate. If we replace interest or dividends excluding this interest from retirement accounts, we see slightly lower income across the distribution.Together, interest and dividend replacement with survey responses lowers income by 1.3 percent at the 25th percentile and 0.5 percent at the 75 percentile, shown in Figure 8.

Next, we look at transfer income, including Social Security (OASDI), SSI, and TANF income, shown in detail in Figure A3 Panel B. If we just replace SSI income with survey responses, we observe increases in income at the bottom of the distribution, primarily because of misclassification of Social Security and SSI, effectively double counting Social Security for individuals that reported Social Security income as SSI. If we replace Social Security only, we observe big declines in income at the bottom and smaller declines higher in the income distribution. If we replace both together, we observe slightly smaller declines at the bottom because we are preserving the misclassified income (SSI reported as Social Security on the survey, for example). Replacing TANF with survey responses results in small declines in income that are only significantly different at a handful of points.Replacing both Social Security and SSI together lowers income by 1.0 perent at the 25th percentile, but the difference is not statistically significant at the 75th percentile, shown in Figure 8.

Figure 8 also shows the the impact of replacing retirement, survivor, disability, and pension income (retirement income, from Form 1099-R) with the corresponding survey items. Even for overall income, the retirement income replacement has the biggest impact across much of the income distribution, including 8.7 percent at the 25th percentile and 4.1 percent at the 75th percentile.

As shown in Figure 9, overall poverty is higher when using survey reports for interest and dividends. It is much higher if we use survey-reported retirement income. Likewise, replacing administrative with survey earnings has a large effect on poverty, particularly if we ignore positive administrative earnings when the survey reports are zero.

6 Release and Future Research

6.1 Transparency and Data Availability

An integral goal of the NEWS project is to be as transparent and open about the data we use, how we clean them, and how we combine them to generate the NEWS income, poverty, and resource estimates. Clarity and transparency are especially important in this context, as there are many decisions about how to clean, process, and combine survey and administrative data that can have major effects on the results. These choices can be relatively opaque and "in-the-weeds" for even a well-informed outsider. For example, using the maximum of survey and administrative income reports, as shown in Figure A5 Panel C, would drastically bias our income and poverty estimates in a way that is not consistent with the survey reporting noise in Figure 6. Transparency about our methods, code, and estimates is required for readers to understand the implications of those kind of detailed data choices.

As such, we commit to making all of the code and as much of the data as we are permitted available to researchers through the Federal Research Data Center (FSRDC) system.[44] We also commit to making the code publicly available, with as few edits as possible as required by the rules on the disclosure of code to abide by Titles 13 and 26 and our agreements with data providers.

With each run of the NEWS code, we also plan to log any changes to input extracts so we can track any changes to input data (such as data provided by the IRS or an updated version of a survey file) that may affect our estimates. We also use git, a software version control system, to ensure that the code that generated the results in this paper (or any future paper with updated data, code, and methods) can be replicated.[45]

We also have written documentation for nearly all the files and functions involved in loading and cleaning the data, creating the address and person extracts, implementing the reweighting, imputation, and earnings measurement error model, generating the final person and tax unit income variables, and estimating income and poverty. While no documentation is perfect, we have endeavored to be as detailed as possible in this documentation, detailing what each section of code is doing, including references to particular line numbers. This is

[44]Subject to the constraints of our data agreements with the various state and federal agencies and commercial data providers.

[45]Up to the limit of what is possible in the software we use. Unfortunately, there are functions we currently use, such as Stata's `rmcoll` function to remove collinear variables from a regression that do not necessarily remove the same variables even when run with the same random seed. The exact set of variables kept can then affect the results from subsequent steps, such as LASSO regression feature selection. A goal for future releases is to remove our dependence on any function that has this property as we would like to ensure that a rerun of the code with the same data and initial seeds generates exactly the same estimates.

in addition to the regular commenting provided within the code itself.

6.2 Future Plans

This release represents version 1.0 of the NEWS project. There are many aspects of this work that we were not able to include in this release and have left for future work. In this section we discuss our goals for version 2.0 and beyond.

First, we have estimated income and poverty in a single year, 2018, as a proof of concept and first step in this work. We plan to expand this to include more years, both earlier years and years up to the present. This will introduce additional challenges. Some administrative data are not available before a specific year. For example, the Census Bureau currently only has access to the universe of W-2 earnings starting in 2005. Likewise, not all administrative data are available in time for estimates of income in the prior year. For example, we might get data from SSA or state agencies with a lag of a year or more. Creating historical or preliminary estimates in the absence of complete data is an important direction for future research.

Second, we have only estimated income and poverty statistics at the national level. In the future we plan to extend the estimates to smaller geographic units, including states, counties, and possibly census tracts. However, to do so would require changes to how the estimates are generated. First, we would likely move to the ACS as the main source of survey information for subnational estimates. However, the ACS has less detailed income information, which makes this work more challenging and would require our using a different approach to estimating various income sources. For example, we do not have separate survey reports of interest, dividends, rental income, unemployment compensation, workers' compensation, etc., because these items are reported as part of questions that ask about several income items simultaneously. Therefore, it will be difficult to know whether the respondent was also reporting another type of income that is not well-covered by available administrative data. In the long term, we may even move beyond the survey sample (while using survey information in the process) to better estimate statistics for small areas using the available administrative, decennial census, and commercial data.

Third, we have generated estimates only for pre-tax money income, as measured in the Census Bureau's annual income and poverty release (Semega et al., 2019). However, there is considerable interest in how in-kind benefits, taxes, and credits affect measures of material wellbeing. We plan on expanding the notions of resources we measure and as well as the set of wellbeing and deprivation statistics we report. For example, we could measure the distribution of disposable income, disposable income plus the cash value of some (or all) in-kind transfers, improved measures of compensation that include employer matches to retirement contributions and employer contributions to health insurance premiums, the Supplemental Poverty Measure (SPM), etc. This will entail estimating taxes and credits and/or addressing household roster disagreement between administrative and survey data (Unrath, 2022; Meyer et al., 2022), incorporating additional data on housing assistance from the Department of Housing and Urban Development and from states on the Special Supplemental Nutrition Assistance Program for Women, Infants, and Children (WIC), and potentially improved imputation and misreporting corrections for other programs such as the National School Lunch program, etc.

Finally, there are dimensions of misreporting and measurement error that we were not able to address in this version. For example, we have discussed how self-employment earnings are underreported in both survey and administrative data (Hurst, Li and Pugsley, 2014; Internal Revenue Service, Research, Analysis & Statistics., 2016) and how much survey and administrative reports disagree on the extensive margin (Abraham et al., 2021). It is not settled in the literature how to adjust for this underreporting (Auten and Splinter, 2018; Piketty, Saez and Zucman, 2017), much less how one would do so and get unbiased estimates by subgroup. We plan to extend our measurement error model to self-employment earnings for which different assumptions about misreporting would be necessary. Likewise, it may be the case that survey samples, even those as large as the ACS, do not adequately capture the incomes of the top individuals and households. Imputation, combination, or reweighting may be insufficient to address this issue to estimate unbiased inequality statistics from a survey sample. We plan on also researching methods to better estimate inequality statistics that account for the far-right tail of the income distribution.

We would also like to further investigate how our adjustments affect estimates for subgroups that may be challenging to reach or be unlikely to be present in the administrative data, such as non-citizens. Weighting and imputation, in particular, assume that the data is missing at random conditional on the observable information. However, there may be limited observable information in the address-linked administrative records to identify and adjust for selection into response by citizenship status. Likewise, our weighting adjustment for linkage uses survey response information to reweight individuals and households that can be linked to administrative data to be representative of the full sample. However, it may be that conditional on the observable survey information (and the address-linked administrative data), the data are not missing at random and that our final estimates for this group are biased. Similarly, there are difficult to reach subgroups that are not in sample for the CPS ASEC that we would like to estimate wellbeing statistics for, such as individuals in group quarters and the homeless or unhoused.

7 Conclusion

This release under the NEWS project is a first step toward integrating what we know about bias and measurement error in survey and administrative data into a set of “best possible” estimates of income, poverty, and resource statistics. We have attempted to address as many of the sources of bias as possible, including nonresponse bias (unit and item), selection into linkage to administrative data, misreporting of survey and administrative income, and incomplete data. However, much work remains to be done to address additional potential sources of error. As we and other researchers advance our understanding of how to address these measurement challenges, we will revise these estimates.

This work also suggests several additional avenues of possible research at the Census Bureau. For example, estimating income and poverty from linked survey and administrative data could impact the information we depend on surveys to provide. Surveys could focus less on items that are well captured in administrative data (such as Social Security payments) and more on items that improve linkage and those that are less well captured by administrative data (self-employment income, etc.). The Census Bureau could also increase efforts to collect survey responses from hard-to-reach groups who may be less well covered by administrative data.

The focus of this project is on improving our estimates of income and poverty. However, much of our planned future work entails trying to understand the quality of various data sources. This commitment promises many potential benefits to users of both survey and administrative data who are not primarily focused on income and poverty measurement. We hope to extend our work, particularly on earnings, to help characterize the data quality issues that other researchers may confront.

References

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

Abowd, John M, Kevin L McKinney, and Nellie L Zhao. 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.

Abraham, Katharine G, John C Haltiwanger, Claire Hou, Kristin Sandusky, and James R Spletzer. 2021. "Reconciling survey and administrative measures of self-employment." *Journal of Labor Economics*, 39(4): 825–860.

Alvey, Wendy, and Cynthia Cobleigh. 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.*

Ambler, Gareth, Rumana Z. Omar, and Patrick Royston. 2007. "A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome." *Statistical methods in medical research*, 16(3): 277–298.

Auten, Gerald, and David Splinter. 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.*

Bee, Adam. 2013. "An Evaluation of Retirement Income in the CPS ASEC Using Form 1099-R Microdata." *Unpublished U.S. Census Bureau Working Paper.*

Bee, Adam, and Jonathan Rothbaum. 2019. "The Administrative Income Statistics (AIS) Project: Research on the Use of Administrative Records to Improve Income and Resource Estimates." *U.S. Census Bureau SEHSD Working Paper #2019-36.*

Bee, Adam, and Joshua Mitchell. 2017. "Do Older Americans Have More Income Than We Think?" *U.S. Census Bureau SEHSD Working Paper #2017-39.*

Bee, Adam, Graton Gathright, and Bruce D. Meyer. 2015. "Bias from Unit Non-response in the Measurement of Income in Household Surveys." *Unpublished U.S. Census Bureau Working Paper.*

Bee, Adam, Joshua Mitchell, and Jonathan Rothbaum. 2019. "Not So Fast? How the Use of Administrative Earnings Data Would Change Poverty Estimates." *Unpublished U.S. Census Bureau Working Paper.*

Bee, Adam, Joshua Mitchell, Nicolas Mittag, Jonathan Rothbaum, Carl Sanders, Lawrence Schmidt, and Matthew Unrath. 2023. "Addressing Measurement Error in Income Reports by Combining Survey and Administrative Earnings." *Unpublished U.S. Census Bureau Working Paper.*

Benedetto, Gary, Joanna Motro, and Martha Stinson. 2016. "Introducing Parametric Models and Administrative Records into 2014 SIPP Imputations."

Benedetto, Gary, Jordan C. Stanley, and Evan Totty. 2018. "The Creation and Use of the SIPP Synthetic Beta v7.0." *U.S. Census Bureau Working Paper.*

Benedetto, Gary, Martha Stinson, and John M Abowd. 2013. "The creation and use of the SIPP Synthetic Beta." *U.S. Census Bureau Working Paper.*

Bhaskar, Renuka, James M Noon, Brett O'Hara, and Victoria Velkoff. 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, Renuka, Rachel Shattuck, and James Noon. 2018. "Reporting of Indian Health Service Coverage in the American Community Survey." *U.S. Census Bureau CARRA Working Paper #2018-14.*

Bingley, Paul, and Alessandro Martinello. 2017. "Measurement Error in Income and Schooling and the Bias of Linear Estimators." *Journal of Labor Economics*, 35(4): 1117–1148.

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

Bollinger, Christopher R, and Barry T Hirsch. 2006. "Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching." *Journal of Labor Economics*, 24(3): 483–519.

Bollinger, Christopher R, Barry T Hirsch, Charles M Hokayem, and James P Ziliak. 2019. "Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch." *Journal of Political Economy*, 127(5): 2143–2185.

Bondarenko, Irina, and Trivellore E Raghunathan. 2007. "Multiple Imputations Using Sequential Semi and Nonparametric Regressions." American Statistical Association Alexandria, VA.

Bond, Brittany, J David Brown, Adela Luque, Amy O'Hara, 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, John, and Alan B Krueger. 1991. "The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?" *Journal of Labor Economics*, 9(1): 1–24.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of Econometrics*. Vol. 5, 3705–3843.

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

Brummet, Quentin. 2014. "Comparison of Survey, Federal, and Commercial Address Data Quality." *U.S. Census Bureau CARRA Working Paper #2014-06.*

Brummet, Quentin, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride. 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.

Carr, Michael D, Robert A Moffitt, and Emily E Wiemers. 2022. "Reconciling Trends in Male Earnings Volatility: Evidence from the SIPP Survey and Administrative Data." *Journal of Business & Economic Statistics*, 1–10.

Chapin, William, Sandra Clark, Amanda Klimek, Christopher Mazur, Chase Sawyer, and Ellen Wilson. 2018. "Housing Administrative Records Simulation." *U.S. Census Bureau ACS Research and Evaluation Report #ACS18-RER-07*.

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

Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon. 2010. "Quantile and probability curves without crossing." *Econometrica*, 78(3): 1093–1125.

Chow, Melissa C, Teresa C Fort, Christopher Goetz, Nathan Goldschlag, James Lawrence, Elisabeth Ruth Perlman, Martha Stinson, and T Kirk White. 2021. "Redesigning the Longitudinal Business Database." *NBER Working Paper #28839*.

Corinth, Kevin, Bruce D Meyer, and Derek Wu. 2022. "The Change in Poverty from 1995 to 2016 Among Single Parent Families." *National Bureau of Economic Research Working Paper #29870*.

Deming, W Edwards, and Frederick F Stephan. 1940. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *The Annals of Mathematical Statistics*, 11(4): 427–444.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration estimators in survey sampling." *Journal of the American statistical Association*, 87(418): 376–382.

Duncan, Greg J, and Daniel H Hill. 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, Jonathan. 2021. "Comparing Respondents and Nonrespondents in the ACS: 2013-2018." *Unpublished U.S. Census Bureau Working Paper*.

Eggleston, Jonathan, and Ashley Westra. 2020. "Incorporating Administrative Data in Survey Weights for the Survey of Income and Program Participation." *U.S. Census Bureau SEHSD Working Paper #2020-07*.

Eggleston, Jonathan, and Lori Reeder. 2018. "Does Encouraging Record Use for Financial Assets Improve Data Accuracy? Evidence from Administrative Data." *Public Opinion Quarterly*, 82(4): 686–706.

Estevao, Victor M, and Carl-Erik Säarndal. 2006. "Survey estimates by calibration on complex auxiliary information." *International Statistical Review*, 74(2): 127–147.

Fox, Liana E, Misty L Heggeness, and Kathryn Stevens. 2017. "Precision in measurement: Using SNAP administrative records to evaluate poverty measurement." *U.S. Census Bureau SEHSD Working Paper #2017-49*.

Fox, Liana, Jonathan Rothbaum, Kathryn Shantz, et al. 2022. "Fixing Errors in a SNAP: Addressing SNAP Underreporting to Evaluate Poverty." *AEA Papers and Proceedings*, 112: 330–334.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1): 1–22.

Giefer, Katherine, Abby Williams, Gary Benedetto, and Joanna Motro. 2015. "Program confusion in the 2014 SIPP: Using administrative records to correct false positive SSI reports." *FCSM 2015 Proceedings*.

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

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 25–46.

Harris, Benjamin Cerf. 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*.

He, Yulei, Alan M Zaslavsky, MB Landrum, DP Harrington, and P Catalano. 2010. "Multiple imputation in a large-scale complex survey: a practical guide." *Statistical methods in medical research*, 19(6): 653–670.

He, Yulei, and Trivellore E Raghunathan. 2006. "Tukey's gh distribution for multiple imputation." *The American Statistician*, 60(3): 251–256.

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

Hokayem, Charles, Trivellore Raghunathan, and Jonathan Rothbaum. 2022. "Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data In the CPS ASEC." *Journal of Survey Statistics and Methodology*, 10(1): 81–114.

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

Imboden, Christian, John Voorheis, and Caroline Weber. 2019. "Measuring Systematic Wage Misreporting by Demographic Groups." *Unpublished U.S. Census Bureau Working Paper.*

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, Stephen P, and Fernando Rios Avila. Forthcoming. "Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data." *Journal of the Royal Statistical Society.*

Joint Committee on Taxation. 2022. "Linking Entity Tax Returns and Wage Filings." *JCT Publication #JCX-5-22.*

Jones, Margaret R., and James P. Ziliak. 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.*

Kang, Joseph DY, and Joseph L Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science*, 22(4): 523–539.

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

Kilss, Beth, and Frederick J Scheuren. 1978. "The 1973 CPS-IRS-SSA exact match study." *Social Security Bulletin*, 41: 14.

Larrimore, Jeff, Jacob Mortenson, and David Splinter. 2020. "Presence and persistence of poverty in US tax data." *National Bureau of Economic Research Working Paper #26966.*

Larrimore, Jeff, Jacob Mortenson, and David Splinter. 2021. "Household incomes in tax data using addresses to move from tax-unit to household income distributions." *Journal of Human Resources*, 56(2): 600–631.

Larrimore, Jeff, Jacob Mortenson, and David Splinter. 2022. "Unemployment Insurance in Survey and Administrative Data."

Little, Roderick J, and Sonya Vartivarian. 2005. "Does weighting for nonresponse increase the variance of survey means?" *Survey Methodology*, 31(2): 161.

McKinney, Kevin L, and John M Abowd. 2022. "Male Earnings Volatility in LEHD before, during, and after the Great Recession." *Journal of Business & Economic Statistics*, 1–8.

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

Meijer, Erik, Susann Rohwedder, and Tom Wansbeek. 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.

Meng, Xiao-Li. 1994. "Multiple-imputation inferences with uncongenial sources of input." *Statistical Science*, 538–558.

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

Meyer, Bruce D, and Nikolas Mittag. 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, Bruce D, and Nikolas Mittag. 2021. "An empirical total survey error decomposition using data combination." *Journal of Econometrics*, 224(2): 286–305.

Meyer, Bruce D, Angela Wyse, Alexa Grunwaldt, Carla Medalia, and Derek Wu. 2021*a*. "Learning about homelessness using linked survey and administrative data." *National Bureau of Economic Research Working Paper #28861.*

Meyer, Bruce D, Derek Wu, Grace Finley, Patrick Langetieg, Carla Medalia, Mark Payne, and Alan Plumley. 2022. "The Accuracy of Tax Imputations: Estimating Tax Liabilities and Credits Using Linked Survey and Administrative Data." In *Measuring Distribution and Mobility of Income and Wealth.* , ed. Raj Chetty, John N Friedman, Janet C Gornick, Barry Johnson and Arthur Kennickell, Chapter 15, 459–498. University of Chicago Press.

Meyer, Bruce D, Derek Wu, Victoria Mooers, and Carla Medalia. 2021*b*. "The Use and Misuse of Income Data and the Rarity of Extreme Poverty in the United States." *Journal of Labor Economics*, 39(S1): S5–S58.

Mittag, Nikolas. 2019. "Correcting for Misreporting of Government Benefits." *American Economic Journal: Economic Policy*, 11(2): 142–164.

Moffitt, Robert, and Sisi Zhang. 2022. "Estimating trends in male earnings volatility with the Panel Study of Income Dynamics." *Journal of Business & Economic Statistics*, 1–6.

Moffitt, Robert, John Abowd, Christopher Bollinger, Michael Carr, Charles Hokayem, Kevin McKinney, Emily Wiemers, Sisi Zhang, and James Ziliak. 2022. "Reconciling trends in US male earnings volatility: Results from survey and administrative data." *Journal of Business & Economic Statistics*, 1–11.

Murray-Close, Marta, and Misty L Heggeness. 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, James, Leticia Fernandez, and Sonya Porter. 2016. "Response Error and the Medicaid Undercount in the Current Population Survey." *U.S. Census Bureau CARRA Working Paper #2016-11.*

O'Hara, Amy, Adam Bee, and Joshua Mitchell. 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, Thomas, Emmanuel Saez, and Gabriel Zucman. 2017. "Distributional national accounts: Methods and estimates for the United States." *The Quarterly Journal of Economics*, 133(2): 553–609.

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

Raghunathan, Trivellore E, James M Lepkowski, John Van Hoewyk, Peter Solenberger, et al. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey methodology*, 27(1): 85–96.

Roemer, Marc. 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.*

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1): 41–55.

Rothbaum, Jonathan. 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, Jonathan. 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.*

Rothbaum, Jonathan. 2023. "Research on Creating Synthetic Data to Better Model the Income of Nonfilers through the Release of Public-Use Parameters." *Unpublished U.S. Census Bureau Working Paper.*

Rothbaum, Jonathan, and Adam Bee. 2022. "Addressing Nonresponse Bias in Household Surveys using Linked Administrative Data." *U.S. Census Bureau SEHSD Working Paper #2020-10, Update for 2021 and 2022 unpublished.*

Rothbaum, Jonathan, Jonathan Eggleston, Adam Bee, Mark Klee, and Brian Mendez-Smith. 2021. "Addressing Nonresponse Bias in the American Community Survey During the Pandemic Using Administrative Data." *U.S. Census Bureau SEHSD Working Paper #2021-24.*

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika*, 63(3): 581–592.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *The annals of statistics*, 130–134.

Rubin, Donald B. 1996. "Multiple imputation after 18+ years." *Journal of the American statistical Association*, 91(434): 473–489.

Schmidt, Lawrence D.W., Yinchu Zhu, Brice Green, and Luxi Han. 2022. "quantspace: Quantile Regression via Quantile Spacing." R package version 0.2.1.

Semega, Jessica, Melissa Kollar, John Shrider, Creamer, and Abinash Mohanty. 2019. "Income and Poverty in the United States: 2018." *U.S. Census Bureau Current Population Reports*.

Shantz, Kathryn, and Liana E Fox. 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*.

Slud, Eric V, and Leroy Bailey. 2010. "Evaluation and selection of models for attrition nonresponse adjustment." *Journal of Official Statistics*, 26(1): 127.

Unrath, Matthew. 2022. "Married... With Children? Assessing Alignment between Tax Units and Survey Households." *Unpublished U.S. Census Bureau Working Paper*.

U.S. Census Bureau. 2009. "Estimating ASEC Variances with Replicate Weights Part I: Instructions for Using the ASEC Public Use Replicate Weight File to Create ASEC Variance Estimates." URL: `http://usa.ipums.org/usa/resources/repwt/Use_of_the_Public_Use_Replicate_Weight_File_final_PR.doc`, Accessed: 2022-08-11.

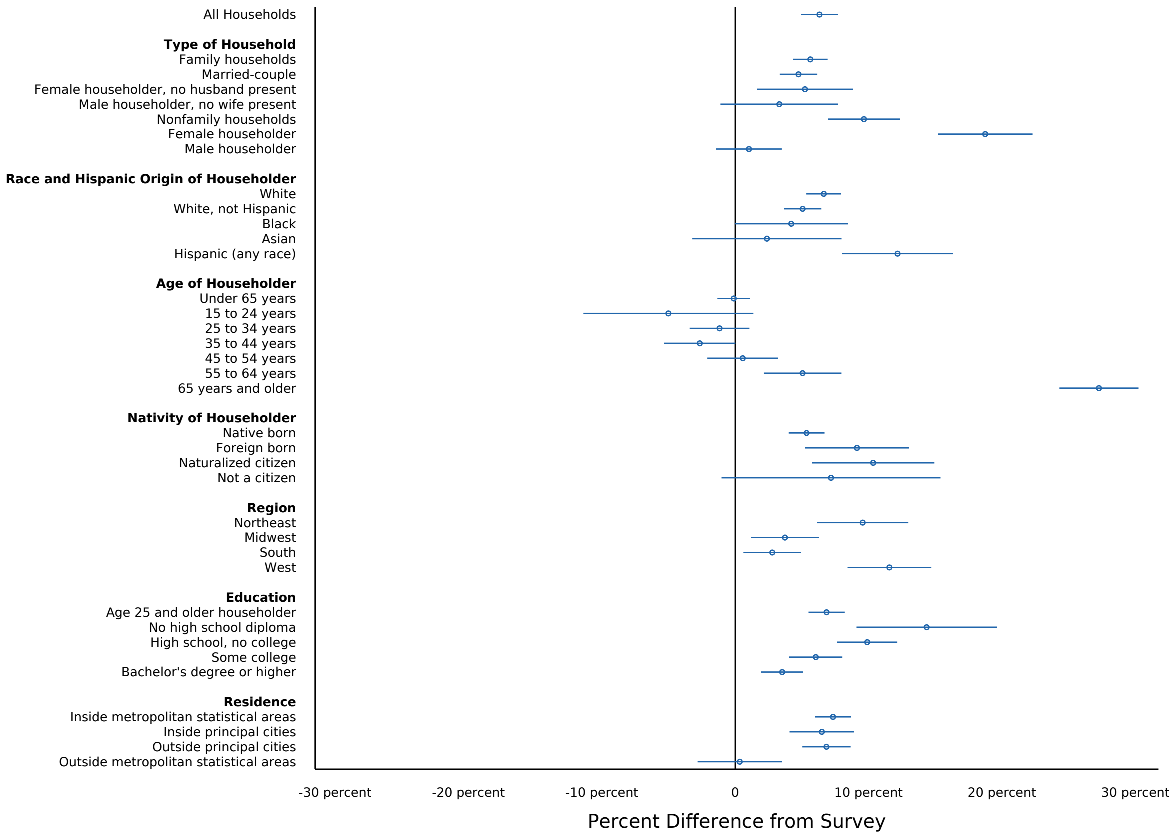
Van Buuren, Stef. 2007. "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical methods in medical research*, 16(3): 219–242.

Wagner, Deborah, and Mary Layne. 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*.

Woodcock, Simon D, and Gary Benedetto. 2009. "Distribution-preserving statistical disclosure limitation." *Computational Statistics & Data Analysis*, 53(12): 4228–4242.

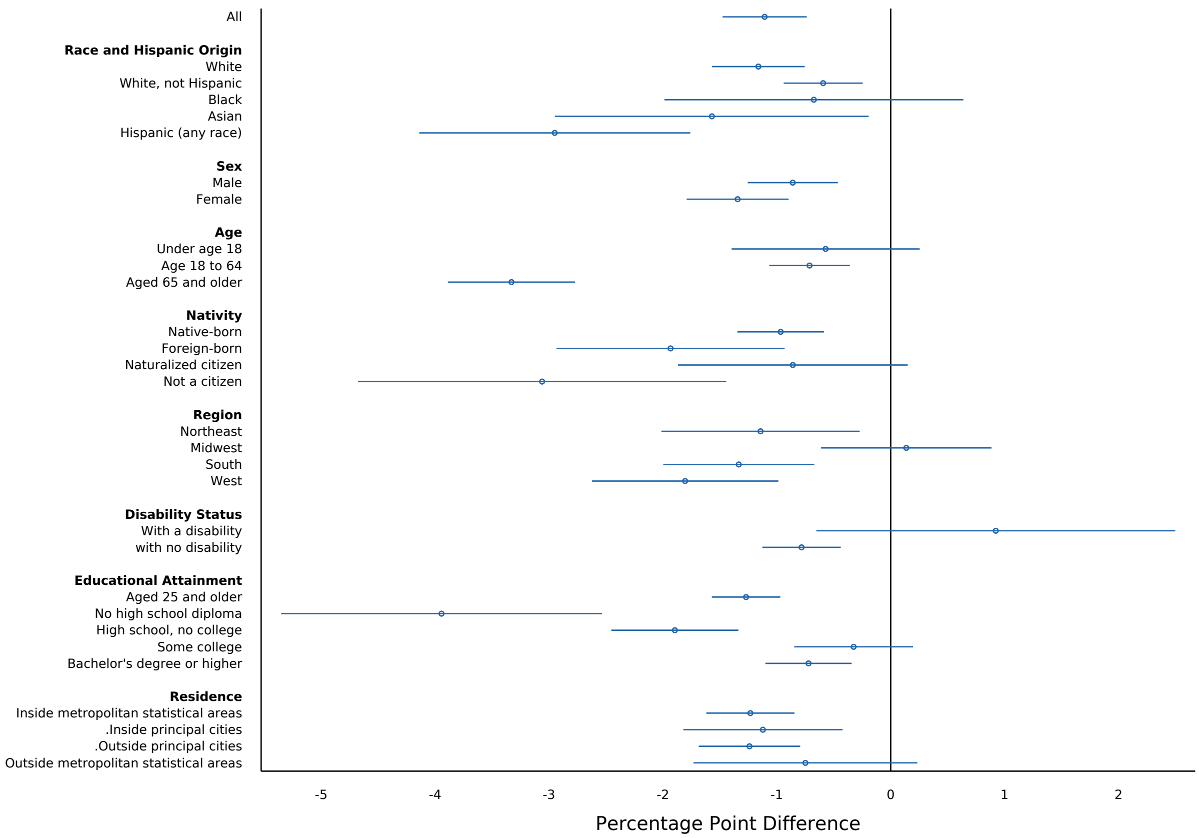
Zhao, Qingyuan, and Daniel Percival. 2017. "Entropy Balancing is Doubly Robust." *Journal of Causal Inference*, 5(1).

Figure 1: NEWS Estimate of Median Household Income Relative to Survey in 2018



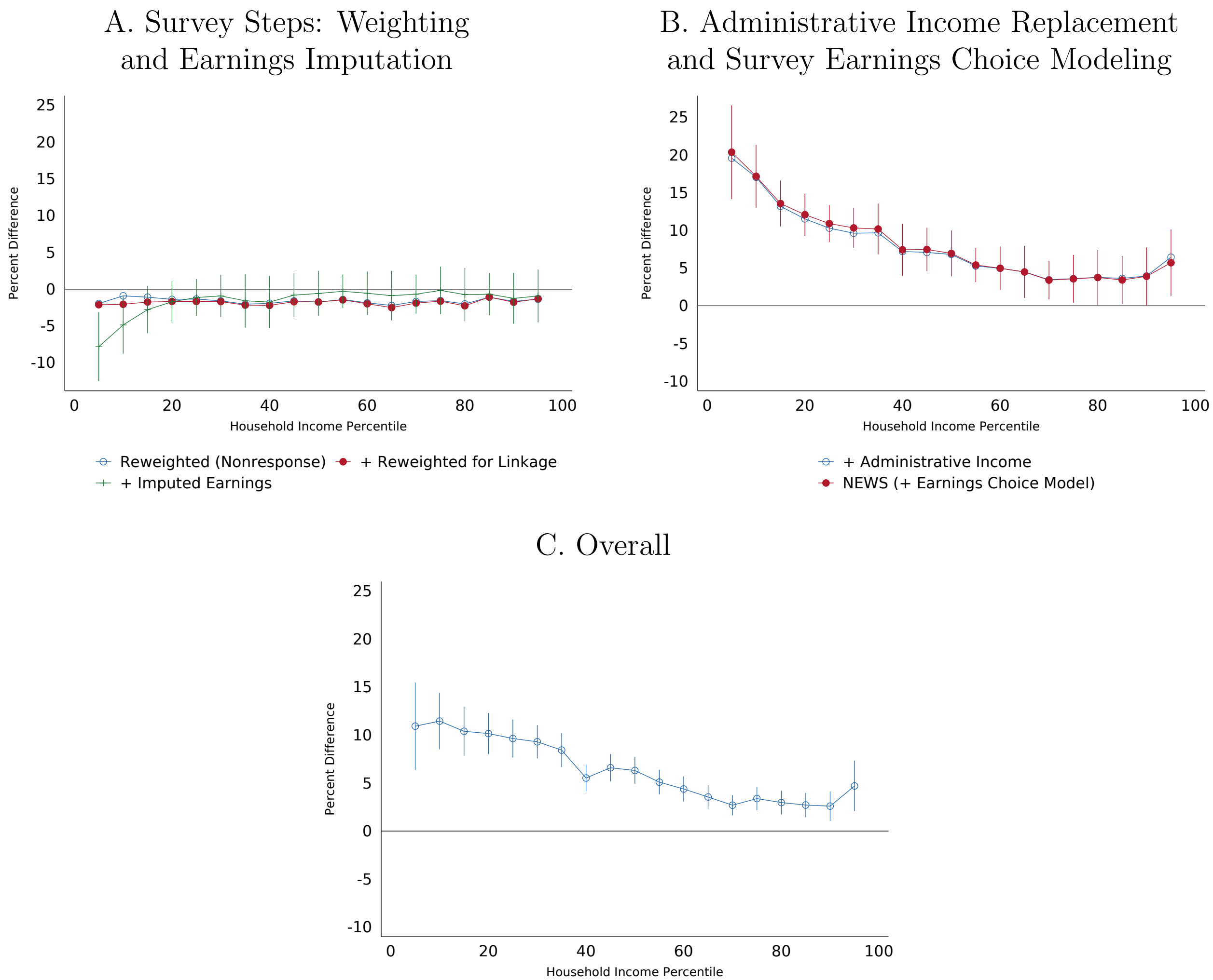
Notes: This figure shows the percent difference between the NEWS estimates of median household income compared to the survey estimates in 2018, also shown in Table 1.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure 2: NEWS Estimate of Poverty Relative to Survey in 2018



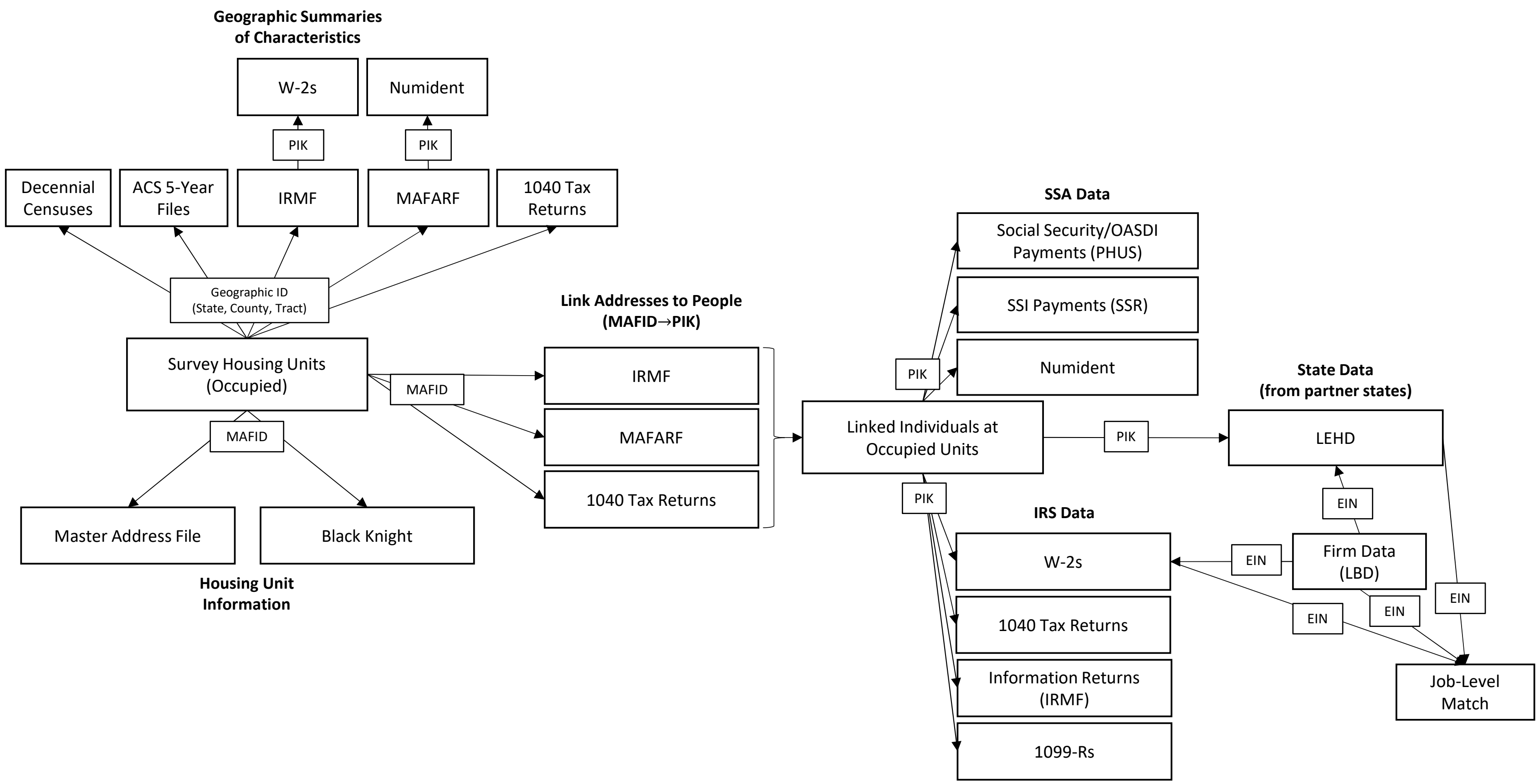
Notes: This figure shows the percentage point difference between the NEWS estimates of poverty compared to the survey estimate in 2018, also shown in Table 2.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure 3: Decomposition of NEWS Processing Steps: Household Income



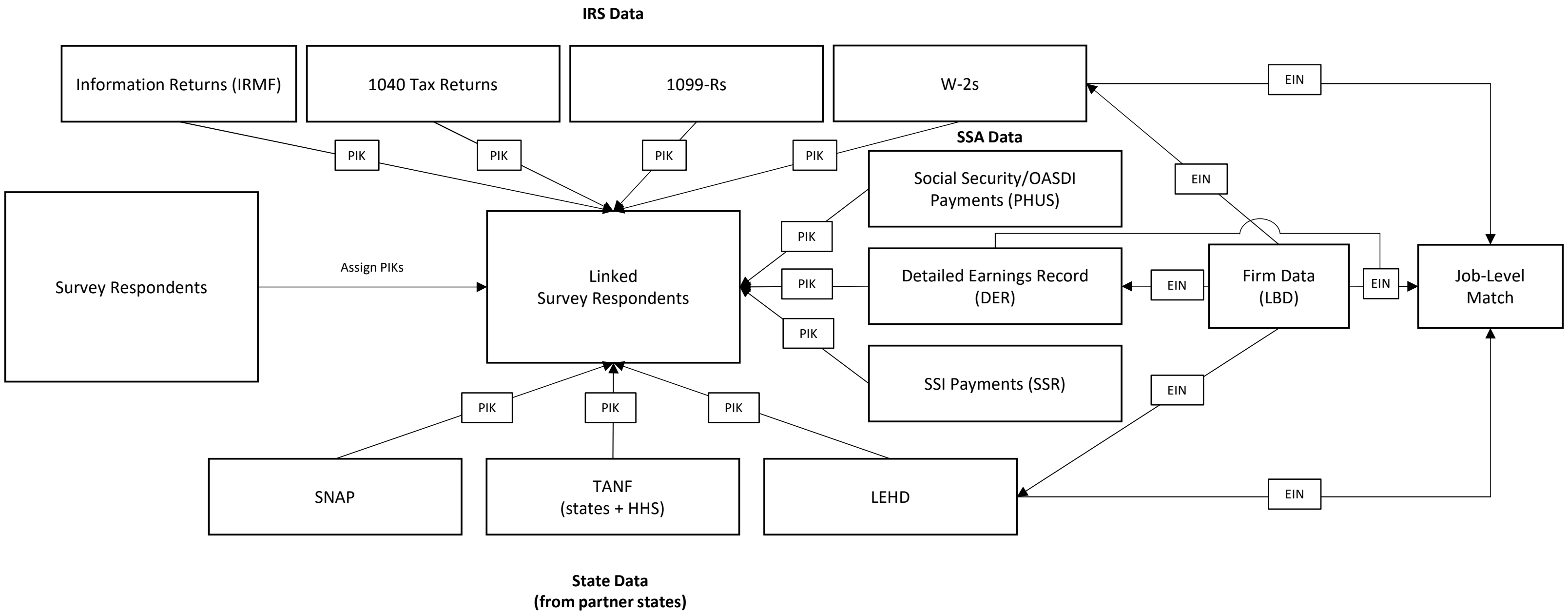
Notes: This figure decomposes the impact of the NEWS processing steps on household income. In Panel A, the figure shows the adjustments made to the survey data, including reweighting and improved earnings imputation comparing household income after the adjustment to the survey estimate. In Panel B, the figure shows impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The full impact of all adjustments is shown in Panel C. The 95 percent confidence interval for the last step is shown in each: for Panel A comparing the estimate after earnings imputation to the survey estimate and for Panel B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure 4: Linkage Diagram for Address File



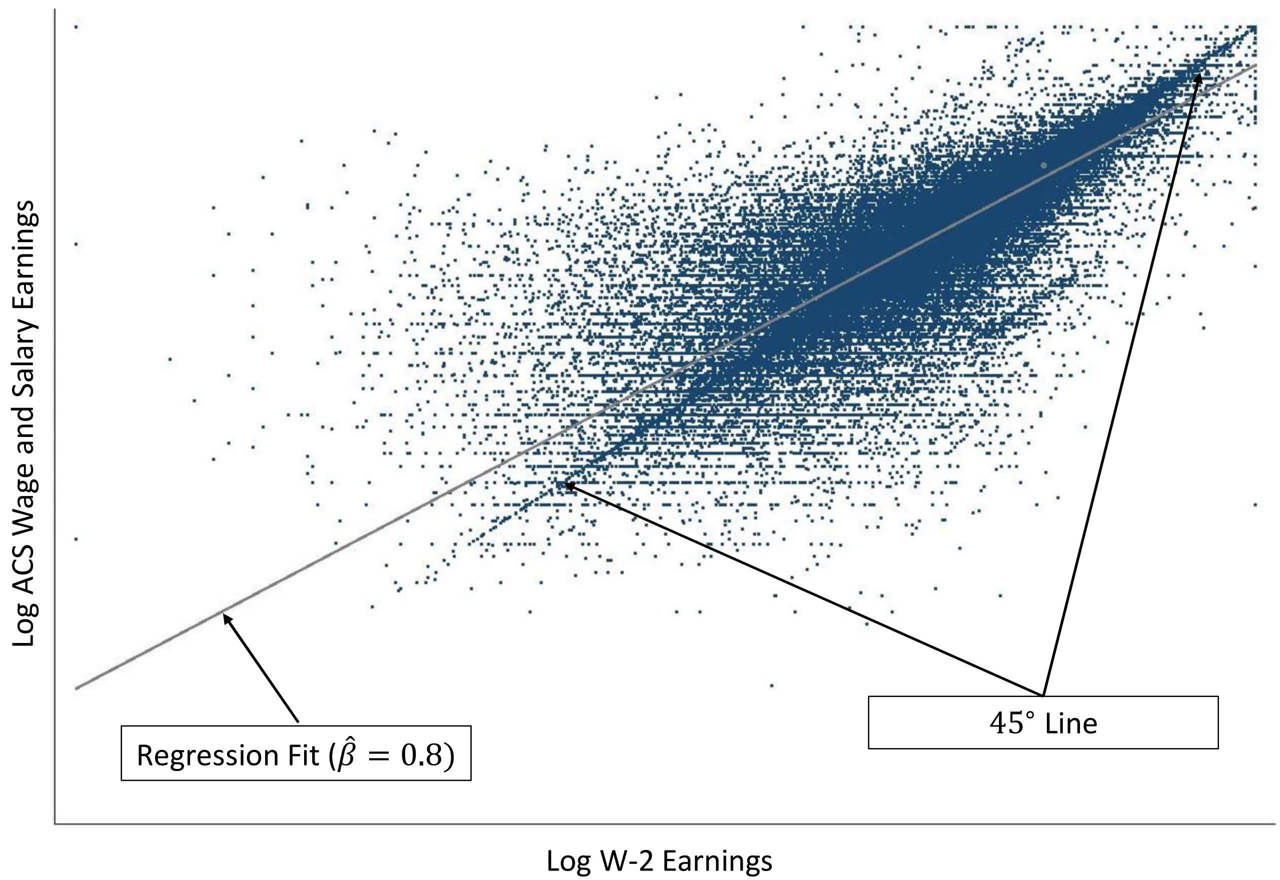
Notes: This diagram shows the linkage used to create the address-based extract file used for weighting. The file starts with the set of occupied addresses in the survey. That file is linked to three sets of files: (1) Geographic summaries of characteristics (by state, county, and tract identifiers), (2) housing unit information from the Master Address File and Black Knight data, and (3) files to link the addresses to people living in them (MAFID $\rightarrow$ PIK). From the third set of files, we create a roster of all individuals found in the occupied surveyed units and link them to the files shown to the right.

Figure 5: Linkage Diagram for Person File



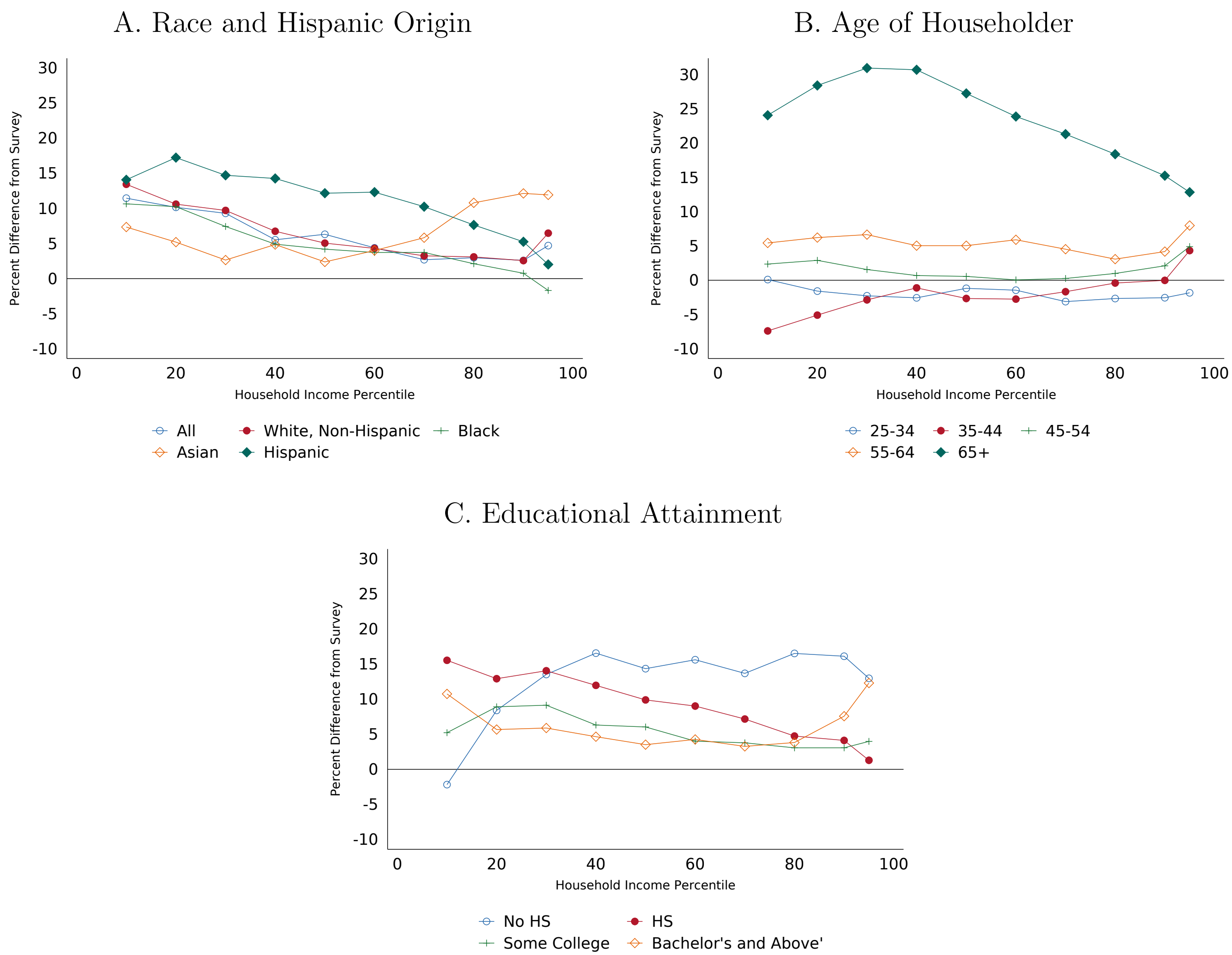
Notes: This diagram shows the linkage used to create the person-level extract file. The file starts with the set of respondents in the survey. For those respondents that can be linked to their Social Security Numbers and therefore assigned a Protected Identification Key (PIK), we link them to the administrative records shown.

Figure 6: Intensive Margin Disagreement in Wage and Salary Earnings



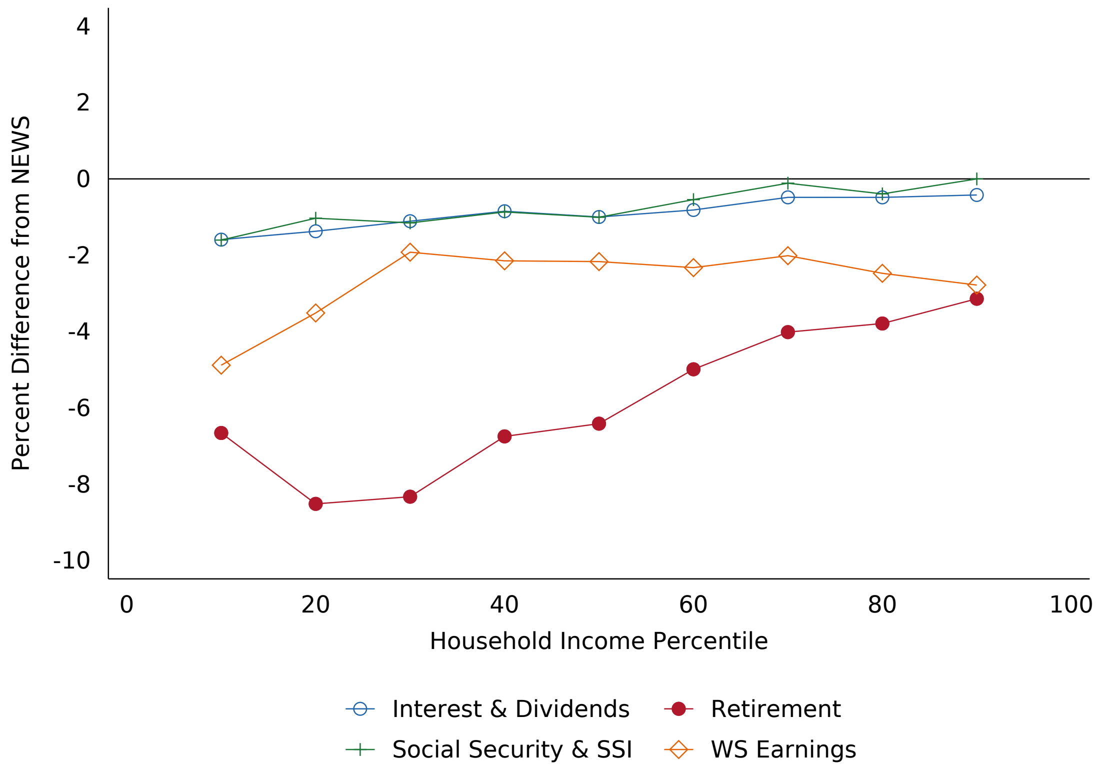
Notes: This figure was published in O'Hara, Bee and Mitchell (2017) and is replicated here with permission, as it is no longer possible to disclose scatter plots of individual earnings reports. The figure compares individual survey wage and salary earnings reports to W-2 earnings from the 2011 ACS. The regression fit line is shown and the 45° is visible in the clustering of points below the regression line on the left side of the figure and above the regression fit on the right. While the survey reports cluster around the 45° line, there is considerable noise in the survey relative to the administrative reports, and the figure is consistent with mean-reversion of survey relative to administrative reports (both in the location of points relative to the diagonal and the fact that $\hat{\beta} < 1$). The axes are unlabeled as a condition of the original release.
Source: O'Hara, Bee and Mitchell (2017) using 2011 American Community Survey data linked to 2010 W-2s.

Figure 7: NEWS Estimate of Household Income Relative to Survey by Subgroup in 2018



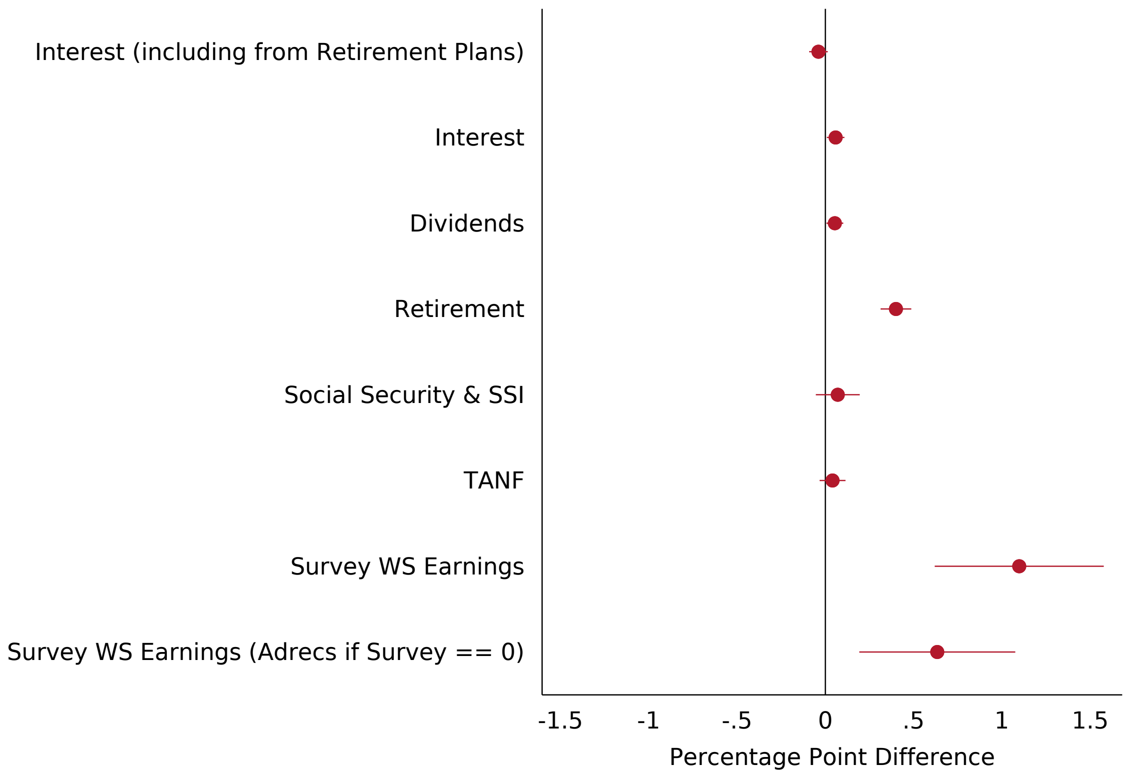
Notes: This figure shows the percent difference between the NEWS estimates of household income compared to the survey estimate at the 10th, 25th, 50th, 75th, and 90th percentiles in 2018.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure 8: Effect of Removing Individual Administrative Income Items on Household Income



Notes: In this figure, we replace individual income items from the NEWS estimates with the corresponding survey information and compare the estimate after replacement with the NEWS estimate. An estimate below the zero line indicates that administrative item increases income at that percentile. We replace: (1) interest and dividends, (2) retirement income, including withdrawals from Defined Contribution plans and retirement, survivor, and disability pensions. For interest and dividends, we exclude survey-reported interest earned in Defined Contribution retirement plans. For wage and salary earnings, we replace administrative wage and salary earnings with survey responses in all cases where the individual does not have administrative self-employment earnings, even if the individual reported no earnings on the survey. More detailed decompositions are available in Figure A3.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure 9: Effect of Removing Individual Administrative Income Items on Poverty



Notes: In this figure, we replace individual income items from the NEWS estimates with the corresponding survey information, including for interest, dividends, retirement income, Social Security, SSI, TANF, and survey wage and salary earnings. An estimate above the zero line indicates that administrative item decreases overall poverty. For survey interest, we show two measures, including and excluding the interest earned in Defined Contribution retirement plans such as 401(k)s. We replace Social Security and SSI together to address misclassification across programs, as discussed in Bee and Mitchell (2017). We replace administrative wage and salary earnings with two survey-based earnings measures. In the first, we use survey responses in all cases where the individual does not have administrative self-employment earnings, even if the individual reported no earnings on the survey. In the second, we only replace administrative wage and salary earnings if the survey report was positive. Retirement includes Defined Contribution plan withdrawals, pensions, and survivor and disability pensions.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 1: NEWS Median Household Income Estimates Compared to Survey in 2018

	Survey			NEWS			Percent Difference (NEWS - Survey)	
	Number	Median Income (dollars)		Number	Median Income (dollars)			
Characteristic	(thousands)	Estimate	95 percent CI	(thousands)	Estimate	95 percent CI	Estimate	95 percent CI
HOUSEHOLDS								
All Households	128,600	63,180	823	133,700	67,170	962	6.3***	1.4
Type of Household								
Family households	83,480	80,660	791	85,840	85,210	1,221	5.6***	1.3
.Married-couple	61,960	93,650	1,340	63,950	98,100	1,402	4.7***	1.4
.Female householder, no husband present	15,040	45,130	1,329	15,250	47,490	1,754	5.2***	3.6
.Male householder, no wife present	6,480	61,520	1,485	6,644	63,550	2,798	3.3	4.4
Nonfamily households	45,100	38,120	983	47,890	41,800	846	9.6***	2.7
.Female householder	23,510	32,010	794	24,860	38,010	1,201	18.7***	3.6
.Male householder	21,580	45,750	1,034	23,030	46,230	1,212	1.0	2.5
Race and Hispanic Origin of Householder								
White	100,500	66,940	769	104,000	71,390	984	6.6***	1.3
..White, not Hispanic	84,730	70,640	777	87,370	74,210	1,166	5.1***	1.4
Black	17,170	41,360	1,079	18,290	43,100	2,058	4.2*	4.3
Asian	6,981	87,190	3,342	7,019	89,270	5,614	2.4	5.6
Hispanic (any race)	17,760	51,450	876	18,400	57,710	2,314	12.2***	4.2
Age of Householder								
Under 65 years	94,420	71,660	683	99,370	71,580	1,001	-0.1	1.2
..15 to 24 years	6,199	43,530	3,204	6,961	41,350	2,245	-5.0	6.4
..25 to 34 years	20,610	65,890	1,281	22,080	65,110	1,764	-1.2	2.3
..35 to 44 years	21,370	80,740	1,276	22,490	78,600	2,390	-2.7*	2.7
..45 to 54 years	22,070	84,460	2,198	23,000	84,940	2,017	0.6	2.7
..55 to 64 years	24,170	68,950	1,720	24,840	72,430	1,975	5.0***	2.9
65 years and older	34,160	43,700	972	34,360	55,610	1,370	27.3***	3.0
Nativity of Householder								
Native born	108,600	64,240	848	114,100	67,680	981	5.3***	1.3
Foreign born	20,020	58,780	1,891	19,670	64,140	2,322	9.1***	3.9
..Naturalized citizen	11,040	65,520	2,682	10,480	72,290	2,877	10.3***	4.6
..Not a citizen	8,976	51,940	1,254	9,193	55,670	4,458	7.2*	8.3
Region								
Northeast	22,050	70,110	2,247	22,840	76,810	2,876	9.6***	3.4
Midwest	27,690	64,070	1,722	28,730	66,460	1,726	3.7***	2.5
South	49,740	57,300	978	52,470	58,890	1,418	2.8**	2.2
West	29,100	69,520	1,900	29,700	77,560	2,366	11.6***	3.1
Residence								
Inside metropolitan statistical areas	110,800	66,160	725	112,600	71,010	1,049	7.3***	1.4
..Inside principal cities	42,980	59,360	1,457	43,040	63,210	1,653	6.5***	2.4
..Outside principal cities	67,810	70,930	902	69,520	75,780	1,522	6.8***	1.8
Outside metropolitan statistical areas	17,790	49,870	1,941	21,170	50,040	1,722	0.3	3.2
Education								
Age 25 and Above	122,400	64,760	806	126,800	69,200	963	6.8***	1.4
No HS	11,230	28,330	1,260	11,850	32,400	1,599	14.4***	5.3
HS	31,810	46,070	870	33,270	50,630	999	9.9***	2.3
Some College	33,940	60,940	918	35,090	64,620	1,432	6.0***	2.0
Bachelor's and Above	45,410	101,800	1,135	46,550	105,400	1,940	3.5***	1.6

Notes: This table compares the NEWS median household income estimates to the survey estimates by subgroup in 2018. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for percent differences. Federal surveys give respondents the option of reporting more than one race. Therefore, two basic ways of defining a race group are possible. A group, such as Asian, may be defined as those who reported Asian and no other race (the race-alone or single-race concept) or as those who reported Asian regardless of whether they also reported another race (the race-alone-or-in-combination concept). This table shows data using the first approach (race alone). The use of the single-race population does not imply that it is the preferred method of presenting or analyzing data. The Census Bureau uses a variety of approaches. About 2.9 percent of people reported more than one race in the 2010 Census.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.
separately.

Table 2: NEWS Poverty Estimates Compared to Survey in 2018

	Survey		NEWS		Change in poverty (NEWS - Survey)	
Characteristic	**Percent**	**95 percent CI**	**Percent**	**95 percent CI**	**Difference**	**95 percent CI**
PEOPLE						
....Total	**11.78**	**0.29**	**10.67**	**0.39**	**-1.11*****	**0.37**
Race and Hispanic Origin						
White	10.07	0.30	8.91	0.40	-1.16***	0.41
...White, not Hispanic	8.07	0.28	7.48	0.35	-0.59***	0.35
Black	20.77	1.16	20.10	1.46	-0.67	1.31
Asian	10.10	0.94	8.52	1.41	-1.57**	1.38
Hispanic (any race)	17.56	0.80	14.61	1.14	-2.95***	1.19
Sex						
Male	10.57	0.32	9.71	0.40	-0.86***	0.40
Female	12.94	0.33	11.59	0.48	-1.34***	0.45
Age						
Under 18 years	16.20	0.67	15.62	0.86	-0.57	0.83
18 to 64 years	10.68	0.29	9.97	0.37	-0.71***	0.35
65 years and older	9.75	0.46	6.42	0.45	-3.33***	0.56
Nativity						
Native-born	11.45	0.31	10.48	0.40	-0.97***	0.38
Foreign-born	13.79	0.67	11.86	0.97	-1.93***	1.01
...Naturalized citizen	9.93	0.75	9.07	0.99	-0.86*	1.01
...Not a citizen	17.46	1.01	14.40	1.59	-3.06***	1.63
Region						
Northeast	10.28	0.66	9.14	0.86	-1.14**	0.87
Midwest	10.37	0.66	10.51	0.83	0.14	0.75
South	13.57	0.55	12.24	0.66	-1.33***	0.66
West	11.22	0.64	9.41	0.83	-1.80***	0.83
Residence						
Inside metropolitan statistical areas	11.34	0.32	10.11	0.43	-1.23***	0.39
...Inside principal cities	14.59	0.65	13.47	0.74	-1.12***	0.70
...Outside principal cities	9.42	0.40	8.18	0.47	-1.24***	0.45
Outside metropolitan statistical areas	14.68	0.99	13.93	1.14	-0.75	0.98
Disability Status						
....Total, aged 18 to 64	10.68	0.29	9.97	0.37	-0.71***	0.35
With a disability	25.72	1.32	26.64	1.66	0.92	1.58
With no disability	9.46	0.25	8.68	0.36	-0.78***	0.35
Educational Attainment						
....Total, aged 25 and older	9.90	0.24	8.62	0.32	-1.27***	0.30
No high school diploma	25.90	1.05	21.96	1.36	-3.94***	1.41
High school, no college	12.73	0.47	10.83	0.56	-1.90***	0.56
Some college	8.38	0.38	8.05	0.51	-0.33	0.52
Bachelor's degree or higher	4.37	0.32	3.65	0.33	-0.72***	0.38

Notes: This table compares the NEWS poverty estimates to the survey estimates by subgroup in 2018. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 3: NEWS Inequality Estimates Compared to Survey in 2018

	Survey		NEWS		Percent Difference (NEWS - Survey)	
Measure	**Estimate**	**95 percent CI**	**Estimate**	**95 percent CI**	**Estimate**	**95 percent CI**
Shares of Aggregate Income						
1st Quintile	0.036	0.001	0.037	0.001	0.001	0.001
2nd Quintile	0.091	0.001	0.089	0.002	-0.002*	0.002
3rd Quintile	0.148	0.001	0.142	0.003	-0.005***	0.003
4th Quintile	0.227	0.002	0.215	0.004	-0.012***	0.004
5th Quintile	0.498	0.004	0.516	0.009	0.018***	0.008
Top 5 Percent	0.218	0.005	0.252	0.012	0.034***	0.012
Summary Measures						
Gini Index	0.459	0.004	0.476	0.009	0.017***	0.009
90/10 percentile ratio	12.52	0.34	11.52	0.36	-1.00***	0.35
90/50 percentile ratio	2.92	0.04	2.82	0.04	-0.10***	0.05
50/10 percentile ratio	4.29	0.10	4.09	0.10	-0.20***	0.11

Notes: This table compares NEWS inequality statistics to the survey estimates in 2018. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for percent differences.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 4: Data Sources

File	Data Source	Description
Current Population Survey Annual Social and Economic Supplement (CPS ASEC)	Census	Annual survey fielded in February to April with household structure and characteristics at the time of interview and income from the prior calendar year. About 95,000 housing units sampled each year.
American Community Survey (ACS)	Census	Rolling survey fielded throughout the year about income from prior 12 months. About 3.5 million housing units sample each year.
Short Form Decennial Census	Census	Complete count decennial census data from 2000 and 2010.
Master Address File (MAF)	Census	File of residential addresses used to support census survey and decennial operations. Survey samples are drawn from this file for both the CPS ASEC and ACS.
Master Address File Auxiliary Reference File (MAFARF)	Census	Comingled file constructed from administrative records, including the IRMF, postal service change of address information, program data, etc. that links individuals (identified by Protected Identification Keys) to addresses in the Master Address File (identified by MAFIDs).
Longitudinal Business Database (LBD)	Census	Database of private non-farm establishments with employees from 1976 forward. For each establishment the LBD has information on industry, payroll, employment, and a firm identifier to group establishments into firms.
Information Returns Master File (IRMF)	IRS	Universe file with flags for whether an individual received each of the following information returns forms: 1098, 1099-DIV, 1099-INT, 1099-G, 1099-MISC, 1099-R, 1099-S, SSA-1099, and W-2. No income information is available. Also contains address information which has matched to the MAF to get a MAFID for each form.
Form 1040 Tax Returns (1040s)	IRS	Universe tax filings with a subset of the information on the complete Form 1040. The extracts provided by the IRS include information on tax-unit wage and salary income, gross rental income, taxable social security income, taxable and tax-exempt interest income, interest income, dividends, Adjusted Gross Income, and a constructed measure of Total Money Income (TMI). TMI is the sum of taxable wage and salary income, interest (taxable and tax-exempt), dividends, gross social security income, unemployment compensation, alimony received, business income or losses (including for partnerships and S-corps), farm income or losses, and net rent, royalty, and estate and trust income. Self-employment income is not available (except as a component of TMI), but flags exist for the filing of different 1040 schedules (such as C, D, E, F, SE).
Form W-2 (W-2s)	IRS	Universe data with a subset of information from the Form W-2. The extracts provided by the IRS include select boxes from the form, including wages and salary net of pre-tax deductions for health insurance premiums and deferred compensation (boxes 1 and 5), as well as the total amount of deferred compensation (summed values from Box 12 Codes D-H). Employee and employer pre-tax contributions to health insurance premiums are not available in the W-2 data.
Form 1099-R (1099-Rs)	IRS	Universe data with a subset of information from the Form 1099-R. The extracts provided by the IRS include information on amounts of defined-benefit pension payments (including for survivor and disability pensions) and withdrawals from defined-contribution retirement plans.
Numerical Identification System (Numident)	SSA	The Numident contains information for anyone ever to have received a Social Security Number. It includes information on date and place of birth, date of death, sex, and some information on citizenship.
Payment History Update System (PHUS)	SSA	Monthly Old Age, Survivors, and Disability Insurance (OASDI) payments from 1984 to the present. The PHUS exists for several subsamples of individuals including 1) those receiving payments in 2020 and 2021, 2) CPS ASEC respondents in linked years, and 3) ACS respondents in linked years (currently only 2019).
Supplemental Security Record (SSR)	SSA	Monthly Supplemental Security Income (SSI) payments from 1984 to the present for federally SSI and federally administered state SSI. The SSR exists for several subsamples of individuals including 1) those receiving payments in 2020 and 2021, 2) CPS ASEC respondents in linked years, and 3) ACS respondents in linked years (currently only 2019).
Detailed Earnings Record (DER)	SSA	Annual job-level income (by Employer Identification Number, EIN) from Form W-2s and annual positive self-employment income (from Form 1040 Schedule SE). The DER exists for several subsamples: 1) CPS ASEC respondents in linked years and 2) ACS respondents in linked years (currently only 2019)
Longitudinal Employer Household Dynamics (LEHD)	States	Quarterly job earnings reports from firms to state Unemployment Insurance offices for participating states. For covered jobs, the LEHD includes gross earnings - this includes employee contributions for health insurance premiums not available on the W-2 extracts. Coverage in the LEHD is not complete as many government employees, such as federal civilian employees, postal workers, and Department of Defense employees are not covered by state UI benefits. Some private-sector employees, including those employed by religious organizations, are not covered by UI, and are therefore not present in the LEHD data.
Supplemental Nutrition Assistance Program	States	SNAP participant data from partner states. In 2018, SNAP data is available for 17 states.
Temporary Assistance for Needy Families (TANF)	States + HHS	TANF participant data from partner states as well as from the Department of Health and Human Services (HHS) for additional states. In 2018, TANF data is available for 36 states.
Black Knight Home Value (Black Knight)	Black Knight	Third party data on home values and housing unit characteristics.

Notes: This table describes the data used in this project, including the source of the data and a short description. The name for the data used in Figures 4 and 5 is in parenthesis.

Table 5: Measurement and Estimation Steps

Section	Step	Inputs	Category	Measurement Challenge	Description	Related Work
A. Weighting	1. Weight respondents	Address and Person Files	Survey	Survey unit nonresponse Selection into administrative data Administrative data “nonresponse”	Use linked information on all occupied housing units and population controls to weight respondent sample to be representative of the target universe of households	Rothbaum et al. (2021); Rothbaum and Bee (2022)
	2. Weight respondents with all adults assigned a PIK	Address and Person Files	Survey	Survey unit nonresponse Selection into administrative data Administrative data “nonresponse” Selection into Linkage	Use information from A1 and reweight households with all adults assigned a PIK to be representative of the target universe of households	
B. Imputation	1. Impute survey earnings	Person File	Survey	Survey item nonreponse	Impute survey earnings conditional on survey and administrative information	Hokayem, Raghunathan and Rothbaum (2022)
	2. Impute LEHD gross earnings	Person File	Admin	Administrative data “nonresponse” Conceptual misalignment Incomplete data coverage	Impute LEHD earnings when missing or there is large disagreement between W-2s and LEHD	
	3. Impute missing means-tested program benefits	Person File	Admin	Incomplete data coverage	Impute means-tested program data (TANF and SNAP) for states for which administrative data is not available	Fox et al. (2022)
	4. Impute administrative income for nonfilers	Person File and nonfiler income parameters	Admin	Selection into administrative data Incomplete data coverage	Impute unemployment insurance compensation, interest, and dividends for nonfilers	Rothbaum (2023)
C. Estimation	1. Earnings Measurement Error Model	Person File (for CPS ASEC and ACS)	Admin	Survey misreporting Administrative misreporting	Combine survey and administrative wage and salary earnings according to the earnings measurement error model	Bee et al. (2023)
	2. Income replacement	Person File	Admin	Survey misreporting Administrative misreporting	Use survey and administrative data, imputed income, and earnings from the measurement error model to construct household and family income	Bee and Mitchell (2017)
	3. Estimate income and poverty statistics	Person File	Admin			

Notes: This table describes the processing steps used to address measurement error and estimate income and poverty. For each step, we include the Category (Survey or Administrative) matching the breakdown used in the decomposition used in Figure 3. Each step also references the relevant measurement challenges discussed in Section 2 and related work done at the Census Bureau that is being integrated into the NEWS project and extended.

Table 6: Rates of Missing Data for Imputed Income Items

	Missingness Rate
Survey	
Earnings from Primary Job	0.456
	(0.003)
Earnings from Other Employers	
Wage and Salary	0.367
	(0.007)
Self Employment	0.445
	(0.014)
Farm Self Employment	0.574
	(0.020)
Usual Hours Worked Per Week	0.260
	(0.003)
Weeks Worked Last Year	0.250
	(0.003)
Administrative	
Job 1 LEHD (gross earnings) missing | W-2 or DER not missing	0.080
	(0.001)
or large disagreement between LEHD and W-2	0.178
	(0.002)
Job 2 LEHD (gross earnings) missing | W-2 or DER not missing	0.120
	(0.002)
or large disagreement between LEHD and W-2	0.184
	(0.003)
SNAP administrative data unavailable	0.695
	(0.001)
TANF administrative data unavailable	0.474
	(0.001)

Notes: This table shows the share of the 2019 CPS ASEC sample that is missing information for the various items imputed in this work, as discussed in Section 4.2. Standard errors in parenthesis. Jobs are ordered in the administrative data (Job 1, Job 2, etc.) from highest to lowest earnings across the three sources of job-level earnings (W-2, DER, and LEHD).
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 7: Sources of Administrative and Survey Earnings

A. All Individuals

Administrative Earnings Sources				Share with Unimputed Survey:	
W-2	DER	LEHD	N	Wage and Salary Earnings	Self-Employment Earnings
X	X	X	72,000	0.887	0.029
				(0.002)	(0.001)
X	X		5,900	0.704	0.033
				(0.010)	(0.003)
X		X	400	0.105	0.034
				(0.018)	(0.011)
X			300	0.804	0.024
				(0.036)	(0.011)
	X	X	30	1.000	Z
				Z	Z
	X		<15	Z	Z
				Z	Z
		X	500	0.244	0.058
				(0.026)	(0.016)
			75,000	0.045	0.027
				(0.001)	(0.001)

B. Citizenship and DER Earnings

Administrative Earnings Sources			N (Survey Earnings Respondents Only)		Share Reporting: Wage and Salary Earnings		Share Reporting: Self-Employment Earnings	
W-2	DER	LEHD	In Numident	Not In Numident	In Numident	Not In Numident	In Numident	Not In Numident
X	X	Yes or No	47,000	<15	0.874	Z	0.029	Z
					(0.002)	Z	(0.001)	Z
X		Yes or No	350	200	0.093	0.847	0.035	0.023
					(0.018)	(0.033)	(0.011)	(0.011)

Notes: This table shows the counts and share of adults with each possible administrative earnings data source (W-2, DER, and LEHD) as well as the share in each group that reported survey earnings (among those that responded to the survey earnings questions). Panel A shows the estimates for all individuals in the CPS ASEC. Panel B shows how the presence or absence of DER earnings given W-2 earnings is related to differential probability of reporting survey earnings for individuals who can be assigned PIKs that have SSNs (In Numident) and do not (Not In Numident). Z indicates an estimate rounds to zero. Standard errors in parenthesis.

Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 8: Combining Administrative and Survey Earnings: Use of Survey Earnings by Group

A. Race and Hispanic Origin

	Share Survey Earnings	
Race/Hispanic Origin	Overall	Relative to Average
All	20.6	Z
	(2.7)	(0.2)
Black	13.8	-6.8*
	(2.9)	(3.1)
Hispanic	22.1	1.5
	(2.9)	(1.2)
White Non-Hispanic	22.6	2.0
	(3.0)	(1.2)

B. Age

	Share Survey Earnings	
Age	Overall	Relative to Average
18-24	6.3	-14.3**
	(1.4)	(3.5)
25-34	29.0	8.4**
	(4.4)	(2.5)
34-44	26.8	6.3**
	(3.5)	(1.8)
45-54	20.5	-0.1
	(4.1)	(2.1)
55-64	16.2	-4.3*
	(3.3)	(2.1)
65+	8.7	-11.9***
	(2.6)	(2.4)

Notes: This table shows the share of individuals in each subgroup where survey earnings are used from the measurement error model for choosing survey or administrative earnings discussed in Section 4.3.1 and in more detail in Bee et al. (2023) Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences relative to average.

Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table 8 Combining Administrative and Survey Earnings: Use of Survey Earnings by Group, Continued

C. Occupation

Occupation (Last Week)	Share Survey Earnings: Overall	Share Survey Earnings: Relative to Average
Unemployed	14.2	-6.4
	(5.3)	(6.0)
Management	30.3	9.7**
	(5.6)	(3.0)
Business and Financial Operations	25.2	4.6
	(2.8)	(3.2)
Computer and Mathematical	41.5	20.9**
	(7.2)	(6.9)
Architecture and Engineering	52.3	31.7***
	(4.0)	(2.9)
Life, Physical, and Social Science	9.1	-11.5***
	(2.1)	(2.2)
Community and Social Services	3.1	-17.5***
	(1.8)	(3.3)
Legal	11.0	-9.6
	(11.0)	(8.5)
Education, Training, and Library	8.8	-11.8***
	(4.2)	(2.5)
Arts, Design, Entertainment, Sports, and Media	7.5	-13.1**
	(2.7)	(3.5)
Healthcare Practitioners and Technical	21.9	1.3
	(3.8)	(2.0)
Healthcare Support	4.1	-16.4***
	(1.6)	(3.8)
Protective Service	15.4	-5.2
	(3.5)	(5.8)
Food Preparation and Serving Related	10.2	-10.4
	(9.8)	(7.7)
Building and Grounds Cleaning and Maintenance	15.1	-5.5
	(6.1)	(3.9)
Personal Care and Service	8.8	-11.8*
	(4.0)	(4.8)
Sales and Related	11.9	-8.7***
	(1.2)	(1.8)
Office and Administrative Support	16.9	-3.7
	(1.9)	(1.9)
Farming, Fishing, and Forestry	61.1	40.5
	(24.3)	(22.3)
Construction Trades and Extraction Workers	42.2	21.6
	(11.3)	(10.6)
Installation, Maintenance, and Repair Workers	38.4	17.8**
	(4.6)	(5.9)
Production Occupations	20.5	-0.1
	(5.1)	(3.7)
Transportation	11.9	-8.7**
	(2.6)	(2.7)
Material Moving	29.9	9.3*
	(5.4)	(4.2)

D. Industry

Industry (Last Week)	Share Survey Earnings: Overall	Share Survey Earnings: Relative to Average
Unemployed	14.2	-6.4
	(5.3)	(6.0)
Agriculture, Forestry, Fishing, and Hunting	64.1	43.5
	(30.7)	(28.7)
Mining	29.2	8.6
	(11.2)	(8.8)
Construction	58.6	38.0**
	(12.1)	(11.5)
Manufacturing	18.9	-1.7
	(6.6)	(5.1)
Wholesale Trade	13.5	-7.1
	(7.6)	(8.4)
Retail Trade	4.2	-16.4***
	(1.5)	(2.8)
Transportation and Warehousing	17.2	-3.4
	(6.6)	(5.8)
Utilities	6.8	-13.8*
	(5.9)	(6.4)
Information	23.9	3.3
	(8.4)	(8.2)
Finance and Insurance	43.8	23.2*
	(8.1)	(10.2)
Real Estate and Rental and Leasing	79.0	58.4***
	(11.3)	(11.7)
Professional, Scientific, and Technical Services	36.2	15.7
	(11.6)	(11.1)
Management of companies and enterprises	2.0	-18.6***
	(3.6)	(4.5)
Administrative and support and waste management services	22.8	2.2
	(11.2)	(9.2)
Educational Services	9.8	-10.8***
	(3.7)	(2.1)
Health Care and Social Assistance	10.9	-9.7***
	(2.3)	(1.7)
Arts, Entertainment, and Recreation	39.3	18.7
	(24.6)	(23.7)
Accommodation and Food Service	14.4	-6.2
	(14.5)	(12.4)
Other Services	27.0	6.4
	(9.3)	(10.4)
Public Administration	7.4	-13.2
	(4.7)	(7.0)

Notes: This table shows the share of individuals in each subgroup where survey earnings are used from the measurement error model for choosing survey or administrative earnings discussed in Section 4.3.1 and in more detail in Bee et al. (2023) Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences relative to average.

Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A1: Simple Job Linkage Example

W-2 Jobs

PIK	EIN	Earnings
1	100	10,000
2	100	20,000
2	400	12,000
3	100	5,000
3	500	200
3	600	2,600

LEHD Jobs

PIK	EIN	Earnings
1	200	11,000
2	200	20,005
2	400	12,000
3	200	5,200
3	500	225

Direct Matches

	W-2		LEHD	
PIK	EIN	Earnings	EIN	Earnings
2	400	12,000	400	12,000
3	500	200	500	225

Indirect Matches

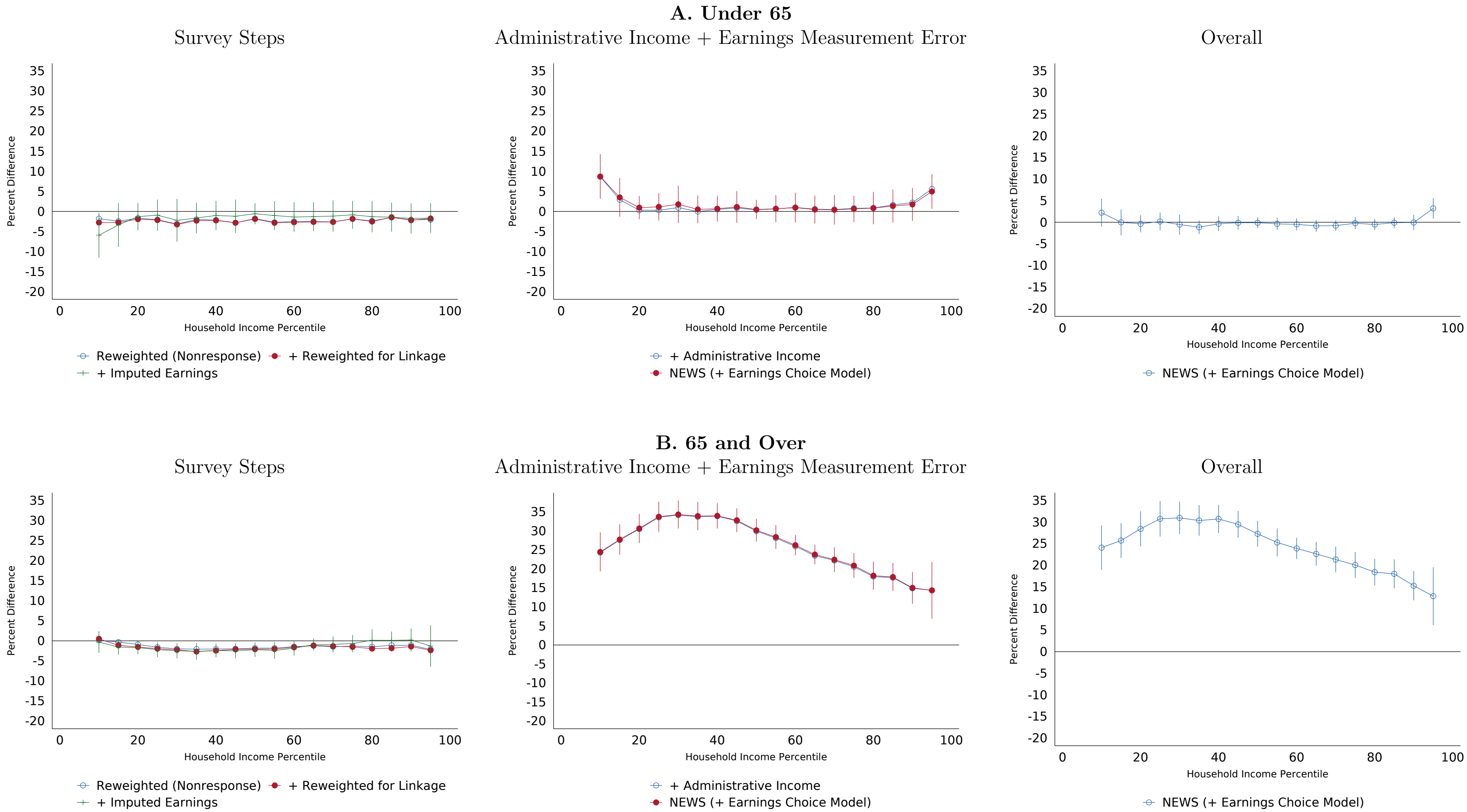
	W-2		LEHD	
PIK	EIN	Earnings	EIN	Earnings
1	100	10,000	200	11,000
2	100	20,000	200	20,005
3	100	5,000	200	5,200

Unmatched

	W-2		LEHD	
PIK	EIN	Earnings	EIN	Earnings
3	600	2,600		

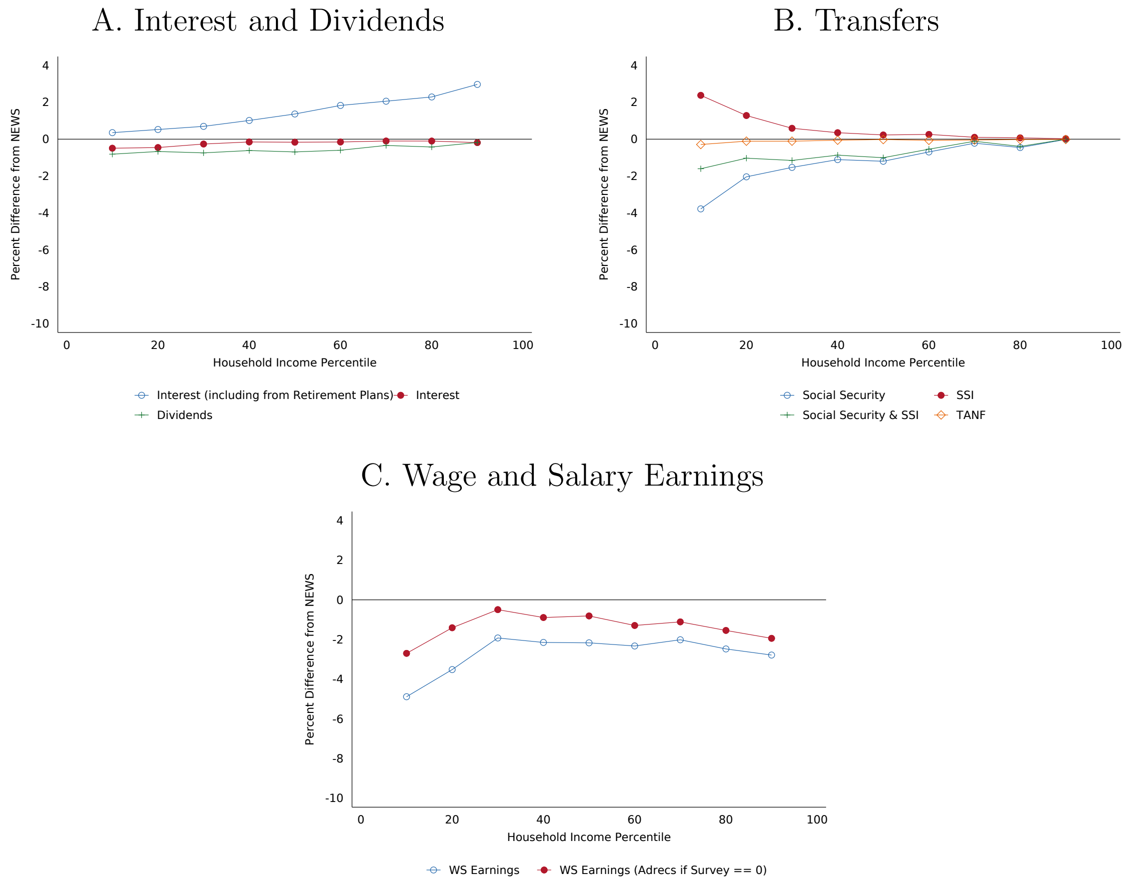
Notes: This is an example of how jobs are linked between W-2s and the LEHD (all PIKS, earnings, and EINs in the example are made up and do not correspond to actual individuals or firms). First and easiest are the jobs that match on PIK and EIN (same person, same firm identifier), which we call direct matches. Next, we find the indirect matches, where each person has one EIN on the W-2s and another on the LEHD (same person, but different firm identifiers on the two files). In this example, everyone with W-2 EIN = 100 has a job with similar earnings on the LEHD, but with EIN = 200. Finally, there are jobs that remain unmatched and only exist on one file or the other.

Figure A2: Decomposition of NEWS Processing Steps By Age: Distribution of Household Income



Notes: This figure decomposes the impact of the NEWS processing steps on household income. In the first column, the figures show the adjustments made to the survey data, including reweighting and improved earnings imputation comparing household income after the adjustment to the survey estimate. In the second column, the figures show impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The full impact of all adjustments is shown in the third column. The 95 percent confidence interval for the last step is shown in each: for A comparing the estimate after earnings imputation to the survey estimate and for B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

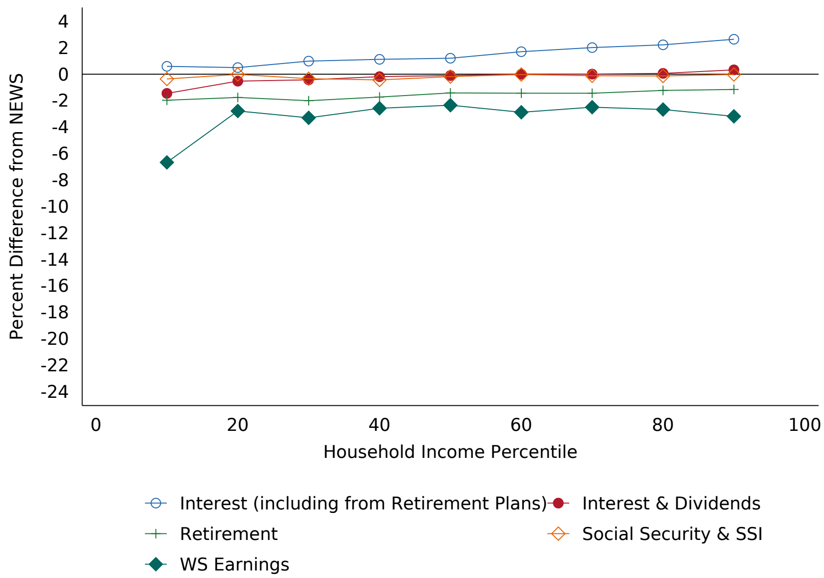
Figure A3: Effect of Removing Individual Administrative Income Items on Household Income, Additional Detail



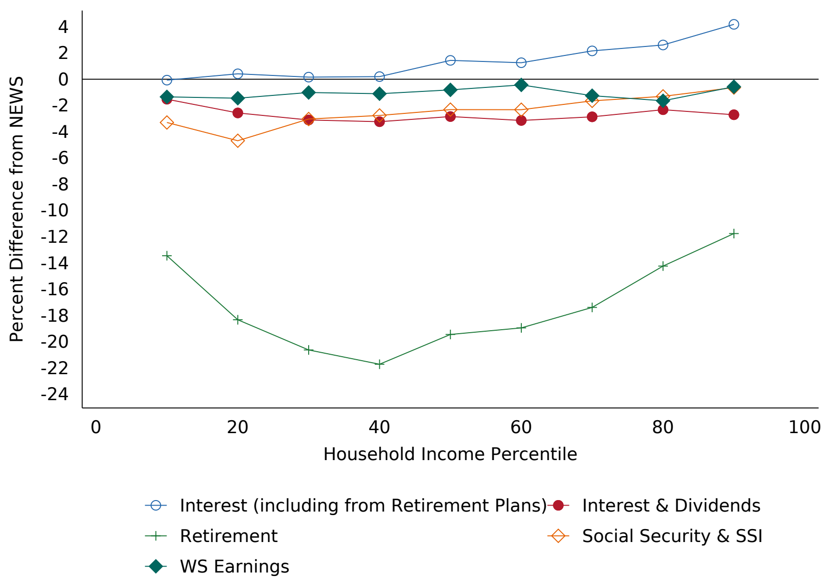
Notes: In this figure, we replace individual income items from the NEWS estimates with the corresponding survey information and compare the estimate after replacement with the NEWS estimate. An estimate below the zero line indicates that administrative item increases income at that percentile. In Panel A, we replace interest and dividend income with survey responses. For survey interest, we show two measures, including and excluding the survey-reported interest earned in Defined Contribution retirement plans such as 401(k)s. In Panel B, we replace Social Security and SSI separately and together (to address misclassification across programs, as discussed in Bee and Mitchell (2017)) and TANF with survey-reported public assistance income. In Panel C, we replace administrative wage and salary earnings with two survey-based earnings measures. In the first, we use survey responses in all cases where the individual does not have administrative self-employment earnings, even if the individual reported no earnings on the survey. In the second, we only replace administrative wage and salary earnings if the survey report was positive.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A4: Effect of Removing Individual Administrative Income Items on Household Income by Householder Age

A. Under 65

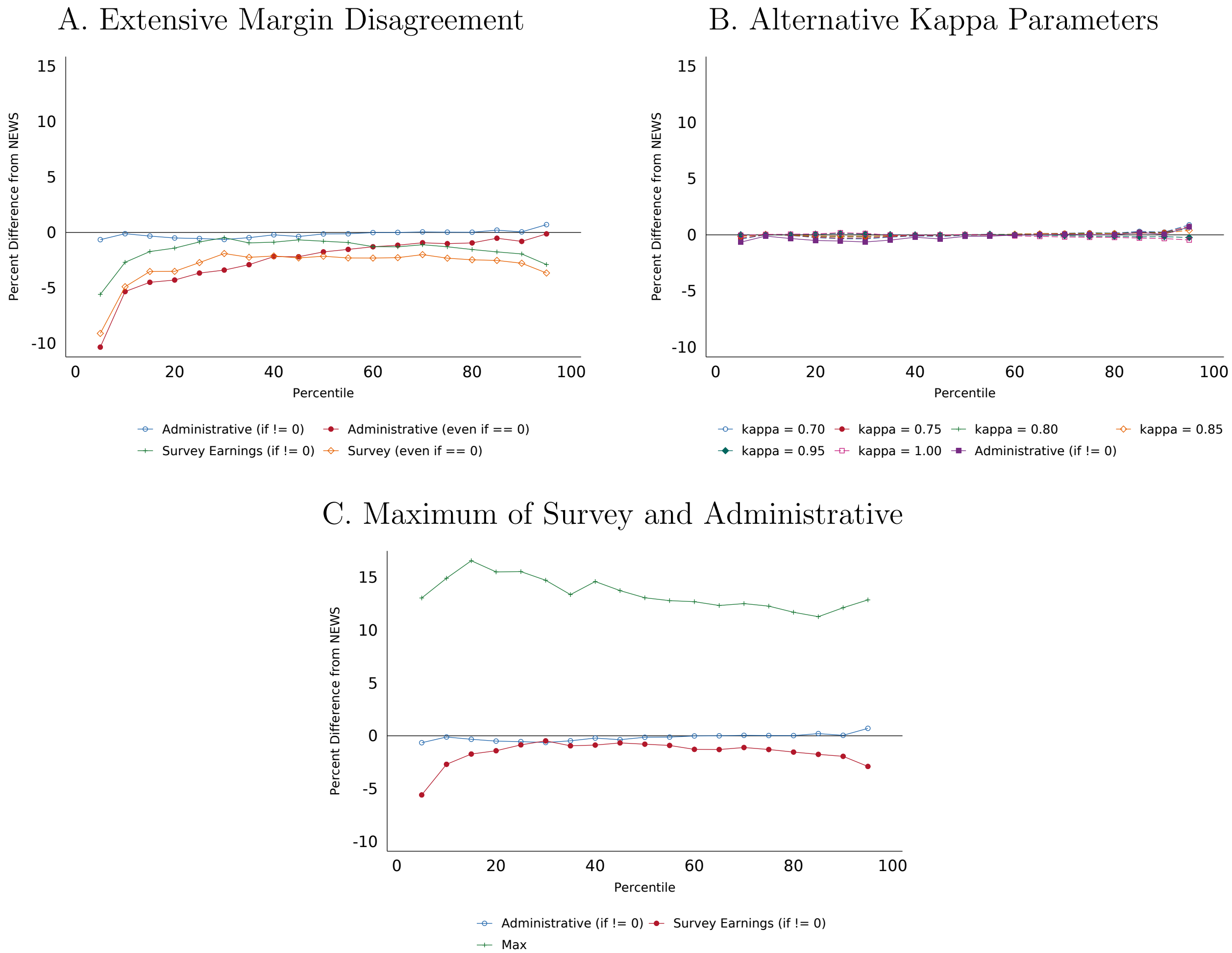


B. 65 and Over



Notes: In this figure, we replace individual income items from the NEWS estimates with the corresponding survey information and compare the estimate after replacement with the NEWS estimate. An estimate below the zero line indicates that administrative item increases income at that percentile. We show each of the major administrative income items, including (1) interest (including and excluding the interest earned in Defined Contribution, DC, retirement plans such as 401(k)s), (2) interest (without DC plan interest) and dividends, (3) DC plan withdrawals, pensions, and survivor and disability pensions (Retirement), (4) Social Security and SSI, and (5) wage and salary earnings.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A5: Alternative Uses of Survey and Administrative Earnings

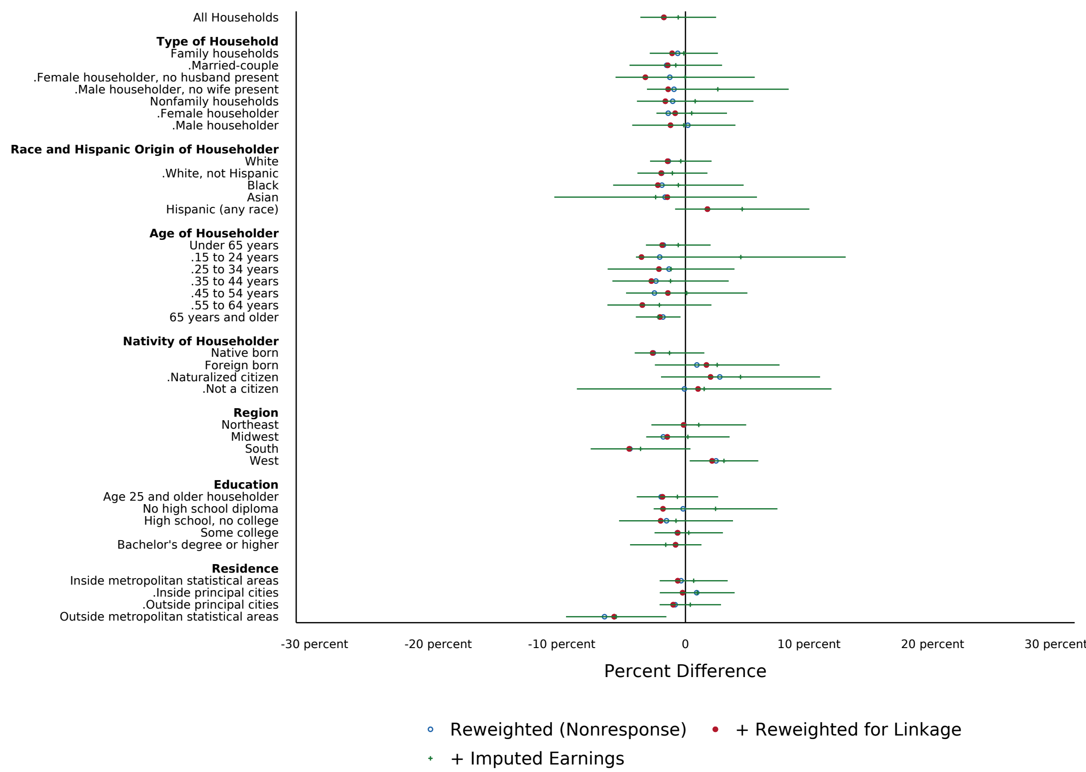


Notes: This figure shows the impact on household income (relative to the baseline NEWS estimates) of alternative uses of survey and administrative earnings in the income estimates. In Panel A, we show how income estimates vary when survey or administrative wage and salary earnings were used for individuals indicated as "Measurement error model" in Table A8. The four options in Panel A include: (1) Administrative earnings if they are not equal to 0, (2) administrative earnings even if they are equal to 0 and survey earnings are positive, (3) survey earnings if they are not equal to 0, and (4) survey earnings even if they are equal to zero and administrative earnings are positive. Panel B shows the impact on household earnings of alternative mean-reversion kappa parameters in the measurement error model (with the share of individual's whose survey earnings are used under each shown in Table A9). Panel B also includes (1) from Panel A, with administrative earnings if they are not equal to 0. Panel C compares the NEWS estimates to simpler uses of survey and administrative earnings, including (1) and (3) from Panel A and using the maximum of administrative and survey earnings.

Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A6: Decomposition of NEWS Processing Steps By Subgroup: Median Household Income

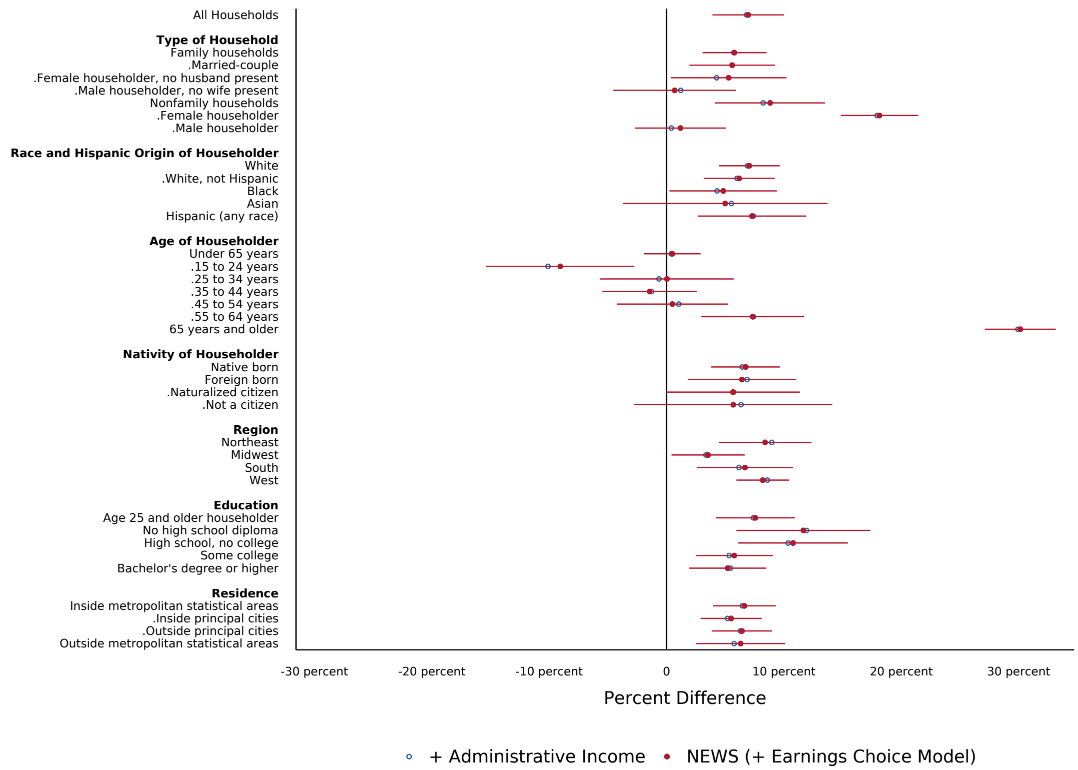
A. Survey Steps: Weighting and Earnings Imputation



Notes: This figure decomposes the impact of the NEWS processing steps on median household income. In Panel A, the figure shows the adjustments made to the survey data, including reweighting and improved earnings imputation comparing median household income for each group after the adjustment to the survey estimate. In Panel B, the figure shows impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The 95 percent confidence interval for the last step is shown in each: for Panel A comparing the estimate after earnings imputation to the survey estimate and for Panel B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A6: Decomposition of NEWS Processing Steps By Subgroup: Median Household Income, Continued

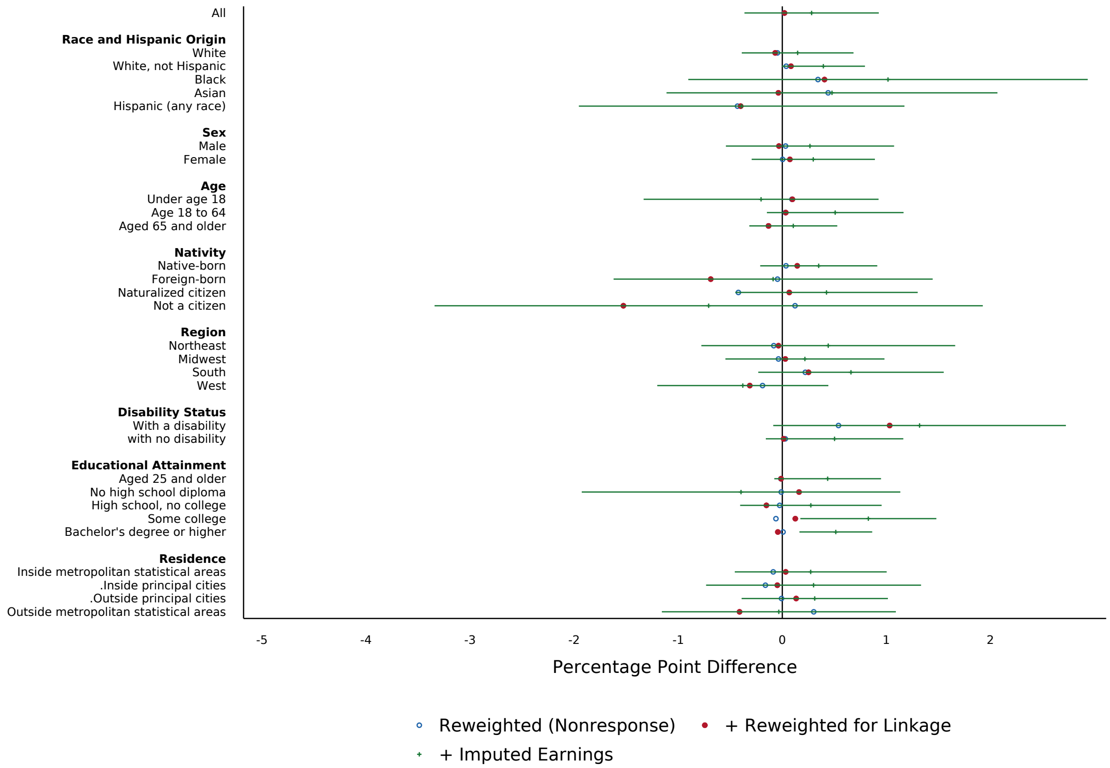
B. Administrative Income Replacement and Survey Earnings Choice Modeling



Notes: This figure decomposes the impact of the NEWS processing steps on median household income. In Panel A, the figure shows the adjustments made to the survey data, including reweighting and improved earnings imputation comparing median household income for each group after the adjustment to the survey estimate. In Panel B, the figure shows impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The 95 percent confidence interval for the last step is shown in each: for Panel A comparing the estimate after earnings imputation to the survey estimate and for Panel B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A7: Decomposition of NEWS Processing Steps By Subgroup: Poverty

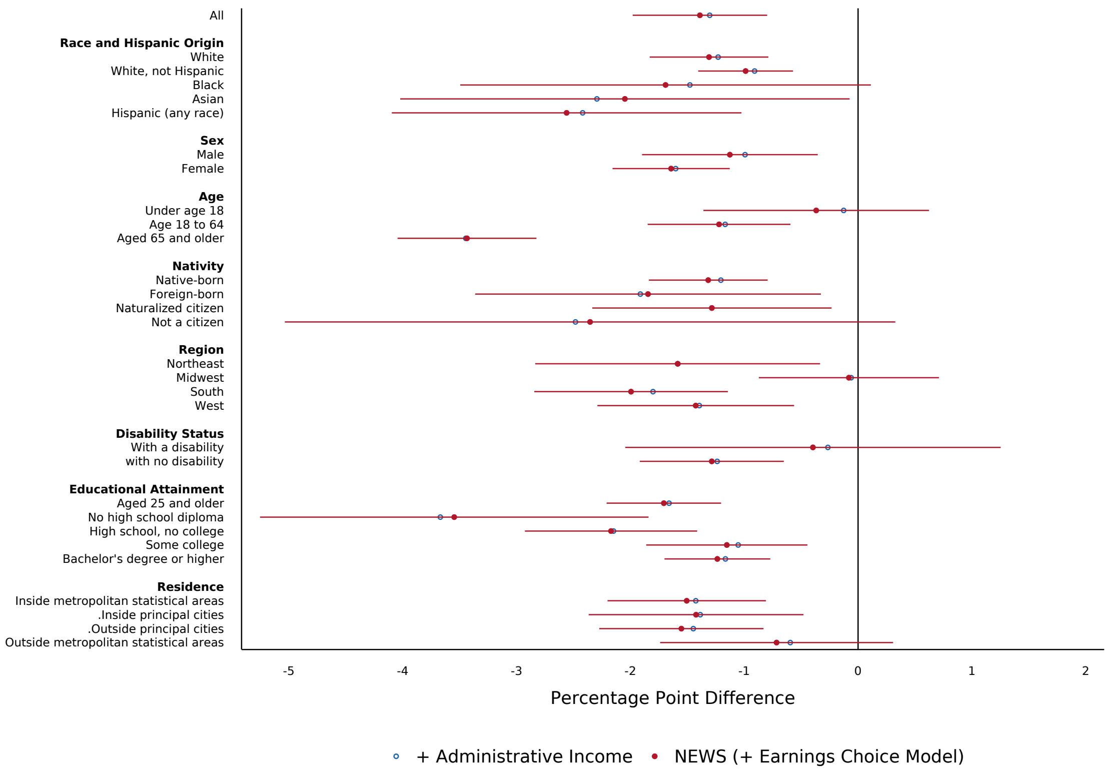
A. Survey Steps: Weighting and Earnings Imputation



Notes: This figure decomposes the impact of the NEWS processing steps on poverty. In Panel A, the figure shows the adjustments made to the survey data, including reweighting and improved earnings imputation comparing poverty for each group after the adjustment to the survey estimate. In Panel B, the figure shows impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The 95 percent confidence interval for the last step is shown in each: for Panel A comparing the estimate after earnings imputation to the survey estimate and for Panel B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

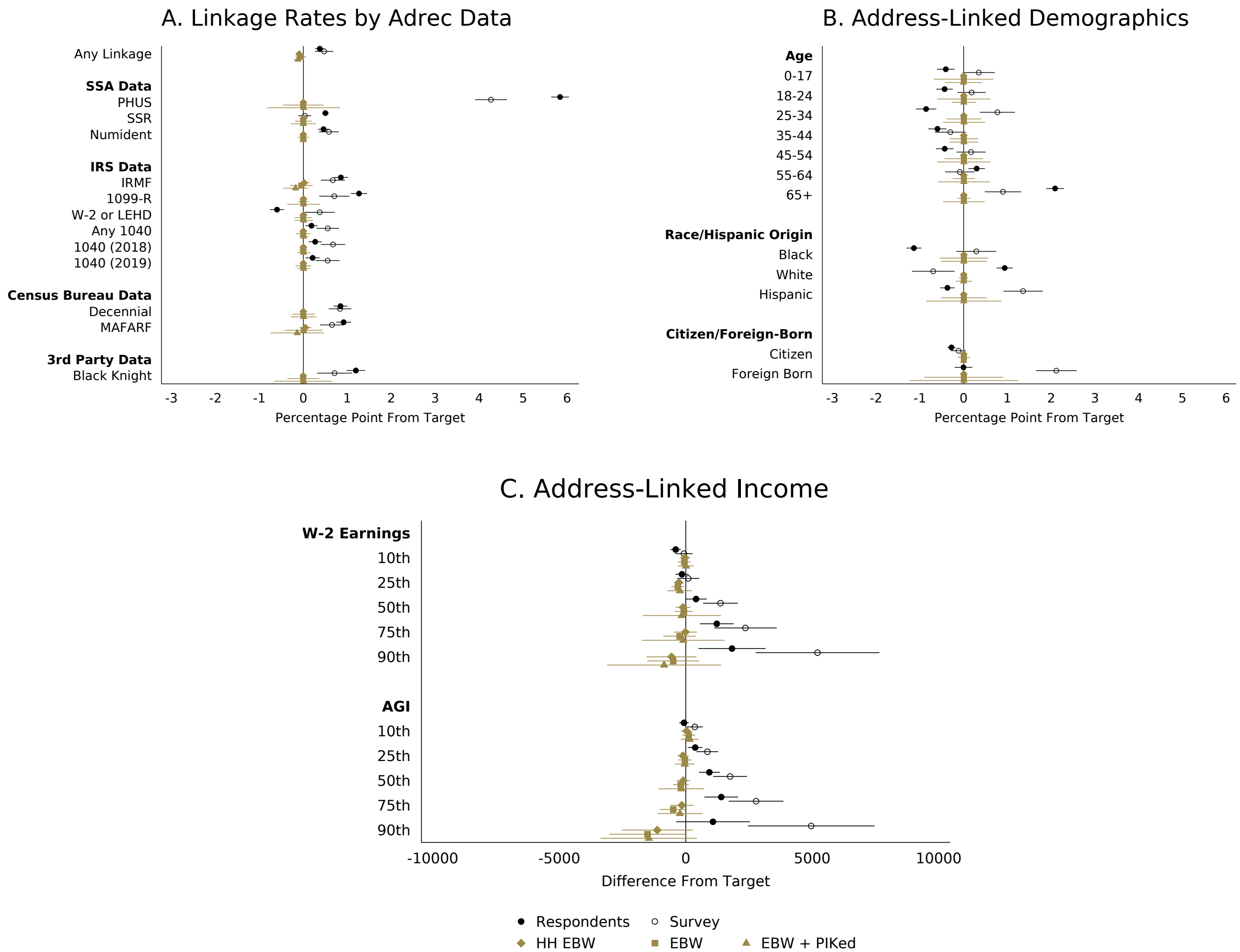
Figure A7: Decomposition of NEWS Processing Steps By Subgroup: Poverty, Continued

B. Administrative Income Replacement and Survey Earnings Choice Modeling



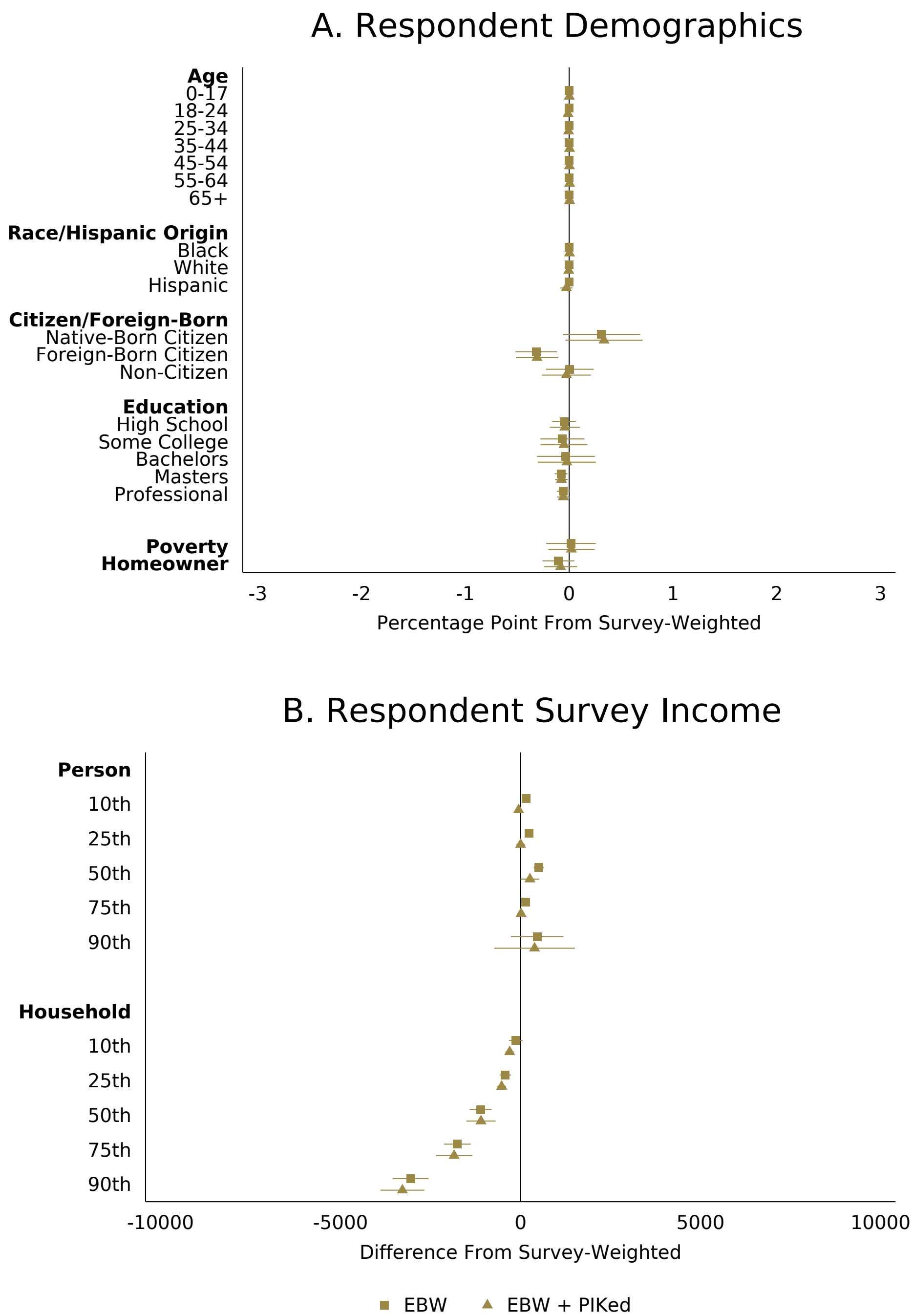
Notes: This figure decomposes the impact of the NEWS processing steps on poverty. In Panel A, the figure shows the adjustments made to the survey data, including reweighting and improved earnings imputation comparing poverty for each group after the adjustment to the survey estimate. In Panel B, the figure shows impact of replacing survey income responses with administrative income, comparing the estimates after each step to the estimates after reweighting and earnings imputation. The 95 percent confidence interval for the last step is shown in each: for Panel A comparing the estimate after earnings imputation to the survey estimate and for Panel B comparing the final NEWS estimate to the estimate after earnings imputation.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A8: Comparing Bias in Linked Administrative Characteristics with Different Weights



Notes: This figure shows various statistics of address-linked administrative, decennial census, and commercial data (refer to Section B.1) using different weights compared to the weighting targets (discussed in Appendix C and shown in Table A5). "Respondents" uses the base weights which adjust only for probability of selection into the sample. "Survey" uses the survey weights. "HH EBW" are the Stage 1 weights that adjust for selection into response at the household level. "EBW" are the Stage 2 weights that further adjust to population controls and "EBW + PIKed" are the Stage 3 weights that further adjust for selection into linkage.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Figure A9: Comparing Survey Characteristics with Different Weights



Notes: This figure shows various statistics of survey demographics and survey-reported income using the entropy balance weights (discussed Appendix C) relative to the survey-weighted estimates. "EBW" are the Stage 2 weights that further adjust to population controls and "EBW + PIKed" are the Stage 3 weights that further adjust for selection into linkage.

Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A1: Comparing Job-Level LEHD and W-2 Earnings

		Health Insurance		
LEHD-W-2 Comparison	All	Yes	No	Yes - No
LEHD < W-2	8.7	9.7	3.9	5.85***
	(0.2)	(0.2)	(0.3)	(0.30)
LEHD ≥ W-2				
0-1 percent greater	66.9	61.8	89.3	-27.52***
	(0.3)	(0.3)	(0.4)	(0.54)
1-3 percent greater	6.4	7.5	2.0	5.51***
	(0.1)	(0.2)	(0.2)	(0.26)
3-5 percent greater	4.9	5.8	1.3	4.50***
	(0.1)	(0.2)	(0.1)	(0.20)
5-10 percent greater	6.8	8.0	1.6	6.32***
	(0.1)	(0.2)	(0.2)	(0.24)
10+ percent greater	6.3	7.3	2.0	5.34***
	(0.1)	(0.2)	(0.2)	(0.25)
Observations	47,000	39,000	8,100	

Notes: This table shows basic summary statistics on job-level comparisons of LEHD earnings to W-2 earnings (including deferred compensation) for the highest earning job. Jobs are classified by the ratio of LEHD to W-2 earnings. The first category, W-2 > LEHD, indicates that W-2 earnings exceed LEHD earnings by more than a trivial amount ($100). The other categories indicate that LEHD gross earnings exceeded W-2 earnings + deferred compensation by specific percent ranges. Because LEHD gross earnings should exceed W-2 taxable earnings + deferred compensation primarily due to employee pre-tax contributions to health insurance premiums, the sample in this table includes only individuals that responded to the health insurance question in the CPS ASEC, i.e., whose health insurance status was not imputed. The first column shows the share in each LEHD-W-2 bin for all workers with a job in both data sources. The next two columns show estimates for those that reported having and not having private health insurance, respectively. The last column shows the difference between the share in each bin between those having and not having private health insurance. Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A2: Direct and Indirect Job Linkage Statistics

		EIN Matches Only		EIN and Indirect Matches	
	All Jobs	Unmatched Jobs	Share of Implied Total	Unmatched Jobs	Share of Implied Total
Total Jobs					
W-2	256,800,000	40,720,000	0.146	25,680,000	0.097
LEHD	237,900,000	21,780,000	0.078	6,744,000	0.026
EIN Matches	216,100,000		0.776		0.820
Indirect Matches	15,040,000				0.057
Implied Total Jobs		278,600,000		263,600,000	

Notes: This table shows the count of jobs that could be directly linked by Employer Identification Number (EIN) and indirectly linked as discussed in Section A.3.
Source: 2018 W-2 and Longitudinal Employer-Household Dynamics data.

Table A3: Weighted Linkage Rates by Administrative Data Source in the Address Data

	Target Estimate	Difference from Target			
	Base-Weighted	Base-Weighted	Survey Weighted	EBW-Weighted	
	Occupied Units	Respondent Units	Respondent Units	Respondent Units	Respondent + All Adults PIKed Units
Any Linkage	0.932***	0.0037***	0.0047***	-0.0006	-0.0012***
	(0.002)	(0.0006)	(0.0010)	(0.0006)	(0.0005)
SSA Data					
PHUS	0.402***	0.0584***	0.0427***	Z	0.0001
	(0.002)	(0.0010)	(0.0019)	(0.0024)	(0.0043)
SSR	0.050***	0.0050***	0.0003	Z	Z
	(0.001)	(0.0004)	(0.0007)	(0.0010)	(0.0015)
Numident	0.921***	0.0046***	0.0058***	Z	Z
	(0.002)	(0.0006)	(0.0012)	(0.0007)	(0.0004)
IRS Data					
IRMF	0.837***	0.0085***	0.0067***	-0.0005	-0.0018
	(0.002)	(0.0008)	(0.0014)	(0.0013)	(0.0014)
1099-R	0.436***	0.0127***	0.0070***	Z	Z
	(0.002)	(0.0010)	(0.0018)	(0.0006)	(0.0019)
Any 1040	0.856***	0.0018***	0.0055***	Z	0.0001
	(0.002)	(0.0007)	(0.0013)	(0.0009)	(0.0006)
1040 (2018)	0.828***	0.0027***	0.0068***	Z	0.0001
	(0.002)	(0.0008)	(0.0014)	(0.0005)	(0.0008)
1040 (2019)	0.835***	0.0021***	0.0055***	Z	0.0001
	(0.002)	(0.0008)	(0.0014)	(0.0009)	(0.0007)
W-2 or LEHD	0.751***	-0.0060***	0.0037**	Z	0.0001
	(0.002)	(0.0008)	(0.0017)	(0.0010)	(0.0010)
Census Bureau Data					
Decennial	0.867***	0.0084***	0.0083***	Z	0.0001
	(0.002)	(0.0008)	(0.0013)	(0.0013)	(0.0015)
MAFARF	0.822***	0.0092***	0.0065***	Z	-0.0014
	(0.002)	(0.0009)	(0.0014)	(0.0022)	(0.0031)
3rd Party Data					
Black Knight	0.644***	0.0119***	0.0071***	Z	Z
	(0.003)	(0.0011)	(0.0020)	(0.0019)	(0.0034)

Notes: This table shows statistics on selection into response at the household level by data source that can be linked to occupied housing units, as discussed in Section B.1. The target estimate is calculated on the base-weighted set of all occupied housing units in the March monthly CPS. The other estimates show differences from the target (evidence of selection into the sample unaddressed by weighting if $\neq 0$) for the indicated samples of respondents and weights. Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences. Z indicates an estimate rounds to zero.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A4: Linkage Rates by Administrative Data Source in the Person Data

	Full Sample		NEWS Sample (All Survey-Adults in HH Assigned PIK)	
	Survey-Adults (15+)	Survey-Children (<15)	Survey-Adults	Survey-Children
Assigned PIK	85.8	79.4	100.0	89.4
	(0.18)	(0.33)		(0.30)
Any Adrec Linked to Address				
If Assigned PIK	94.7	95.6	93.9	95.0
	(0.15)	(0.22)	(0.16)	(0.26)
If Not Assigned PIK	89.9	92.6		92.3
	(0.40)	(0.48)		(0.88)
Present In | Assigned PIK				
Any Administrative Record	98.1	85.2	98.0	87.4
	(0.05)	(0.30)	(0.07)	(0.33)
IRS Data				
Tax Filing (1040)	84.6	83.2	84.4	85.6
	(0.17)	(0.30)	(0.19)	(0.34)
IRMF	89.4	7.8	88.2	7.5
	(0.10)	(0.22)	(0.12)	(0.24)
W-2	64.3	1.0	63.9	1.0
	(0.16)	(0.07)	(0.17)	(0.08)
1099-R	21.1	0.1	20.1	Z
	(0.14)	(0.02)	(0.13)	(0.02)
SSA Data				
DER	67.6	0.3	67.2	0.3
	(0.16)	(0.04)	(0.17)	(0.05)
PHUS	37.8	3.9	35.2	3.5
	(0.16)	(0.16)	(0.16)	(0.16)
SSR	3.6	1.3	3.4	1.2
	(0.09)	(0.10)	(0.09)	(0.10)
State Data				
LEHD	64.3	1.0	63.9	1.0
	(0.16)	(0.07)	(0.17)	(0.08)

Notes: This table shows statistics on the individuals that can be assigned a PIK as well as the households in which those 15 and over (survey-adults) can be assigned a PIK. For all households and the 82 percent of households with all survey-adults assigned a PIK (the NEWS analysis sample), we show the share of survey-adults and survey-children that can be linked to various data sets. Estimates and standard errors that are 0 by construction are omitted. Z indicates an estimate rounds to zero. Standard errors in parenthesis.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A5: Entropy Balance Reweighting Procedure

Stage/Step	Moment Variables	Moment Sample	Reweighted Sample
1. Housing-unit level	Linked survey, administrative, and census variables	Non-vacant housing units in March Basic CPS (respondents and nonrespondents)	Respondent housing units
2. Person level			
A. Preserve distribution of housing unit characteristics	Linked survey, administrative, and census variables	Householders and householder-partners, using the housing-unit level weights from Stage 1	Householders and householder partners
B. Spousal equivalence	Linked survey, administrative, and census variables	Married couples and cohabiting partners	Married couples and cohabiting partners
C. External population targets	State-level population estimates by race, Hispanic-origin, gender, and age	External population estimates	All individuals
D. Match distribution of household characteristics in March Basic Sample	Subset of linked survey, administrative, and census variables and state-level population controls	Householders and householder partners in the March Basic File	Householders and householder partners in the full CPS ASEC sample
3. Address Selection into PIK assignment (for all adults in HH)			
A. Preserve distribution of respondent and housing unit characteristics	Linked survey, administrative, and census variables. Additional moments for survey-only and linked survey-administrative characteristics from full respondent sample	Respondent sample with weights from step 2.	Households where all individuals asked income questions (age 15+) are linked to a PIK.
B. External population targets	State-level population estimates by race, Hispanic-origin, gender, and age	External population estimates	

Notes: This table describes the entropy balance reweighting procedure. In the first stage, respondent housing units are reweighted to control for selection into response. This is done by reweighting them to match the characteristics of the target population – all nonvacant housing units in sample. In the second stage, we estimate individual weights that preserve the distribution of housing-unit characteristics from the first stage, while also matching external population totals and approximating the spousal equivalence of weights that are a part of the existing CPS ASEC weights, as in Rothbaum and Bee (2022). To address selection into PIK assignment (and the availability of administrative data), we add a third-stage weighting adjustment.

Table A6: Imputation Summary Statistics: Survey Earnings

	W-2 Earnings	Respondents	Imputed Estimate Survey	Imputed Estimate SRMI	SRMI - Survey (Percent difference for dollar values)
Has Survey Earnings	= 0	0.181	0.282	0.230	-0.052***
					(0.007)
	!= 0	0.908	0.860	0.907	0.046***
					(0.005)
	q = 1	0.676	0.623	0.706	0.083***
					(0.014)
	q = 2	0.924	0.842	0.921	0.079***
					(0.009)
	q = 3	0.967	0.928	0.961	0.033***
					(0.008)
	q = 4	0.984	0.960	0.978	0.018***
					(0.006)
	q = 5	0.985	0.960	0.973	0.013**
					(0.006)
Average Wage and Salary Earnings	= 0	45,760	43,550	40,440	-0.071
(from main job)					(0.061)
	!= 0	55,520	52,470	53,330	0.016
					(0.047)
	q = 1	11,960	22,010	20,840	-0.053
					(0.084)
	q = 2	23,540	29,810	26,300	-0.118*
					(0.055)
	q = 3	37,750	43,950	37,910	-0.137**
					(0.045)
	q = 4	57,340	62,050	56,790	-0.085
					(0.058)
	q = 5	120,300	100,000	124,900	0.248***
					(0.061)
Median Wage and Salary Earnings	= 0	25,900	30,210	31,360	0.038
(from main job)					(0.092)
	!= 0	41,200	37,690	37,090	-0.016
					(0.047)
	q = 1	6,747	12,400	13,780	0.111
					(0.158)
	q = 2	20,720	24,660	22,160	-0.102
					(0.055)
	q = 3	35,630	36,250	33,570	-0.074
					(0.055)
	q = 4	55,350	51,490	52,060	0.011
					(0.045)
	q = 5	100,300	78,690	97,460	0.238**
					(0.073)

Notes: This table shows basic summary statistics of survey wage and salary earnings conditional on W-2 earnings (having a W-2 and by W-2 earnings quintile for q = 1,2,3,4,5). Each row shows the relevant survey wage and salary earnings statistic for survey earnings respondents, imputed as part of regular survey production and by SRMI, as discussed in Appendix D. Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A7: Imputation Summary Statistics: Means-Tested Benefits

	Administrative Data Available?		Difference	Diff in Diff
	Yes	No	No - Yes	(Adrec - Survey) and (No - Yes)
TANF				
Survey				
Receipt	1.03	1.05	0.02	0.17
	(0.08)	(0.08)	(0.11)	(0.20)
Amount	3,054	3,937	882**	-975**
	(205)	(331)	(391)	(471)
Administrative				
Receipt	0.78	0.97	0.19	
	(0.06)	(0.16)	(0.18)	
Amount	2,604	2,511	-93	
	(168)	(244)	(293)	
SNAP				
Survey				
Receipt	9.85	9.28	-0.57*	-0.42
	(0.32)	(0.22)	(0.38)	(0.51)
Amount	2,363	2,345	-18	73
	(70)	(51)	(87)	(120)
Administrative				
Receipt	16.11	15.12	-0.99*	
	(0.44)	(0.39)	(0.58)	
Amount	2,807	2,862	55	
	(60)	(80)	(100)	

Notes: This table shows basic summary statistics of means-tested benefits imputed for incomplete state-level administrative data. For both TANF and SNAP, the first rows show how survey responses vary across states with and without administrative records and the next set of rows show the administrative and imputed estimates. For each, we then compare the states without administrative data (No) to the states with (Yes) and take the difference in difference by comparing the administrative (No - Yes) to the survey (No - Yes). The means-tested benefit imputation is discussed in Appendix D. Standard errors in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent levels and are only shown for differences.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A8: Combining Survey and Administrative Earnings

A. By Reported Earnings Type and Source

Survey		Administrative		Rule		Percent of Sample	
Wage and Salary	Self Employment	Wage and Salary	Self Employment	Wage and Salary	Self Employment	All Adults	Any Earnings
X	X	X	X	Job-level administrative	1040 (from TMI)	0.4	0.6
	X	X	X	Job-level administrative	1040 (from TMI)	0.4	0.6
X		X	X	Job-level administrative	1040 (from TMI)	4.1	5.7
		X	X	Job-level administrative	1040 (from TMI)	0.4	0.5
X	X		X	None (administrative)	1040 (from TMI)	0.7	1.0
	X		X	None	1040 (from TMI)	1.5	2.1
X			X	None (administrative)	1040 (from TMI)	1.3	1.7
			X	None	1040 (from TMI)	1.2	1.7
X	X	X		Measurement error model	Survey	1.8	2.4
	X	X		Measurement error model		0.8	1.1
X		X		Measurement error model	None	50.5	70.1
		X		Job-level administrative	None	5.6	7.7
X	X			Survey	Survey	0.8	1.1
	X			None	Survey	1.0	1.4
X				Survey	None	1.6	2.3
				None	None	28.0	

B. By Combination Rule

	Percent of Sample	
Combination Rule	All Adults	Any Earnings
Simple - no earnings or only earnings in one source	38.6	14.7
Earnings Choice	53.0	73.6
Default to administrative data due to data issues (potential misclassification, missing self-employment, etc.)	8.4	11.7

Notes: This table describes the possible combinations of survey and administrative reports of wage and salary and self-employment earnings as well as our rules for when we use survey and administrative reports for each. If the administrative wage and salary earnings on the 1040 is positive but there are no reported job-level administrative earnings, then we use the 1040 value when the rule indicates use of the job-level data. "All adults" includes anyone 15 or over as they are asked survey earnings questions. The sample only includes individuals in the NEWS sample.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A9: Combining Administrative and Survey Earnings: Share with Survey Earnings by Mean Reversion Parameter Kappa

Kappa	Share Survey Earnings
0.7	5.8
	(1.1)
0.75	8.4
	(1.5)
0.8	11.8
	(2.0)
0.85	16.0
	(2.3)
0.9	20.6
(NEWS)	(2.7)
0.95	25.8
	(3.4)
1	30.9
	(3.8)

Notes: This table shows how variation in the mean-reversion kappa parameter in the measurement error model affect the share of individuals whose survey wage and salary earnings are used. Figure A5 shows how the household income distribution differs under these alternatives. Standard errors in parenthesis.
Source: 2019 Current Population Survey Annual Social and Economic Supplement linked to administrative, decennial census, and commercial data.

Table A10: Income Type by Source for Filers and Nonfilers

	Source		
Income Type	**Filers**	**Nonfilers**	**Notes**
Wage and Salary Earnings	W-2 DER LEHD 1040	W-2 DER LEHD	Administrative data may miss unreported ”under-the-table” earnings. Current W-2s and DER do not include pre-tax employee contributions to health insurance premiums. LEHD does not have complete coverage. Survey has potential for misreporting and underreporting.
Self-Employment Earnings	1040 DER	Survey only	Under-reported substantially on surveys and in administrative records. Considerable disagreement between extensive margin reporting on surveys and administrative data (Abraham et al., 2021).
Social Security	1040 PHUS	PHUS	
Supplemental Security	SSR	SSR	
Unemployment Insurance	1040	Survey only	Included in 1040 Total Money Income. Imputed for nonfilers using disclosed results from more detailed 1099-G data.
Worker's Compensation	Survey only	Survey only	Not available federal administrative data.
Public Assistance	TANF	TANF	Current data only covers some states. TANF data does not cover all possible cash assistance programs.
Veteran's Benefits	Survey only	Survey only	Potential for VA data use in the future
Disability, Survivor, and Retirement Income	1099-R	1099-R	
Interest	1040	Survey only	Imputed for nonfilers using disclosed results from more detailed 1099-INT data.
Dividends	1040	Survey only	Imputed for nonfilers using disclosed results from more detailed 1099-DIV data.
Rent and Royalty Income	1040	Survey only	Net rent and royalty income included in 1040 Total Money Income. Gross rent and royalty income available as a separate variable.
Educational Assistance	Survey only	Survey only	
Financial Assistance	Survey only	Survey only	
Alimony	1040	Survey only	Included in 1040 Total Money Income
Gambling Winnings	1040	Survey only	Included in 1040 Total Money Income. Potentially available on survey as ”other income.”

Notes: This table describes the available data sources for the various types of income, including notes about the limitations of various sources. The availability of income varies between filers and nonfilers, with more income sources available in the currently available administrative records for filers.

Appendices

A Data Linkage

A.1 Person Linkage[46]

The Census Bureau developed the Person Identification Validation System (PVS) to probabilistically match individuals' records in survey and other data to their SSN or Individual Taxpayer Identification Number (ITIN) using personally identifying information (PII), such as name, date of birth, and residential address (Wagner and Layne, 2014). Linked records are assigned a Protected Identification Key (PIK) and the PII and SSN or ITIN are removed. The PIK serves as the anonymized linkage key to match individuals across data sets.

As a result, if PVS is unable to assign a PIK to a given survey respondent, no administrative data are available for that respondent. Bollinger et al. (2019) found a linkage rate in their CPS ASEC sample (2006-2011) of 86 percent, which matches our estimate for the 2019 CPS ASEC. Because observable characteristics, such as race, ethnicity, citizenship status, etc., are correlated with PIK assignment (Bond et al., 2014), we must account for this selection into linkage in our estimates, which we discuss in Section C.

A.2 Address Linkage

Brummet (2014) describes the development and performance of the system used to link household records, via residential address fields, to the Master Address File (MAF), called the "MAF Match." Information such as house number (and suffix, such as apartment number), street name (and prefix/suffix, such as rural routes or state highway identifiers), city, state, ZIP code, etc. is used to link addresses in each data set to the MAF, to assign them MAFIDs.

As with PIKs, this means that if the MAF Match process is unable to assign a MAFID to an address, the information associated with that address in that data source cannot be linked to other address-level data. For recent years of surveys such as the ACS, CPS ASEC, and SIPP, every

[46] The discussion in this section follows Bee and Rothbaum (2019) closely.

housing unit has a MAFID because the sample was drawn directly from the MAF.

A.3 Job Linkage

The W-2, DER, and LEHD files all have information on individual jobs. However, unlike the LEHD, the W-2s and DER do not capture gross earnings. The Census Bureau receives W-2 extracts from the IRS that include Box 1 "Wages, tips, and other compensation," Box 3 "Social Security wages," and the sum of deferred compensation in Box 12 codes D-H.[47] We only observe taxable earnings and deferred compensation, but not other non-taxable earnings. We therefore do not have information on pre-tax employee payments for health insurance and other forms of pre-tax compensation not available in the extract provided by the IRS, such as contributions to Health Savings Accounts. In most of this section, we will primarily discuss W-2s and not the DER, as the two are identical for most workers for whom the DER is available.

Not all jobs are covered by unemployment insurance, and thus some jobs are out of universe for the LEHD. This includes all federal government employees and some private sector employees.[48]

In the earnings question on the CPS ASEC and ACS, respondents are asked to report "money income", which includes gross wage and salary earnings. To match this concept, we would like gross earnings for each individual job, which we could then use to estimate person-level gross earnings. However, we have gross earnings for only a subset of jobs (from the LEHD) and taxable earnings + deferred compensation from the universe of jobs (from W-2s). Because the LEHD includes a subset of jobs we should observe in W-2s, it is possible for an individual to have one job

[47]These codes include elective deferrals to plans under Box 12 codes D: 401(k), E: 403(b), F: 408(k)(6), G: 457(b), and H: 501(c)(18)(D). These boxes cover 96.3 percent of all elective retirement contributions on W-2s, calculated from IRS Statistics of Income Tax States for Individual Information Return Form W-2 Statistics, Table 7.A at `https://www.irs.gov/statistics/soi-tax-stats-individual-information-return-form-w2-statistics`, accessed 11/17/2021.

[48]For example, Maryland's Department of Labor lists the following jobs as exempt: barbers and beauticians, taxicab drivers, owner-operated tractor drivers in certain E and F classifications, maritime employment, election workers, church employees, clergy, certain governmental employees, railroad employment, newspaper delivery, insurance sales, real estate sales, messenger service, direct sellers, foreign employment, other state unemployment insurance programs, work-relief and work-training, family members, hospital patients, student nurses or interns, yacht salespersons who work for a licensed trader on solely a commission basis, services of aliens who are students, scholars, trainees, teachers, etc., who enter the U.S. solely to pursue a full course of study at certain vocational and other non-academic institutions, recreational sports officials, home workers, and casual labor. Refer to `https://www.dllr.state.md.us/employment/empfaq.shtml` accessed 11/1/2022.

in the LEHD and two in the W-2s. Therefore, we cannot just sum the earnings from both sources and take the maximum, because the one with the higher value (in this case, W-2 earnings from two jobs) may understate this individual's true gross earnings.

Therefore, we would like to combine the LEHD and W-2 records at the job level. For an individual with one LEHD job and two W-2 jobs, we would then observe gross earnings for one job and taxable earnings plus deferred compensation for the other. For the second job, we could impute gross earnings conditional on the other information observed about them (discussed in Appendix D) and then sum the job-level gross earnings to estimate their administrative gross earnings.

However, linking LEHD and W-2 jobs is not trivial. In the simplest case, a firm files a W-2 and reports the job to the UI office with the same EIN. We can link these "direct matches" by PIK and EIN. However, some firms do not file their W-2s and UI reports under the same EIN, and some firms use multiple EINs in one source but a single EIN in the other (i.e., a separate EIN for each state's employment in the LEHD but one EIN in the W-2s). Other firms use other identifiers, such as state EINs, when they report jobs to UI offices. Therefore, we cannot directly link many jobs between the LEHD and W-2 files using PIK/EIN combinations. Since nearly all jobs in both files include a PIK, we can create a set of possible matches that match on PIK but not EIN. We can then identify the W-2 EINs that correspond to a different EIN or state EIN in the LEHD by looking across all workers with unmatched jobs. We create a W-2 EIN to LEHD EIN crosswalk of these "indirect match" jobs.

An example of how we find direct and indirect matches is shown in Figure A1. In the example, we have three workers (PIK = 1, 2, 3) and their W-2 and LEHD jobs. For EIN = 400 and 500, the jobs match at the PIK-EIN level. However, EINs 100 and 600 in the W-2s and 200 in the LEHD do not match. Each worker with EIN = 100 in the W-2s also has a job with EIN = 200 in the LEHD and each of those jobs has similar earnings on the two files. We use this information to infer that W-2 EIN 100 is the same firm as LEHD EIN 200. We would then be left with the W-2 job at EIN = 600 that does not match to any job in the LEHD, perhaps representing a job that is not covered by unemployment insurance.

To create a crosswalk of all indirect matches between W-2 and LEHD EINs, we develop an iterative algorithm using three pieces of information:

1. The diference in earnings reported on the W-2 and LEHD for the possible job match,
2. The share of jobs in the W-2 EIN that match to the same LEHD EIN and the share of jobs from the LEHD EIN match to the same W-2 EIN, and
3. The number of likely matches between a W-2 EIN and an LEHD EIN

For the first rule, we can identify matches as likely if the W-2 and LEHD earnings are within some percent of each other. For the second, we can only keep matches in the crosswalk if many or most of the jobs in a W-2 or LEHD EIN are identified as likely matches to a single EIN on the other file. For the third, we may be more confident of a possible match if 100 jobs are all flagged as likely matches than if two are.

We create an iterative process to create our indirect matches where we set the thresholds for each of these three possible rules to identify likely matches. We identify the W-2 EIN-LEHD EIN combinations that match under these thresholds, add those combinations to our crosswalk and then remove the matched jobs from our possible match dataset. The removed jobs include all jobs with those pairs of EINs, not just the ones flagged as likely matches by our percent difference cutoff. We then repeat the process with the remaining jobs after adjusting the thresholds used to identify possible matches. The goal of the iterative process is to first add the matches we are sure of from the set of unmatched jobs (large firms, for example) before we match jobs from smaller firms or with larger differences in earnings across the files.

For example, in the first pass at identifying indirect matches, we flag jobs as likely matches if the W-2 and LEHD earnings are within 10 percent of each other. We then keep the W2 EIN-LEHD EIN combinations where 50 percent or more of them match in one direction or the other - i.e., 50 percent of jobs at a W-2 EIN match to the same LEHD EIN or 50 percent of jobs at the LEHD EIN match to the same W-2 EIN. Finally, we only keep EIN matches for the crosswalk if at least 5 jobs match.

In the example in Figure A1, there are three jobs at W-2 EIN = 100 and LEHD EIN = 200 that are within 10 percent of each other and flagged as likely matches. All jobs in W-2 EIN = 100 match to LEHD EIN = 200 (and vice versa). This combination meets the first two conditions. However, the number of matches is 3, which is less than the threshold of 5 so this combination of EINs would not be flagged as a match. These jobs would be kept in the set of unmatched jobs for the next round of the process.

In subsequent rounds, we can (1) increase the tolerance on likely matches (i.e., from 10 to 20 percent difference in earnings), (2) reduce the share matched needed within W-2 or LEHD EINs (i.e., from 50 percent to 25 percent), or (3) lower the threshold of likely matches needed to confirm a match (i.e., from 5 to 3). From Figure A1, if we lowered the number of likely matches to 3, then we would count W-2 EIN = 100, LEHD EIN = 200 as an indirect match, add that match to our crosswalk, and remove the matches under Indirect Matches from the set of unmatched jobs.[49]

Finally, we implement a series of additional steps to match the remaining set of jobs. First, we try to find jobs that have multiple EINs in the LEHD but one EIN in the W-2s, for example if a firm changed EIN mid-year for any reason (restructuring, acquisition, etc.). In that case, the LEHD might have multiple EINs during the year as the firm filed its quarterly reports, but only one EIN for the workers' W-2s. We then flag remaining unmatched jobs as ad hoc likely matches if their earnings are within a certain percent of each other, but they were not matched by the iterative process.

In Table A2, we show summary statistics from the linkage process. In the W-2s, there are 257 million unique jobs in 2018, with 238 million in the LEHD. Of those, 216 million are direct matches by PIK-EIN combination. This leaves 41 million unmatched W-2 jobs and 22 million unmatched LEHD jobs. However, we find an additional 15 million indirect matches through our matching algorithm, covering 70 percent of the unmatched LEHD jobs and 37 percent of the unmatched W-2 jobs. We then have 82 percent of jobs matched directly by PIK-EIN, 6 percent matched indirectly, 10 percent unmatched from W-2s, and 3 percent unmatched from the LEHD. We use this linked

[49] In practice, we first increase the earnings percent difference threshold for likely matches from 10 percent to 20 percent to 25 percent. We also decrease the share of matches within an EIN that must match from 50 percent to 25 percent to 10. Finally, we also decrease the minimum number of matches from 5 to 2 to 1. We make each of these changes separately from the initial thresholds and then change them simultaneously.

job information to better estimate gross earnings at the job and person level for use in our income estimates.

Since LEHD earnings should exceed W-2 taxable earnings + deferred compensation in large part due to employee pre-tax payments for health insurance premiums, we compare them in our CPS ASEC sample for individuals who reported whether they have private health insurance coverage.[50] As shown in Table A1, individuals with private coverage are less likely to have LEHD earnings that are approximately the same as their W-2 earnings + deferred compensation (LEHD $\geq$ W-2 by 0-1 percent), and covered individuals are 3 to 5 times more likely to have LEHD values that exceed the W-2 amounts by 1-3 percent, 3-5 percent, 5-10 percent, and 10+ percent. This likely reflects the missing gross earnings for employee pre-tax contributions to health insurance premiums on W-2s.

However, Table A1 also shows that there is a substantial number of jobs whose W-2 taxable earnings + deferred compensation exceeds LEHD gross earnings. At present, we treat these jobs as having measurement issues in the LEHD and default to the taxable earnings + deferred compensation from the W-2 and impute gross earnings for those jobs as discussed in Appendix D. We plan to investigate this issue further in future NEWS releases.

A.4 Firm Linkage

Our firm identifier in the employment data is the EIN. However, as we noted when crosswalking the job-level data between the W-2 and LEHD, an EIN does not necessarily correspond to a firm. Some firms have multiple EINs, for example in each state of operation, which can make matching individual workers to their firm (rather than subunits of the firm) difficult.

This is a challenge for all users of EIN-based administrative data (Joint Committee on Taxation, 2022; Chow et al., 2021). Chow et al. (2021) redesigned the Longitudinal Business Database (LBD) in part to help bridge this gap and to make linkages between various worker- and firm-level datasets easier. We use this redesigned LBD to map EINs to LBD firm identifiers (LBDFID). In the LBD,

[50] Note that the CPS ASEC variable we use indicates receipt of private coverage, but not necessarily that the individual's job (rather than a spouse, partner, or other family member) was the source of the coverage.

each establishment is associated with one or more EINs and also to a LBDFID. We create a crosswalk of all EIN to LBDFID combinations by year. If a firm restructures during a given year, it is possible for the same EIN to map to different LBDFIDs in the same year. When that happens, we assign the EIN to the associated LBDFID in the subsequent year. From that, we create a year-by-year EIN-LBDFID crosswalk for all firms in our data. We can then merge the job-level data by EIN to an LBDFID to match each worker to a firm. At the firm level (by LBDFID), we can then use LBD data or create our own summary statistics on firm employment and payroll from the linked job-level data. At present, we use this firm information for modeling, imputation, and weighting.

B File Construction

B.1 Address File

The first file we create from the data in Sections 3.1-3.6 is the Address File. We link the sample of occupied (non-vacant) housing units in the survey to the aforementioned sources of administrative, survey, census, and commercial data, as shown in Figure 4. By starting with addresses, we have information from all occupied units, including respondents *and* nonrespondents. In the address file, we do not use any information from survey responses other than whether the unit responded. This file is used to construct the weights that address selection into our sample, discussed in Section C.

First, we link the MAFIDs of occupied housing units to the MAF and Black Knight data to get information on the housing units, such as home value and type (single vs. multi-unit). We then link the same MAFIDs to several files that have both MAFIDs and PIKs, including the IRMF, MAF-ARF, and 1040 tax returns, giving us information on the information returns (W-2, 1099-G, etc.) sent to that address, their income (from tax returns), and PIKs for individuals who are associated with that address. We create a roster of PIKs for the linked individuals in each occupied unit. We then link this roster to various files, including the universe PHUS and SSR files, the Numident, W-

2s, LEHD, and the IRMF and 1040 tax returns.[51] We then link the LEHD and W-2 jobs together using the job crosswalk discussed in Section A.3. We also link those jobs to the characteristics of the employer firm in the LBD using the EIN-firm ID crosswalk discussed in Appendix A.4.

Finally, we create geographic summary files at different levels of aggregation (state, county, and tract) that summarize the characteristics of residents of those locations from different files. These include (1) a summary of demographic characteristics from the 2010 decennial census, (2) demographic and socioeconomic characteristics from 5-year ACS files, (3) earnings and information return receipt from the IRMF and W-2 files, (4) citizenship information from the MAF-ARF linked to the Numident, and (5) income and marital status information from 1040 tax returns.

This gives us information on the income, earnings, industry, race, Hispanic origin, marital status, presence of children, home value, housing unit type, etc., as well as information about the neighborhoods in which each household lives. However, data coverage is not perfect. As shown in Table A3, we can link 93 percent of occupied CPS ASEC addresses to at least one data set (excluding the MAF, from which the addresses were sampled). That leaves 7 percent of addresses that we cannot link to any data other than the MAF. For these, we have no additional address-level information, and we cannot link the address to possible residents, which means that we cannot observe any address-level demographic or socioeconomic characteristics for these households (apart from the survey responses). For them, we only have information about their communities from the geographic summary files and about their housing unit from the MAF. Furthermore, we do not directly observe some characteristics that may be related to wellbeing and survey response, such as educational attainment, health insurance status, disability status (except if receiving SSI or OASDI), etc.[52]

[51] For the IRMF and tax return link, we do this in case an individual associated with the address received an information return at a different address or was on a 1040 tax return filed from a different address.

[52] Rothbaum and Bee (2022) evaluate how well weighting can control for differences between respondents and nonrespondents by one of the dimensions unobserved in our linked data, educational attainment, by linking the subset of housing units to prior ACS responses. They find that most, but not all, of the selection into response by educational attainment is addressed by weights created using similar linked data.

B.2 Person File

The second file we create from the data in Sections 3.1-3.6 is the Person File. We create this file by linking survey respondents to administrative data, as shown in Figure 5. In combination with the weights created using the Address File, the Person File is used to create our income and poverty estimates.

The Person File contains survey responses, including demographics, socioeconomic characteristics, income, etc. as well as administrative information on income on the following files: 1040s, W-2s, DER, LEHD, 1099-Rs, PHUS, SSR, and TANF. Table A10 shows the data sources with information by income type (wage and salary earnings, Social Security, etc.) for tax filers and nonfilers. For tax filers, most income types are available in the administrative data, either as separate variables or as part of 1040 Total Money Income. For nonfilers, we observe wages and salary earnings (W-2s, DER, and LEHD), OASDI benefits (PHUS), SSI (SSR), retirement, survivor and disability income (1099-R), and TANF income (state data), as well as flags for the potential presence (but not amount) of interest income (1099-INT), dividends (1099-DIV), and unemployment compensation (1099-G). Several types of income are only available on the survey, regardless of tax filing status, including workers' compensation, veterans benefits, educational assistance, and inter-household financial assistance. Table A4 shows the share of the sample that can be assigned a PIK and the share of individuals with a PIK that can linked to each of the administrative data sources.

C Weighting

Weighting is one method for addressing missing data, where variables are completely unobserved for a subset of the sample.[53] Let R be an indicator for whether the information is available for an individual or unit (i.e., response to a survey). Given a set of k variables $X = \{x_1, x_2, \ldots, x_k\}$ for n units (individuals, households, firms). These covariates are observed for some units, but not others, $X = \{X_O, X_M\}$, where O indicates observed ($R = 1$) and M indicates missingness ($R = 0$).

There are several possible relationships between missing data and the individual and household

[53] The discussion in this section follows Rothbaum and Bee (2022) closely.

characteristics we are interested in estimating. The simplest possible pattern of missingness (for the analyst) is if the data are missing completely at random (MCAR). In this case, nonresponse is completely random and not related to X_O or X_U, or $R \perp (X_O, X_M)$. For example, if a unit flips a coin when deciding whether to respond to the survey, nonresponse would be MCAR. If the data are MCAR, then the solution is easy – we do not need any adjustment to the data to get an unbiased estimated. We can just drop missing observations. Only precision is affected by MCAR data, as the sample is smaller than if all individuals were observed.

Another possibility is that the data are missing at random (MAR), conditional on the observable information. Given a distribution $f(\cdot)$, data are MAR if $f(R|X) = f(R|X_O)$, which means that missingness is conditionally independent of the unobserved information (X_U). This is the underlying assumption of most nonresponse bias adjustments, such as survey weights.

However, another possibility is that the data are not missing at random (NMAR), where $f(R|X) \neq f(R|X_O)$. This is much more challenging to address. Suppose the probability of information availability varies with income, which is in X. Then $f(R|X) \neq f(R|X_O)$, and we cannot easily recover the true underlying income distribution from the observed data in X_O without strong, generally difficult to verify assumptions about $f(R|X)$.

However, MAR is an independence assumption conditional on X. Suppose there is another set of variables A that are observed for the full sample, independent of response. In that case it is possible that the data are NMAR with respect to X, but MAR with respect to A, or more formally $f(R|X) \neq f(R|X_O)$ but $f(R|X, A) = f(R|X_O, A)$. Rothbaum and Bee (2022) found that from 2020 to 2022, nonresponse in the CPS ASEC was NMAR with respect to X and that income statistics were biased by 2-3 percent as a result. They used additional information from administrative data linked at the address level to the addresses of respondent *and* nonrespondent households to adjust the weights for nonresponse.[54]

There are several aspects of our data that lend themselves to weighting to address missing information — where a subset of variables is completely missing for some units. For survey nonresponse, none of the survey information is observable for the nonresponding units. For incomplete linkage,

[54]Rothbaum et al. (2021) did the same to address nonresponse bias in the 2020 ACS.

none of the administrative data is available for the unlinkable individuals. If survey nonresponse or linkage are MAR, we can address the bias through weighting.

To include additional characteristics in the weighting model, we use entropy balancing (Hainmueller, 2012). Entropy balancing is an application of exponential empirical calibration. Empirical calibration has a long history of use in survey weighting (Deming and Stephan, 1940; Deville and Särndal, 1992) – the existing weighting models (using raking) in the ACS and CPS ASEC are applications of empirical calibration.[55]

We use the unobservable information (in the survey) from the linked administrative and decennial census data, which are available for all linkable households regardless of whether they responded as well as the geographic summary information. Entropy balancing estimates weights that match a specified set of moment constraints (i.e., to adjust the weights according to $f(R|X_O, A)$) while keeping the final weights as close as possible to the initial weights.

Entropy balancing has several appealing features for this application. The first is flexibility. Inverse probability weighting (or any simple regression-based reweighting technique) is only amenable to matching characteristics of the distribution in the sample, but not external targets. Empirical calibration will adjust the weights to match any properly specified target moment, whether that moment was estimated on the sample or with external data. The second is statistical efficiency, which is achieved by keeping the final weights as close as possible to the initial probabilities of selection.[56] Third, entropy balancing directly adjusts the weights to the moment conditions, like with raking but unlike single-index propensity score weighting approaches (such as inverse probability weights). In propensity score approaches, the adjustment is made to the single index generally estimated from a regression. The resulting balance must be assessed to evaluate the success and quality of the propensity score model. In some cases, a misspecified propensity score model can make balance worse on a given set of dimensions. As entropy balancing directly targets those moments, balance is assured. Fourth, unlike raking, or cell-based empirical calibration methods,

[55]Raking, also called iterative proportional fitting, adjusts the weights for each group to match the population total for that group. It is solved by iterating across groups to match the different population targets in stages.

[56]Through the minimization in equation C.1.

entropy balancing allows for the inclusion of continuous variables in the weighting model.

The fifth is computational efficiency – entropy balancing allows matching to a high-dimensional vector of moment constraints. In terms of our MAR assumption, if A or X is high dimensional, then the computational efficiency makes it feasible to include all of A and X in the weighting model. As in Rothbaum and Bee (2022), we use state-level population controls that include estimates of the share of the population in 20 separate groups in each of the 50 states and the District of Columbia. That yields 1,020 separate target population moments before even considering information from the linked administrative data. The computational efficiency of the entropy balancing optimization algorithm allows us to match to both the linked administrative and population control targets simultaneously. This eliminates the need for an additional population control raking step that can undo the balance from the nonresponse adjustment.[57]

Next we discuss entropy balancing in detail. Suppose we have n observations, where $i = 1, 2, \ldots, n$ with base weights based on sampling probabilities of $q = \{q_1, q_2, \ldots, q_n\}$. Entropy balancing estimates weights $w = \{w_1, w_2, \ldots, w_n\}$ that solve the following minimization problem:

$$\min_{w} \sum_{i=1}^{n} w_i \log(\frac{w_i}{q_i}) \tag{C.1}$$

subject to several sets of constraints. First, we have p moment conditions. Let $X = \{X_1, \ldots, X_p\}$ be a matrix of observable characteristics. For characteristic j, the moment conditions are defined

[57]Several studies have implemented first-stage nonresponse adjustments followed by second-stage raking to population controls that do not condition on the first-stage adjustment. Slud and Bailey (2010) found that for some metrics of weight quality, the benefits of the first-stage adjustment disappeared after the application of the second-stage raking to population controls. Eggleston and Westra (2020) found that for some measures used in the first-stage adjustment, the bias is not improved or can be greater using the final weights after raking to population controls, although most statistics show reduced bias after the second-stage raking. Rothbaum et al. (2021) found something similar in follow-up work on the ACS when applied to the 5-year release. Without including very detailed population controls in the 2020 1-year ACS weights (down to tract-level population), when the 2016-2020 files were combined and raked to the 5-year population controls, the 2020 nonresponse adjustment had little impact on the 5-year estimates. Only when the 2020 file was simultaneously reweighted to detailed population controls and the linked administrative targets, limiting the need for additional raking adjustments, did the nonresponse bias adjustment persist on the final 5-year file.

to match a vector of pre-specified constants $\bar{c}_j$, where:

$$\sum_{i=1}^{n} w_i c_j(X_{i,j}) = \bar{c}_j. \tag{C.2}$$

$c_j(\cdot)$ can be any arbitrary function.

Second, we have constraints on the weights themselves:

$$\begin{aligned} \sum_{i=1}^{n} w_i &= \bar{w} \\ w_i \geq 0, i &= 1, \ldots, n \end{aligned} \tag{C.3}$$

which ensure that the weights sum to some pre-specified total weight $\bar{w}$, which can be the population count or 1. The value of $\bar{w}$ does not affect the relative weights of each observation.

As such the weights can be adjusted to match pre-specified moments such as population means, variances, higher-order moments, moments of any transformed distribution of $X_(i,j)$, etc. In summary, entropy balancing adjusts the weights according to (C.1), subject to the constraints in (C.2) and (C.3).[58]

Entropy balancing was developed as an application of empirical calibration to balance treatment and control groups when estimating causal treatment effects in observational studies. Zhao and Percival (2017) show that, in that context, entropy balancing is equivalent to estimating a logistic model for the propensity score and a linear regression model for the outcome, conditional on the covariates used in the moment conditions. They find that entropy balancing is doubly robust - if at least one of the two models is correctly specified, the estimated population average treatment effect on the treated (PATT) is consistent.[59] Using the notation of that literature, let γ be the PATT, Y be an outcome of interest where $Y(1)$ is the outcome if treated and $Y(0)$ is the outcome if untreated, then:

[58]In practice, as is not necessarily possible to satisfy all constraints simultaneously through weighting adjustment, the analyst sets a tolerance level for the moment constraints. The weighting algorithm adjusts the weights iteratively until all constraints are satisfied subject to the specified tolerance.

[59]Double robustness is not a panacea. Kang and Schafer (2007) show via simulation that doubly robust models for missingness can perform poorly when neither model is correctly specified, or as they write, "in at least some settings, two wrong models are not better than one."

$$\gamma = E[Y(1)|T=1] - E[Y(0)|T=1]. \tag{C.4}$$

In the causal inference literature, the challenge is that $E[Y(0)|T = 1]$ is not observed. Under entropy balancing, given $\sum_{i=1}^{n} q_i = \bar{q}$, the PATT is estimated as:

$$\hat{\gamma}_{ebw} = \frac{1}{\bar{q}} \sum_{T_i=1} q_i Y_i - \frac{1}{\bar{w}} \sum_{T_i=0} w_i Y_i. \tag{C.5}$$

In the case of survey weights, the "treatment" is nonresponse, and the double robustness result applies. Entropy balancing reweights the sample so that the estimate of Y for the weighted respondents is equal to the estimate of Y for the population,[60] or:

$$E[Y] = \frac{1}{\bar{w}} \sum_{i=1}^{n} w_i Y. \tag{C.6}$$

We would like to reweight the respondent sample so that its distribution of characteristics matches the target population from which the sample was drawn. However, some characteristics are not observable for all housing units with the available linked census, survey, and administrative data. For example, we do not observe any demographic information for housing units that are not linked to an information return in the IRMF file, as the IRMF provides the identifier needed (PIK) to link individuals to all other data sources. Therefore, we use a second source of data for our reweighting – the aforementioned external estimates of population by geography. For both the linked data and the external population estimates, we can specify a set of moment conditions, which are intended to capture the distribution of characteristics in the target population. In the language of our MAR assumption, we are concerned that $f(R|A) \neq f(R|X)$ and that we need X_O (the demographic information) in the weighting model as well, such that $f(R|A, X_O) = f(R|X)$.

Our data have one additional complication – the target moments are at separate levels of aggregation. Estimates from the linked administrative and census data are at the housing unit level

[60] Conditional on strong ignorability ($Y(0), Y(1) \perp T|X$) and overlap ($0 < P(T = 1|X) < 1$), from Rosenbaum and Rubin (1983), as well as the proper specification of the moment conditions required for the Zhao and Percival (2017) double robustness result.

whereas the external state-level population moments are at the individual level. Entropy balancing is not amenable to matching moments at different levels of aggregation. Therefore, we proceed with a multi-stage reweighting procedure, which we discuss below and summarize in Table A5. This is analogous to two-step calibration, as discussed in Estevao and Säarndal (2006).

In the first stage, we adjust the household base weights for nonresponse, controlling to moments estimated from the linked administrative and census data. The target distribution is estimated using the nonvacant housing units in the March Basic CPS Sample, which includes both respondent and nonrespondent housing units. Given the known probability of inclusion in the sample (from the base weights), these are estimates of the underlying population moments for each of the included characteristics. The moments include housing-unit-level summary statistics on race, Hispanic origin, age, marital status, income, sources of income (through information return dummies), citizenship, and nativity.

Entropy balancing adjusts the housing unit weights so that the weighted estimates from respondent units match the moments estimated from all nonvacant households. Let us designate the housing-unit moment constraint variables as $X^L_{i,j}$, where L indicates linked data. Let w^1_i be the output weights of the first-stage reweighting. Given n respondent households, and a set of nonvacant (occupied) households NV, where $i = 1, 2, \ldots, n_{NV}$ with survey base weights q_i, the moment conditions are of the form:

$$\sum_{i=1}^{n} w^1_i c_j(X^L_{i,j}) = \sum_{i=1}^{n_{NV}} q_1 c_j(X^L_{i,j}). \tag{C.7}$$

With these moment conditions, we estimate w^1_i for each household using entropy balancing.

In the second stage, we would like to create weights (denoted $w^2_{m,i}$) for each individual m and household i, where $m = 1, 2, \ldots, M$, that adjust to external population controls while maintaining the household weighting adjustment from the first stage. We do so by simultaneously matching to three sets of target moments (2A-C in in Table A5):

A Preserve the distribution of housing unit characteristics

B Spousal equivalence

C External population targets

In the first set of constraints (A), we calculate person-weighted moments from the stage-1 weights. Given the number of people in household i, n_i^{HH}, we define the moment conditions using the stage-1 weights as follows:

$$\sum_{m=1}^{M} w_{m,i}^{2} \frac{1}{n_i^{HH}} c_j(X_{i,j}^{L}) = \sum_{i=1}^{n} w_i^{1} c_j(X_{i,j}^{L}). \tag{C.8}$$

This ensures that if we take the average weight of household members in household i (HH_i) as $\bar{w}_i^2 = 1/n_i^{HH} \sum_{p \in HH_i} w_{m,i}^2$, the following condition will be satisfied:

$$\sum_{i=1}^{n} \bar{w}_i^{2} c_j(X_{i,j}^{L}) = \sum_{i=1}^{n} w_i^{1} c_j(X_{i,j}^{L}). \tag{C.9}$$

This does not require that $\bar{w}_i^2$ is equal to w_i^1 for any household i, but rather that the specified constraints from stage one hold in the final entropy-balance weights, when the final weights are averaged across all household members. This procedure of dividing the household moments equally among the family members helps ensure that each person contributes to satisfying the moments from the linked administrative and decennial census data, which should reduce the variability of weights among household members. It is particularly important for person-level statistics, such as poverty or health insurance status, that are functions of household or family characteristics. For example, poverty status (poor/non-poor) is defined at an aggregated level (the family), but the share in poverty is estimated from individual weights. By having each household member be part of the moment conditions for the linked data, administrative income affects each member's weight, which affects the poverty estimate.

For the second set of moments in the second-stage reweighting (2.B. in Table A5), we approximate the spousal equalization that is part of existing CPS ASEC weights. We include this set of conditions because household- and family-level statistics should also be invariant to which spouse's weight is used as the family or household weight. Let $S = \{0, 1, 2\}$, where $S = 0$ if an individual is unmarried, 1 if the individual is the first spouse or cohabiting partner on the file, and 2 if the individual is the second spouse or partner on the file. Given an indicator function $I(\cdot)$, the spousal equivalence moment condition for a given characteristic in the linked data is:

$$\sum_{i=m}^{M} \left[I(S=1) w_{i,m}^{2} c_j(X_{i,j}^{L}) - I(S=2) w_{i,m}^{2} c_j(X_{i,j}^{L}) \right] = 0. \tag{C.10}$$

This does not require that each individual's weight be equal to their partner's, as that would require a separate moment condition for each couple. Instead, it requires that the characteristics of the households of partners in the linked data be balanced.

The third set of moment conditions (2.C. in Table A5) reweight the individual observations to match the age by race/Hispanic-origin/gender cells for each state and the District of Columbia, as noted above. These conditions have the simple form of equation (C.2).

With these three sets of conditions, we reweight the March Basic CPS sample to simultaneously match the household-level linked administrative data and the individual-level state population targets. For each individual, the initial weights for the stage 2 reweighting are the household weights from the stage 1 reweighting (w_i^1), so that the minimization from (C.1) becomes:

$$\min_{w^2} \sum_{i=1}^{n} w_i^2 \log\left(\frac{w_i^2}{w_i^1}\right). \tag{C.11}$$

However, for the full CPS ASEC sample, there is one more complication. The full sample includes groups that were oversampled based on characteristics reported in earlier survey responses, including Hispanic origin and the presence of children. Therefore, in the full sample, the weights for these oversampled individuals and households need to be adjusted to reflect their prevalence in the population and characteristics. To do this, we add a fourth set of moment conditions (2.D. in Table A5). We create these conditions from the entropy-balance weighted March Basic sample, because it is a stratified random sample that is not affected by oversampling based on observable characteristics from prior survey responses. Let $w_{i,m}^{2,March}$ be the second-stage weights from the March Basic Sample, $w_{i,m}^{2,Full}$ be the second-stage weights from the full CPS ASEC sample, and M_{Full} and M_{March} be the number of individuals in the full and March Basic CPS samples. This fourth set of conditions has the form:

$$\sum_{m=1}^{m_{Full}} w_{i,m}^{2,Full} c_j(X_{i,k}) = \sum_{m=1}^{m_{March}} w_{i,m}^{2,March} c_j(X_{i,k}). \tag{C.12}$$

This fourth set of moments includes information on race, Hispanic origin, income (from the linked administrative data), and the number of adults and children in the household. Without this set of conditions, estimates of the number of households by type (especially for oversampled groups) differ between the full and March Basic CPS ASEC samples. Additionally, without these constraints, observables-based oversampling in the full CPS ASEC biases estimates for oversampled subgroups relative to estimates from the March Basic sample. Although we focus on the estimates from the full CPS ASEC sample in this paper, we present the results from the Basic March sample in the Appendix as well, because it is a stratified random sample with no oversampling based on observable characteristics from earlier survey responses.

At this point, the weights would adjust for selection into response. However, because we are using administrative data to address survey misreporting, inclusion in our sample is also conditional on linkage to a PIK as that is the key to linking each individual to *every* source of administrative data. We therefore include in our sample only those households in which all those old enough to receive survey income questions (15+) are assigned a PIK. To address this selection, we add a third stage to the entropy balancing weight procedure used in Rothbaum and Bee (2022), as shown in Table A5, Stage 3.

Stages 3A and 3B have the same form as 2A and 2C, but add additional moments to the already specified ones from the linked data and external population controls. In adjusting for selection into linkage, we include moments on survey-reported income, administrative income, and survey poverty status by survey reported demographics such as race, Hispanic-origin, citizenship, and age.

The weights after this third-stage adjustment should adjust the sample for both selection into survey response and selection into linkage, to the extent possible given the observable survey and linked administrative data.

For valid inference, we repeat the above two-stage reweighting procedure 160 additional times using the baseline successive difference replicate factors created during the sampling process, which are available for all households regardless of response status. These replicate factors account for the sampling design of the monthly Basic CPS and CPS ASEC. Also, the first-stage target moments from the March Basic CPS sample are estimates and thus subject to sampling error. By repeating the procedure with the base weights and replicate factors, the target moments for each replicate will vary, and variation in the final weights across the replicates will reflect the uncertainty in our linked data estimates. All standard errors reported using EBW are calculated with these 160 replicate-factor EBW.[61]

As noted in Rothbaum et al. (2021), in addition to changing point estimates, improved weights can also affect standard errors. It is generally understood that increased variability among the survey weights can increase the standard errors, so weighting adjustments aimed at reducing bias are often done at the expense of increasing variance. However, Little and Vartivarian (2005) show that this may not hold true if variables used to adjust for nonresponse are correlated with survey variables of interest, a property they call "super-efficiency." This also has implications for how weighting models should be constructed, as including variables that are not strongly predictive of response, but are correlated with outcomes of interest can reduce variance of an estimate even if they do not affect its bias.

The full reweighting procedure is described in Table A5 . Stage 1 adjusts for nonresponse at the housing unit level by reweighting respondent households to match the characteristics of occupied households estimated from the linked administrative, decennial, and commercial data. Stage 2 creates individual weights that maintain the adjustment from Stage 1, but additionally adjust the person weights to match the external population controls. As in Rothbaum and Bee (2022), the Stage-2 weights adjust the sample for selection into survey response.

However, because we are using administrative data to address survey misreporting, inclusion in our sample is also conditional on linkage to a PIK, as that is the key to linking each individual to *every* source of administrative data. Our final sample includes only those households where all

[61]Refer to "Estimating ASEC Variances with Replicate Weights" (U.S. Census Bureau, 2009) for a discussion of successive difference replication in the CPS ASEC. Note also that at present we do not include uncertainty in the external population targets, but we hope to explore how best to account for that uncertainty in the weights as well in future research.

those old enough to receive survey income questions (15+) are assigned a PIK. To address this selection, we add a third stage to the entropy balancing weighting procedure used in Rothbaum and Bee (2022), as shown in Table A5, Stage 3. The Stage-3 weights maintain the adjustments of the Stage-2 weights, but also control for selection into linkage, to the extent possible given the observable survey and linked administrative data.

For valid inference, we repeat the above two-stage reweighting procedure 160 additional times using the baseline successive difference replicate factors created during the sampling process, which are available for all households regardless of response status. These replicate factors account for the sampling design of the monthly Basic CPS and CPS ASEC. Also, the first-stage target moments from the March Basic CPS sample are estimates and thus subject to sampling error. By repeating the procedure with the base weights and replicate factors, the target moments for each replicate will vary and variation in the final weights across the replicates will reflect the uncertainty in our linked data estimates. All standard errors reported using EBW are calculated with these 160 replicate-factor EBW.

As noted in Rothbaum et al. (2021), in addition to changing point estimates, improved weights can also affect standard errors. It is generally understood that increased variability among the survey weights can increase the standard errors, so weighting adjustments aimed at reducing bias are often done at the expense of increasing variance. However, Little and Vartivarian (2005) showed that this may not hold if variables used to adjust for nonresponse are correlated with survey variables of interest, a property they call "super-efficiency." This also has implications for how weighting models should be constructed, as including variables that are not strongly predictive of response, but are correlated with outcomes of interest, can reduce variance of an estimate even if they do not affect its bias.

Figure A8 shows the bias in estimates of address-linked characteristics using the various weights. In each panel, we compare the five separate weights to the target moments estimated on the set of all occupied housing units. They are:

1. Respondents — the weights only adjust for the probability the housing unit is selected into

the sample

2. Survey — the final survey weights

3. HH EBW — the Stage 1 weights that adjust for response at the household level only

4. EBW — the Stage 2 weights that adjust for response at the household level and to the external population controls

5. EBW + PIKed — the Stage 3 weights that adjust for response at the household level, to external population control, and for selection into linkage.

From Figure A8, we can see that OASDI recipients (linked to the PHUS) are overrepresented with the respondent and survey weights (Panel A), as are housing units with residents that are 65 and over (Panel B). The EBW bias estimates in Panels A and B (those that can be directly targeted in the weighting) are all very close to zero, with few statistically significant differences.[62]

Figure A9 compares statistics estimated on survey responses using the survey weights to those estimated using the Stage 2 (EBW) and Stage 3 (EBW + PIKed) weights. In this case, the survey-weighted and EBW estimates by race, Hispanic origin, and age should match the survey estimates by construction (as they are each weighting to external population controls). However, differences for other statistics for the EBW relative to the survey-weighted estimates reflect potential bias in the survey estimates, which we see, for example, for household income.

D Imputation

Suppose we have two variables Y_i and Y_j with missing values indicated by $R_i = 0$ or $R_j = 0$.[63] Missingness is monotone if $R_j = 0$ in all cases where $R_i = 0$. The pattern of missingness discussed above for weighting is one case of monotone missingness.[64] Missingness is non-monotone if $R_i = 0$

[62] Percentiles cannot be directly matched by entropy balancing. Instead, the weighting model weights respondents to match the share of units in different income bins (i.e., the share of households with address-level W-2 earnings $\leq$ \$25,000.

[63] The discussion in this section follows Hokayem, Raghunathan and Rothbaum (2022) and Fox et al. (2022) closely.

[64] In that case, we are assuming that for all variables in X, $R_i = R$, where $i = 1, \ldots, k$.

does not imply that $R_j = 0$.

While weighting can address missing data for the monotone missingness discussed in the prior section, it is not optimal as a general missing data correction when missingness is non-monotone. For non-monotone missingness, imputation is a better approach as it fully utilizes the available information (Raghunathan et al., 2001). In this section, we discuss imputations models generally followed by our implementation.

Suppose O is a collection of observable variables with no missing values, with $O = (O_1, O_2, \ldots, O_q)$ and $Y_1, Y_2, \ldots, Y_p$ are variables with missing values, with $Y = (Y_1, Y_2, \ldots, Y_p)$. Further, let U be a set of unobserved characteristics. Let $f(Y|O, U, \theta)$ be the conditional joint density, with $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and where θ_j is a vector of parameters in the conditional distribution for Y_j such as regression coefficients and dispersion parameters. An imputation model imposes some assumptions on f and θ to assign plausible values to Y where data are missing.

In this case, Y is MAR if missingness can be accounted for by observable characteristics, which can be written as $f(Y|O, \theta) = f(Y|O, U, \theta)$ (Rubin, 1976).[65] Another way to view imputation is through the lens of a researcher or data user. Consider a statistic Q, which could be a distributional statistic (such as a mean or median), a regression coefficient, or any other statistic or parameter of interest to the researcher. An imputation model is congenial or proper and results in unbiased estimates of Q if $E(\hat{Q}|O, \theta) = E(\hat{Q}|O, U, \theta) = Q$ and has valid confidence intervals for $\hat{Q}$ (Meng, 1994; Rubin, 1996).

This is only true when the imputation model is congenial and proper for the analysis being conducted. There are many examples in the literature where this congeniality condition fails for a given statistic or set of statistics. An example is match bias in the CPS. Bollinger and Hirsch (2006) showed that because the imputation model in the CPS does not include union status, estimates of the relationship between union status and earnings are attenuated in the imputed data. Even in this case, the issue is not that their earnings are misclassified (as very rarely will imputed earnings match the true value for a given individual), but that they are drawn from the wrong distribution – one that does not condition on union status. However, uncongeniality for one statistic does

[65]It is NMAR if $f(Y|O, \theta) \neq f(Y|O, U, \theta)$.

not indicate bias for other related statistics. For example, match bias on union status does not necessarily mean that the CPS imputation model will bias statistics of the unconditional earnings distribution.

It is impossible for congeniality to hold for all possible statistics Q, unless the model perfectly predicts the missing values, i.e., there is no misclassification.[66] However, we could assess the quality of an imputation model by comparing a set of the resulting $\hat{Q}$ estimates against known Q values. Fox et al. (2022) took this approach, using a variety of statistics, including regression coefficients and conditional and unconditional distributional statistics to evaluate their imputation model.

Hokayem, Raghunathan and Rothbaum (2022) addressed survey nonresponse in the CPS ASEC in 2009-2013 by including more covariates in the imputation model than the current CPS ASEC hot deck approach and comparing models with and without administrative data on earnings and income in the model. They find further evidence of match bias. However, with sufficient information in the model, they do not find evidence of nonignorable nonresponse (NMAR) when they compare the estimates of imputes that condition on administrative income to those that do not.

This non-monotone missingness is present in several variables in our data. Income items are particularly prone to survey nonresponse - over 40 percent of earnings (and all income) is imputed in the CPS ASEC due to nonresponse in recent years (Hokayem, Raghunathan and Rothbaum, 2022). We also do not observe gross wage and salary earnings (in the LEHD) for all jobs because not all jobs are covered by unemployment insurance and non-covered jobs are not reported to state UI offices. Gross earnings are also missing for jobs that are not available in the LEHD for other reasons, such as firms that erroneously fail to report jobs and states with no data-sharing agreement in a given year. For the missing survey responses and missing gross earnings, we observe a lot of information (variables in O) that can help us *predict* the missing values, such as W-2 job-level earnings, survey-reported occupation, hours and weeks worked, educational attainment, private health insurance coverage, etc.

[66] In this sense, misclassification can be important. If the imputed value equals true value for all cases, the data are not truly "imputed." However, in practice, imputations are unlikely to have extremely low misclassification rates, and we must evaluate the potential bias of each $\hat{Q}$ with the available information.

We use Sequential Regression Multivariate Imputation (SRMI) to impute plausible values for the missing data (Raghunathan et al., 2001).[67] SRMI is an iterative resampling technique to estimate $f(Y|O,\theta)$ while imposing fewer strong parametric assumptions on the joint conditional distribution f. Under SRMI imputation, We estimate the model for each Y_j iteratively as follows. In the first iteration, Y_1 is regressed on O and the missing values are imputed. Any imputation model can be used to impute values for each Y_j, such as a regression model, a hot deck, or predictive mean matching, with their attendant assumptions about $f(Y|O,\theta)$. Let $Y_1^{(1)}$ denote the filled-in version of the variable Y_1 from the first iteration. Now Y_2 is imputed using $(O, Y_1^{(1)}$ as covariates to generate $Y_2^{(1)}$, the filled in version of Y_2 from the first iteration. This process continues until the missing values in Y_p are imputed using $(O, Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{p-1}^{(1)})$ as predictors.

We cannot stop at iteration 1 because the imputation of $Y_1^{(1)}$, for example, fails to exploit the observed information from $(Y_2, Y_3, \ldots, Y_p)$. Iterations $t = 2, 3, \ldots$ proceed in the same manner except that all other variables (with some filled at the current and the rest in the previous iterations) are used in imputing each variable. Specifically, at iteration 2, Y_1 is re-imputed using $(O, Y_2^{(1)}, Y_3^{(1)}, \ldots, Y_p^{(1)})$ as predictors; Y_2 is re-imputed using $(O, Y_1^{(1)}, Y_3^{(1)}, \ldots, Y_p^{(1)})$ as predictors, etc. In each iteration, we are updating our predictions of θ as well as Y.

In general, at iteration $t > 1$, Y_j is re-imputed using $(O, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)})$ as predictors. The iterations are continued several times in order to fully use the predictive power of the rest of the variables when imputing each variable. Empirical analysis has shown that fewer than 20 and generally as few as 5 to 10 iterations are sufficient to condition the imputed values in any variable on all other variables (Ambler, Omar and Royston, 2007; Van Buuren, 2007; He et al., 2010). By repeating the imputation process in each iteration, SRMI is akin to a Gibbs or MCMC resampling technique that should iteratively converge to the true conditional joint density (if the model is properly specified).

We impute survey earnings, job-level administrative gross earnings (or LEHD-equivalent earnings), and missing state-level means-tested program data. For survey earnings, we impute extensive

[67]SRMI has also been called Fully Conditional Specification and Flexible Conditional Models in the literature.

margin earnings receipt and intensive margin earnings amounts for all earnings variables. In the CPS ASEC this includes the variables ern_yn (earnings receipt), ern_srce (primary job earnings source - wage and salary, self employment, or farm self employment), ern_val (earnings amount from primary job), ws_yn, se_yn, and frm_yn (secondary wage and salary, self employment, for farm self employment earnings?), and ws_val, se_val, and frm_val (amount of secondary earnings in each category). We also impute upstream variables that are highly predictive of earnings, including weeks worked last year (wkswork) and hours worked per week last year (hrswork).

For gross earnings by job (for the two highest earning jobs for each worker), we impute several variables to simplify the imputations and capture important features in the data. First, we impute a dummy variable for whether gross earnings ≈ taxable earnings + deferred compensation, which is true for a large share of workers. For those where gross earnings > taxable earnings + deferred compensation, we then impute a series of dummies for whether gross earnings/(taxable earnings + deferred compensation) falls in several bins, including 1.1 and above, [1.05, 1.1), [1.03, 1.05), [1.02, 1.03), [1.01, 1.02), and (1, 1.01). After assigning each job to a gross earnings/(taxable earnings + deferred compensation) bins, we then impute the amount of gross earnings for each job. We chose this approach because many variables (such as survey-reported private health insurance coverage) are good predictors of whether gross earnings/(taxable earnings + deferred) compensation exceeds specific thresholds while not necessarily being good predictors of the exact value of gross earnings/(taxable earnings + deferred).

For each earning variable, we have separate imputation models by spouse (by sex if an opposite-sex couple, by order on the file if a same-sex couple). This allows for a more flexible imputation model and allows us to condition on spousal income in the SRMI.

For state-level means-tested program data, we impute program receipt ({Program}_yn) and, conditional on receipt, the amount received ({Program}_val) for each program at the household level.

As discussed in Hokayem, Raghunathan and Rothbaum (2022), there are a number of challenges to implementing SRMI in this context. First, many income types do not follow a normal distribution. Second, we must select predictors for the modelling of each income variable from a very large set of possible covariates. Third, we must properly account for uncertainty in our estimates of the parameters in θ. Included in this uncertainty is the selection of variables for our imputation models because when we select predictors for our models, we are imposing the assumption that there is no relationship between the excluded variables and the variable being imputed conditional on the included variables. Next, we discuss how we address each of these issues.

To address non-normality, we transform each continuous variable using the inverse hyperbolic sine, which allows us to include negative values, as in Fox et al. (2022).[68]. As the inverse hyperbolic sine is nearly perfectly correlated with the natural log over most of the defined range of the natural log, one can interpret the regression coefficients of continuous variables as elasticities (for continuous dependent variables) or semi-elasticities (for binary dependent variables).

As a practical matter, there are too many potential variables in O to be used in our model. We reduce the set of variables to be used to impute each Y_j in two stages, both using the Least Absolute Shrinkage Operator (LASSO, Friedman, Hastie and Tibshirani (2010)). In the first stage, we take all of the possible interaction terms we specify in O and use LASSO to prune the list to $\hat{O}_j$ that predict Y_j (including all non-interacted terms in $\hat{O}_j$). The set of variables in $\hat{O}_j$ will generally be large (hundreds of variables and interactions, if the regression sample size is large). In terms of the general notation $f(Y|O,\theta)$, this process places constraints on θ.[69].

During the imputation process, we have a second-stage of regularization when we estimate the values in $\hat{\theta}$. As $\hat{\theta}$ is a set of unknown parameters, we also must incorporate the uncertainty in $\hat{\theta}$ into the imputation process – the third challenge noted above. We do this as follows. In each implicate c (independent run of the imputation model), we start by taking a Bayesian Bootstrap of the sample, we then do a second-stage variable selection process to further reduce the number of variables in $\hat{O}_j$ to $\hat{O}_{j,c}$, again using LASSO regularization.[70] From the regression of Y_j on $\hat{O}_{j,c}$, we estimate $\hat{\theta}_{j,c}$.

[68]Hokayem, Raghunathan and Rothbaum (2022) tested alternative transformations, such as Tukey's gh transformation (He and Raghunathan, 2006) and an empirical normal transformation (Woodcock and Benedetto, 2009). However, as in Fox et al. (2022), they found the inverse hyperbolic sine performed well, and we use that transformation here.

[69]This is primarily done for practical speed considerations. Reducing the number of candidate variables upfront considerably speeds up the process of imputation for each variable in each implicate.

[70]The Bayesian Bootstrap (Rubin, 1981) is the Bayesian analogue of the bootstrap. Each observation is drawn (with replacement) with an expected probability of $1/n$, but with variability. The probabilities of being drawn are defined by taking $n-1$ draws from the uniform distribution (0,1), ordering draws from

Doing this on a Bayesian Bootstrap sample enables us to account for the uncertainty present in each step of this process, including which variables are used as model predictors ($\hat{O}_{j,c}$) and to draw from the distribution of parameters values $\hat{\theta}_{j,c}$. This resampling approach to estimating uncertainty in regression-based imputation has been taken in other data products and research, including SIPP topic flag imputation (Benedetto, Motro and Stinson, 2016), the SIPP Gold Standard and SIPP Synthetic Beta (Benedetto, Stinson and Abowd, 2013), and imputation research on missing income in the CPS ASEC (Hokayem, Raghunathan and Rothbaum, 2022).

With the transformed continuous variables, regularization, and Bayesian Bootstrap-based estimation of the uncertainty of $\hat{\theta}$, we are almost ready to impute missing values. We must also specify the functional form of our imputation models (parametrizing $f(Y|O,\theta)$). Unless otherwise indicated, we use predictive means matching (PMM) to impute both binary and continuous dependent variables.

For binary dependent variables, we use a Linear Probability Model (LPM), regressing the dependent variable on the model selected using the LASSO on the Bayesian Bootstrap sample. We then predict the vector $\hat{p}_j(Y = 1|X, \hat{\theta}_j)$, which includes the estimated probability for all individuals in sample whether $R_j = 0$ or $R_j = 1$. We then take a random draw for each unit i where $R_{i,j} = 0$ from the ten nearest units k where $R_{k,j} = 1$ to assign $Y_{i,j}$ values. We use LPM rather than a logit or probit model as the LPM model more predictor variables. Although LPM does not impose $0 \leq \hat{p}_{i,j} \leq 1$, the $Y_{i,j}$ draws must equal 0 or 1. Fox et al. (2022) used the same approach for imputing SNAP receipt and showed that this PMM model performed well for several conditional and unconditional statistics (Q's such as SNAP receipt, SNAP receipt conditional on earnings and demographics, for example).

For continuous dependent variables, we use Ordinary Least Squares (OLS), regressing the dependent variable on the model selected using the LASSO on the Bayesian Bootstrap sample. We then predict

lowest to highest, where $u = u_0, u_1, u_2, \ldots, u_n$ given $u_0 = 0$ and $u_n = 1$. The probability of being drawn for each observation i is based on the gaps between each adjacent value in u, so that for observation i the probability of being drawn is $g_i = u_i - u_{i-1}$. As noted in Benedetto, Stinson and Abowd (2013), using the Bayesian Boostrap adds additional variability to the imputation process to account for the fact that the sample distribution may not be the same as the population distribution. Without the use of the Bayesian Bootstrap, the confidence intervals would not be proper.

the vector $\hat{Y}_j(Y_{-j}, X, \hat{\theta}_j)$ where Y_{-j} is the matrix Y excluding Y_j, again for all individuals in sample whether $R_j = 0$ or $R_j = 1$. We then take a random draw for each unit i where $R_{i,j} = 0$ from the ten nearest units k where $R_{k,j} = 1$ to assign $Y_{i,j}$ values.

For survey wage and salary earnings from the longest job (ern_val if ern_srce == 1), rather than using PMM, we use a two-stage model that incorporates OLS and quantile regressions. As before, we first use OLS to predict $\hat{Y}_j(Y_{-j}, X, \hat{\theta}_j)$ after LASSO regularization. We then use quantile regression to regress Y_j on binned $\hat{Y}_j$ and several variables from O, including race and Hispanic origin, age, education, and hours worked. We do this for each 5th percentile from the 5th to the 95th. This gives us an estimate for $\widehat{Y_{j,i,q}}$ for each individual i at each quantile q.[71]. From the values of $\hat{Y}_{j,i,q}$, we have a posterior predictive distribution (PPD) of $Y_{j,i}$ for each individual i (after interpolation using Schmidt et al. (2022)). For each individual, we then draw a percentile value from 0 to 1 to impute $Y_{j,i}$ from the PPD. [72]

Using quantile regression to estimate the PPD is useful if there is potential heterogeneity in the relationship between specific variables in O and Y_j. For example, suppose the average relationship between education and earnings reflects a bigger right tail for college graduates (more very high earners), the PMM-based estimate would not necessarily reflect that in the resulting imputes. However, the quantile regression-based PPD would. However, more data (a large sample) is required to use quantile regressions to reliably estimate the PPD. Because of the possibility of heterogeneity and the greater data needs, we implement this approach from survey wage and salary earnings from the primary job (the largest single source of survey income, covering almost 70 percent of total income).

For the means-tested program variables imputed at the household level, we recode the data to summarize the information of household members (such as presence of members by race, total

[71]The regressions do not impose monotonicity, i.e., it does not ensure that for two quantiles q and r where $r > q$, $\hat{Y}_{j,i,r} > \hat{Y}_{j,i,q}$ (the quantile crossing problem). Following Chernozhukov, Fernández-Val and Galichon (2010), we rearrange the curve by sorting the $\hat{Y}_{j,i,q}$ values from lowest to highest and assigning them to the corresponding position's q value. As Chernozhukov, Fernández-Val and Galichon (2010) show, the rearranged curve is closer to the true quantile curve than the original curve in finite samples.

[72]If any part of this process fails (such as from nonconvergence in a quantile regression estimate), we impute using PMM. This is unusual, but possible, in an automated process like SRMI that runs many regressions per iteration repeated across implicates.

household earnings, etc.) and household head variables (such as education, race, etc.) to use as predictors and then impute receipt and amounts using PMM as discussed above.

For nonfilers, we observe whether they received several information returns, including Forms 1099-G, 1099-INT, and 1099-DIV in the IRMF. From these we have information on whether they received UI compensation, interest income, and dividends, respectively. Each of these are vastly underreported on surveys (Rothbaum, 2015). Rothbaum (2023) has been working with more detailed data available under a separate agreement between the Census Bureau and IRS, for limited use. In that work, the 1099-G, 1099-INT, and 1099-DIV data is available, including income amounts. Rothbaum (2023) released coefficients that can be used to impute these amounts for nonfilers conditional on survey responses and the administrative data used in this project.

To release this statistics, Rothbaum (2023) estimated models for the synthesis of four variables:

1. UI compensation receipt conditional on receipt of a Form 1099-G
2. UI compensation amount conditional on receipt of UI compensation
3. Interest income amount conditional on receipt of a Form 1099-INT
4. Dividend income amount conditional on receipt of a Form 1099-DIV

In order to allow the creation of synthetic data to correct for survey underreporting, Rothbaum (2023) released three sets of results for each variable.

For UI compensation receipt, they estimate a Linear Probability Model (LPM) of UI compensation receipt conditional on receiving a Form 1099-G. Individuals receive a 1099-G for various government payments, including (1) UI compensation, (2) state or local income tax refunds, credits, or offsets, (3) reemployment trade adjustment assistance payments, (4) taxable grants, and (5) agricultural payments. This model is estimated as described above using the two-stage LASSO feature selection, with the second stage estimated on a Bayesian Bootstrap. As such, the released parameters are effectively a draw from the distribution of possible parameter estimates that could be used to predict nonfiler UI receipt.

With these regression coefficients, we can estimate the expected probability of UI receipt for each nonfiler ($\hat{p}_j(Y = 1|X, \hat{\theta}_j)$) on a separate sample (or the data without access to the more detailed 1099-G data). However, as they were estimated using a LPM, we cannot directly use them to synthesize UI receipt data (as the $\hat{p}_j(Y = 1|X, \hat{\theta}_j)$ can be < 0 or > 1, which PMM addresses by taking a random draw from individuals with similar $\hat{p}_j(Y = 1|X, \hat{\theta}_j)$, but with observed values for Y_j. Instead, Rothbaum (2023) then separate the expected probability space into bins and released the boundaries between those bins and the empirical probability that an observation received UI compensation in each bin. For example, the top quintile of observations has an expected probability of receipt of 0.87 or higher (the boundary). Within that bin of observations with an expected probability of 0.87 or higher that received UI compensation was 0.98 (the empirical probability in the bin), then we can impute UI receipt for this group by drawing a random number between 0 and 1 and assigning receipt if it is ≤ 0.98.

By releasing regression coefficients, bin boundaries, and empirical probabilities, Rothbaum (2023) implement a semiparametric imputation technique that is similar to the binned imputation proposed by Bondarenko and Raghunathan (2007).

For the income variables – UI compensation, interest income, and dividends – the approach is slightly different. The first step is the same as above for continuous variables – estimate an OLS model to predict expected income amounts conditional on the available information. Again, the models are estimated using the two-stage LASSO feature selection, with the second stage estimated on a Bayesian Bootstrap. The coefficients from this model are released so that the expected income amount can be estimated on a separate sample ($\hat{y}_{i,j}$). To allow the synthesis of continuous variables, Rothbaum (2023) release two set of variables. First, they partition $\hat{y}_{i,j}$ into bins. Then, using quantile regression at various percentiles, the regress income amounts on bin dummies. As with ern_val above, these regression coefficients can be used to estimate a PPD for each individual. By drawing a value from 0 to 1, we can impute income amounts from these PPDs.

In summary, for each income amount synthesized, Rothbaum (2023) release three sets of statistics, regression coefficients, bin boundaries and quantile regression coefficients to enable relatively low dimensional data to be used to synthesize or impute UI compensation amounts, interest income, and dividends.

Finally, we repeat this process five times, to create the five independent implicates. In each implicate, we use SRMI to impute the survey and gross earnings variables, followed, in a separate step, by the imputation of means-tested program variables. For any statistic or parameter estimate, we can account for the uncertainty in the imputation process (Rubin, 1976). To do so, we calculate the total variance by combining the within-implicate variation (for example, the standard error of an estimate in one implicate) with the between-implicate variation (the variance of the estimates for that parameter across the five implicates).

In Table 6, we show the rates of missing data for survey earnings, state program data, and LEHD job-level gross earnings. In the 2019 CPS ASEC, 46 percent of individuals with earnings had their primary job earnings imputed. We do not have state-level administrative TANF data for 47 percent of households. Finally, we impute gross earnings for 18 percent of jobs, either because there is no LEHD information for them (8 percent of highest earning jobs) or because the LEHD and W-2 values disagree substantially (i.e., the LEHD < W-2, about 10 percent of highest earning jobs).

As the imputation models are applications from prior work (Hokayem, Raghunathan and Rothbaum 2022 for earnings, Fox et al. 2022 for means-tested benefits, and Rothbaum 2023 for nonfiler UI, interest, and dividends), we provide limited statistics on the imputation outputs. Table A6 shows summary statistics for survey earnings imputation, comparing the CPS ASEC imputations to the NEWS SRMI imputations conditional on W-2 earnings. The SRMI estimates fewer individuals with zero survey earnings conditional on having zero W-2 earnings. Also, the SRMI estimates higher survey earnings conditional on having higher W-2 earnings (such as in the 5^{th} quintile of W-2 earnings). Table A7 provides some summary statistics for means-tested program imputation.