CARRA Working Paper Series
Working Paper 2018-06

The Opportunities and Challenges of Linked IRS Administrative and Census Survey Records in the Study of Migration

Thomas B. Foster
U.S. Census Bureau

Mark Ellis
University of Washington

Lee Fiorio
University of Washington

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: July, 2018

Abstract

This paper details efforts to link administrative records from the Internal Revenue Service (IRS) to American Community Survey (ACS) and 2010 Census microdata for the study of migration in the United States. Specifically, we (1) document our linkage strategy and methodology for inferring migration in IRS records; (2) model selection into and survival across IRS records to determine suitability for research applications; and (3) gauge the efficacy of the IRS records by demonstrating how they can be used to validate and potentially improve migration responses in ACS microdata. Our results show little evidence of selection or survival bias in the IRS records, suggesting broad generalizability to the nation as a whole. Moreover, we find that the combined IRS 1040, 1099, and W2 records may provide important information on populations that are hard to reach with traditional Census surveys. Finally, while preliminary, the results of our comparison of IRS and ACS migration responses shows that IRS records may be useful in improving ACS migration measurement for respondents whose migration response is proxy, allocated, or imputed. Taking these results together, we discuss the potential applications of our longitudinal IRS dataset to innovations in migration research in the United States.

Introduction:

Administrative records – information collected as part of the regular operation of federal and state programs – are an increasingly important source of data on the U.S. population. Indeed, the U.S. Census Bureau operates with a congressional mandate to use administrative records to improve, cut costs associated with, and reduce respondent burden on Census surveys such as the American Community Survey (ACS), as well as the decennial census (Johnson, Massey, and O'Hara 2015). Among the wide array of administrative records employed in this capacity, records from the Internal Revenue Service (IRS) stand apart in terms of their breadth of coverage, consistency over the long term, and long-standing use in Census operations. IRS 1040 income tax return records are used internally at the U.S. Census Bureau in the construction of yearly population estimates, and the IRS's Statistics of Income Division publishes intercounty and interstate migration flows among 1040 filers, dating back to the 1990 tax year.

IRS records like 1040 income tax returns are a particularly attractive source of administrative migration data. The vast majority of American earners file a tax return each year, and at approximately the same time each year (April 15th). Filers are incentivized to provide accurate address information to avoid IRS scrutiny and ensure timely reimbursement for any excess taxes paid in the preceding year. And, because dependents and spouses claimed can decrease tax burdens, filers are incentivized to accurately report other members of their household. Furthermore, administrative records may offer unparalleled insights into the migration behaviors of populations that can be hard to reach with traditional surveys. These advantages have encouraged the use of 1040 records in innovative academic research exploring links between geography and earnings over the life course (Chetty et al. 2014; Chetty and Hendren Forthcoming).

Despite these advantages, questions remain about the suitability of IRS administrative data, in general, and IRS 1040s, in particular, for research applications. Chief among these, perhaps, is the issue of filers' and earners' self-selection into the universe of IRS administrative records. Answers to this question thus far tend to address selection bias for only subsets of the 1040 records. For example, linking individuals age 25 to 65 found in the 1040 records to the 2010 Census, Akee, Jones, and Porter (2017) find that female, White, and Asian individuals are more likely to be found in the 1040 record, while males and other racial minorities are less likely to be represented. Presumably, a lack of representation in the universe of 1040 filers reflect lower incentives for filing among some groups. Those earning low wages or not participating in the labor force gain little from filing income tax returns, and, because Black, American Indian, Pacific Islander, and Hispanic individual tend to earn lower wages than their White and Asian counterparts (Ramakrishnan and Ahmad 2014), this may explain their underrepresentation in the 1040 record. A corollary concern with respect to migration studies is the extent to which selection bias into the IRS records is exacerbated by survival bias over time. That is, to what extent is any measure of

migration obtained from IRS administrative records biased by differential odds among sociodemographic groups of "dropping out" and, therefore, not being counted in migration rates? Moreover, because studies thus far primarily utilize 1040s, it remains unclear what role, if any, IRS 1099 (reports of income for tax preparation purposes) and W2 earnings records may play in minimizing selection and survival bias in IRS administrative records research.

Thus, the goals of this paper are threefold. First, we describe and create a longitudinal database of earners, filers, and their families found in a combined suite of IRS 1040, 1099, and W2 administrative records. Second, we address selection and survival bias in the suite of IRS records, particularly as these sources of bias relate to the application of our longitudinal migration database to migration research. Third, we link our IRS migration database to restricted-use ACS microdata to demonstrate the efficacy of IRS records in validating and improving ACS migration items.

The motivation for these goals is to explore the utility of IRS records for the longitudinal analysis of migration among population subgroups, across small geographies, and with high levels of precision. Migration is a relatively rare event, and researchers have raised questions about the precision with which surveys such as the ACS can measure these rare events, particularly among relatively small populations and across sub-state geographies (Conway and Rork 2016; Franklin and Plane 2006; Raymer and Rogers 2007; Rogers, Jones, and Ma 2008). Even migrations between small and medium-sized states may be imprecisely measured by annual ACS data (i.e. has large error bands), reducing their utility for demographic estimates and assessment of trends. The population scale of IRS migration data offers a possible solution to these data precision problems. The longitudinal structure of these data provides new opportunities for more precisely assessing the role of factors that perhaps differentially motivate the internal migration of population subgroups, such as changes in family structure (partnering, fertility events) and changes in origin and destination economic conditions and policy environments. The information in these data on pre- and post-migration income and employment status also allows for better measurements of returns to migration than the cross-sectional ACS, which has no information on pre-migration employment conditions and income.

The results of our selection and survival analyses show that the population found in IRS 1040s, 1099s, and W2s is remarkably similar to that found in the 2010 Census with respect to population age, sex, regional, and nativity distributions. Survival rates among individuals identified in the IRS record are also quite high: roughly 96 percent of records identified in a given tax year will survive to the next, and 87.7 percent of all individuals identified in tax year 2000 IRS records are also found in tax year 2015. In general, our findings suggest that the suite of IRS 1040s, 1099s, and W2s is broadly representative of the U.S. population, though linking IRS with other microdata may limit generalizability.

As such, our results suggest that the suite of IRS administrative records provide an innovative source of data on migration in the United States. We demonstrate the efficacy of these data by comparing IRS and ACS microdata in order to benchmark ACS migration responses. Though they remain preliminary and research is ongoing, the results of this comparison find that IRS administrative record migration data may prove particularly useful in improving ACS migration responses that would otherwise be proxy, allocated, or imputed. We conclude this paper with a discussion of the ways in which the longitudinal IRS administrative dataset created, vetted, and demonstrated in this paper facilitate innovative research on migration and related phenomena like segregation, gentrification, and the links between geographic mobility and intra-household earnings dynamics.

Data and Methods:

Administrative Record Sources

This project utilizes four administrative records sources – three sources of IRS Federal Tax Information (FTI) and one sorce from the Social Security Administration (SSA). The first source of FTI, IRS 1040s for the tax years 2000 through 2015, form the backbone of our data. Though scrubbed of Personally Identifiable Information (PII) like names and Social Security Numbers (SSNs), the 1040 records retain much of the data filers report when filing income taxes each year. This includes addresses, adjusted gross income and incomes from various sources, exemptions claimed, information on dependents and spouses, and the sorts of tax credits claimed by the filer (e.g. the Earned Income Tax Credit). [1]

The second source of FTI, IRS 1099 "information returns" for tax years 2003 to 2015, supplement the universe of filers, spouses, and dependents found in IRS 1040s. The 1099s include any number of documents provided to potential income tax filers, such as income and withholdings information for self-employed individuals, as well as other forms delivered to individuals for the purposes of tax preparation. Like the 1040s, the 1099s are scrubbed of PII but do retain address information, making them valuable for this study of migration. In addition, they facilitate insight into a population typically overlooked in studies of migration relying solely on 1040s – namely, those who earn incomes but who, for any number of reasons, choose not to file a tax return. As such, the 1099s allow researchers to capture those populations typically self-selected out of the universe of filers.

The final source of FTI is the universe of IRS W2s for tax years 2005 to 2015. IRS W2s contain detailed earnings information for *individual earners*. While this source of FTI does not contain address information, thereby limiting its use for measuring migration, it does allow insight into individual contributions to gross household earnings. For migration studies, this means that researchers can

[1] Our records identify primary and secondary filers and up to four dependents. As such, up to six individuals can be found in a single IRS 1040 record.

investigate whether and how migration influences the earnings of spouses differently, rather than relying on gross reports of household income.

We also rely on one other administrative record source from the SSA, which provides sex, date of birth, date of death, and place of birth information for the population found in FTI records. Drawn from the SSA's list of all names, Social Security Numbers (SSNs), and Individual Tax Identification Numbers (ITINs) issued in the United States, this file is provided for Census Bureau use and scrubbed of SSNs, names, and other PII. This file is particularly useful in this study because it provides important clues about the demographic composition of the population found in IRS 1040s and 1099s (particularly those who cannot be linked with decennial census records), while also allowing a better understanding of why individuals drop out of the IRS records over time (i.e., death).

Census Data Sources

To the administrative records sources discussed above, we add two sources of restricted-use Census Bureau microdata – 2010 Census records and 1-Year ACS responses from 2010 through 2013. The 2010 Census was the first in recent decades to proceed without the "long form" questionnaire, which included, among other modules, one assessing whether the respondent had moved in the last five years. Nevertheless, that Census collected detailed information on the demographic composition of American households containing over 300 million individuals. As such, the 2010 Census serves as a benchmark against which the population found in IRS FTI records can be compared and contrasted.

Fully implemented in 2005, the ACS is an ongoing survey sampling over 2.8 million housing unit addresses a year and providing a wealth of information on the demographic and socioeconomic characteristics of the U.S. population.[2] It also contains a module about the migration of the respondent in the last year. As described in greater detail below, we use individual 1-Year ACS migration responses from 2010 through 2013 to both demonstrate the viability of the IRS data as a source of migration data and to explore the ways in which the IRS data may help reduce respondent burden and improve ACS estimates.

Linking Administrative and Census Records

As noted above, all administrative and Census records are scrubbed of PII such as names and SSNs before researchers gain access in order to protect privacy and reduce the risk of disclosing personal information; in their stead, the Census Bureau assigns individuals unique, anonymous personal identifiers

[2] The ACS design and methodology report can be found at https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html. Additional information on ACS methodology, variable definitions, code lists, comparisons, statistical testing, and the accuracy of estimates can be found at https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2010.html.

which facilitate individual-level record linkage across data sources. As outlined in Wagner and Layne (2014), these unique identifiers are assigned on SSNs or ITINs, when available, and probabilistically assigned using date of birth and address information provided by respondents when an SSN is not available. Not every record receives these unique identifiers, but rates of unique person identifier assignment differ across data sources. Because SSNs or ITINs are required on tax documents, over 99 percent of individuals found in IRS administrative records receive a person identifier. Over 91 percent of individuals in the 2010 Census were assigned unique person identifiers, while the comparable rate for ACS surveys is approximately 94 percent. We restrict our analyses to the universe of individuals with person identifiers in each of the data sources listed above, as they are required for record linkage.[3]

Our data linkage infrastructure is depicted in Figure 1. The IRS 1040s, shown in the center of the figure, form the backbone of our linked data. The lower half of the figure details the longitudinal linkages between IRS 1040s, 1099s, and W2s. The upper half of the figure details linkages between IRS, ACS, and decennial census data sources. Because the tax records for any given year are not filed until approximately April 15th of the following year, we link IRS records from year *y* to ACS and decennial census records from year *y+1* such that linked records reference roughly the same period in time. This structure provides longitudinal location and income information for the universe of individuals assigned unique identifiers in IRS administrative records for tax years 2000 through 2015, as well as basic demographic information (age, race/ethnicity, gender, nativity) for all individuals found in the 2010 Census and detailed residential, socioeconomic, and household information on the subset of individuals linked to the ACS.

Measuring Migration in ACS and IRS Records

In each year since its full implementation in 2005, the ACS has gauged migration in the United States by asking respondents whether they live in the same house they did one year ago, or whether they have moved. For those who live in a different house, a series of follow-up items inquire as to whether and which administrative boundaries (e.g. state, county) were crossed when the respondent moved.

As with any survey item, response rates on the migration item(s) are not 100 percent, and the ACS uses several different strategies to account for non-response. First, householders or interviewees will often provide proxy responses for other household members, which will be used when available. Second, migration responses will be assigned based on other information provided by the respondent that suggests

[3] Patterns in the failure to assign unique identifiers are non-random and may introduce bias in subsequent analyses. Research on ACS microdata has shown that racial/ethnic minorities and those of lower socioeconomic status are less likely to receive identifiers than non-Hispanic Whites and those of higher socioeconomic status (Bond et al. 2014).

a move occurred in the last year. Finally, in the absence of both proxy and assigned responses, the ACS uses a hot deck technique to impute migration responses.

Migration in the IRS administrative records is inferred by first longitudinally linking individuals across tax years, and then checking for changes in filing or mailing addresses found in 1040 and 1099 records. One method for geolocating individuals is to use address information from these forms. However, the researcher is typically quite limited because the most consistent administrative boundary available is the state. A more sophisticated geolocation method relies on 9-digit postal Zip Codes. Using a special file developed with the United States Postal Service (USPS), 9-digit Zip Codes (which identify areas as small as a building or a block side) found on IRS files can be converted to Federal Information Processing Standard State and County codes. This is essentially the same method used to gauge interstate and intercounty migration by both the IRS's Statistics of Income Branch (to produce public migration estimates) and the U.S. Census Bureau (to estimate the migration components of population change for yearly population estimates for Congressional use).

In this paper, however, we use a third geolocation method that relies on unique, anonymous address identifiers (akin to those assigned individuals in place of PII like names and SSNs) to measure migration. The Census Bureau maintains and uses the Master Address File (MAF), an evolving list of all addresses in the United States. This file serves as the sampling frame for the majority of Census Bureau surveys (e.g. decennial census, ACS) authorized under Title 13 of the U.S. Code.[4] However, it can also be used to provide consistent administrative and geographic boundary information (down to the block level) on historical administrative record addresses. Unique address identifiers are assigned by matching standardized address information found on IRS records with addresses found in the MAF. Using this matching procedure, nearly 90 percent of all addresses found in IRS administrative records are assigned unique address identifiers. We can then infer migration for individuals in the IRS records by comparing unique address identifiers for an individual at two points in time, or use the unique identifiers to gauge migration across any administrative boundary of interest.

We use unique address identifiers to create retrospective migration variables, gauging migration between tax years *y-1* and *y*, for individuals found in the IRS 1040s and 1099s. This retrospective view is advantageous because it approximates the retrospective nature of ACS migration items. Because of death among those found in the 1040 and 1099 records, as well as incomplete address identifier assignment and the tendency for some individuals to periodically "drop out" of the IRS for spells of varying lengths, migration cannot be measured consistently among all individuals identified in IRS records.

[4] See https://www.gpo.gov/fdsys/pkg/USCODE-2007-title13/pdf/USCODE-2007-title13.pdf for the full text of Title 13 of the U.S. Code.

Results:

Selectivity Bias in the Universe of IRS Administrative Records

To understand selection into the universe of IRS 1040, 1099, and W2 records, we link individuals from tax year 2009 records to 2010 Decennial Census (those with both a person and address identifier) and SSA records (see the top portion of Figure 1). This staggering of IRS and decennial census records ensures that the populations in each data source are measured at approximately the same point in time; that is, IRS records for the 2009 tax year are filed by April 15, 2010, while the 2010 Census asks respondents to use April 1, 2010 as a reference point.

As shown in Table 1 (Columns a and c), over 275 million 2010 Census records[5] and 300 million 2009 IRS records were linked to a record in the SSA file. Of those 275 million Census records, 94.4 percent (over 260 million, Columns e and i) were also linked to IRS records. However, this includes individuals found only in IRS W2 records, which do not contain address information necessary for measuring migration. If we limit our analysis to only those IRS records with unique address identifiers, we find a linkage rate of 88.5 percent (nearly 245 million records, Columns g and j). As such, the 2009 IRS records cover the vast majority of all individuals with a unique person identifier in the 2010 Census, even when we limit IRS records to those with a valid unique address identifier.

The population in IRS records that cannot be linked to the 2010 Census, while sizeable, is not altogether unexpected. Prior research linking a broad array of federal, state, and commercial administrative records – including records from the IRS – to the 2010 Census found mismatches in the populations covered in decennial census and administrative records. Rastogi and O'Hara (2012) report that 48.8 million individuals identified in joint administrative records and assigned person identifiers could not be linked to the 2010 Census, while 5.5 million individuals enumerated and assigned person identifiers in the 2010 Census could not be linked to administrative records. A substantial portion of this mismatch is likely attributable to lower rates of person identifier assignment in the decennial census relative to IRS records. Of the over 308 million people enumerated in the 2010 Census, 280 million (91 percent) were assigned person identifiers and, for this analysis, only 276 million could be linked to the SSA administrative record. Therefore, it is likely that the 43 million unlinked IRS records are, by and large, individuals enumerated in the 2010 Census but not assigned person identifiers and/or linked to the SSA administrative record. There are other reasons for the mismatch in IRS and 2010 Census. As prior research utilizing IRS 1040s has noted, the population earning little to no wage income will, of course, be

[5] The 2010 Census enumerated 308,745,538 individuals. The 276.7 million individuals found in the 2010 Census microdata and described in Table 1 do not include the approximately 10 percent of individuals enumerated in the 2010 Census to whom the Census Bureau could not assign a unique person identifier and, therefore, could not be linked to SSA or IRS administrative records.

undercounted in IRS records. Supplementing 1040s with 1099s corrects to some extent for this undercount, but IRS administrative records will inevitably undercount economically marginalized groups. Furthermore, because children born between January 1, 2010 and April 15, 2010 cannot be claimed on tax returns for the 2009 tax year, the population under one year of age will be systematically undercounted in the IRS administrative records. Decennial census undercounts for segments of the U.S. population – such as young children, racial and ethnic minorities, and the foreign born – are also known and well documented (Rastogi and O'Hara 2012) despite significant efforts by decennial operations to mitigate non-response. As such, though imperfect, IRS administrative records may include members of population subgroups traditionally undercounted in the decennial census.

Broadly speaking, Table 1 shows that the populations found in decennial census and IRS records are similar with respect to their distributions across age, sex, region, and nativity categories. IRS and Census population shares in each category are typically within +/- 1 percent point of one another. Comparing age distributions in columns b and d the 18-24, 25-44, and 65+ age groups are slightly overrepresented in IRS records relative to the 2010 Census population, while the 0-2, 3-4, 5-17, and 45-64 age groups are slightly underrepresented. Males are slightly overrepresented in IRS records, while females are slightly underrepresented, likely reflecting the male-breadwinner gender stereotype prevalent in the United States. Regional distributions are also broadly similar, with the South claiming the largest share of both Census and IRS records, followed by the West, Midwest and Northeast. Finally, though the IRS records include a larger number of both native-born and foreign-born individuals than the 2010 Census population with unique person identifiers, foreign-born individuals are slightly overrepresented in the universe of IRS records, making the IRS records a potentially valuable source of information on foreign-born and other difficult to reach populations.

Finally, Table 1 suggests that the extent to which IRS administrative records are representative of the U.S. population, as a whole, depends heavily on how they are used. Taken alone, the distribution of records in the universe of IRS 1040s, 1099s and W2s is remarkably similar to the distribution of records in the 2010 Census with respect to age, sex, nativity, and geography (compare Columns a and b with c and d). These similarities suggest that: (1) augmenting 1040 records with 1099s and W2s pays large dividends in terms of correcting biases introduced by self-selection into the population of tax return filers; (2) research utilizing IRS records alone are, broadly speaking, generalizable to the United States as a whole; and (3) IRS records may offer insight populations difficult to reach in decennial census operations. However, the representativeness of IRS microdata deteriorates when it is linked with records from other sources and/or restricted to records with valid address identifiers. On average, 94.4 percent of all 2010 Census records are also found in the IRS universe (Column i), but these coverage rates are significantly smaller for those under age 5, Hispanics, and all racial groups with the exception of White alone. Gaps in

coverage grow when Census-IRS linkages are limited to those in the IRS universe with a valid address identifier, such that only 88.5 percent of Census records are linked. These patterns may reflect group differences in the likelihood that individual 2010 Census records are assigned unique person identifiers and, therefore, are linked with IRS records (see also Akee, Jones, and Porter 2017). They may also originate from 15 million individuals (Table 1, columns a and e) in the 2010 Census who are not found in the IRS records (see also Rastogi and O'Hara 2012). Regardless, research using IRS microdata relying on detailed geographic identifiers and/or on linking those data with other sources may have limited generalizability.

Survival Bias in the Universe of IRS Administrative Records

To understand survival bias in the population found in IRS 1040s, 1099s, and W2s, we first take a one percent simple random sample of records with unique person identifiers from each tax year. We then attempt to find individuals from each of these samples in IRS records from subsequent tax years. Using date of death information from the SSA administrative record, we distinguish individuals who drop out of the IRS record due to death from those who drop out due to, for example, failing to file. Then to understand the degree to which survival differs by various sociodemographic characteristics, we model "failure" (i.e., dropping out of the IRS records) using a Cox Proportional Hazards regression model (Allison 2014).

Table 2 reports the share of the total population found in IRS 1040s, 1099s, and W2s that can be found in subsequent tax records. Broadly speaking, survival curves for IRS populations change very little over time. On average, roughly 97 percent of IRS records are found in the following year; records "lost" are due in small part to deaths among the IRS population, but represent, for the most part, a simple lack of IRS documentation for those individuals in that year. This "loss" effect diminishes in 2003 when IRS 1099 records become available and is reflected in survival increases in the 2000 and 2001 populations between 2002 and 2003. Finally, survival rates over the long term are quite strong: 87.7 percent of the 2000 IRS population are found in 1040s, 1099s, and/or W2s from 2015.

With respect to survival among particular sociodemographic groups, we find that patterns of survival and "failure" in the IRS population over time generally reflect biases in the distribution of the linked IRS-Census population shown in Table 1. Table 3 presents the results of a Cox Proportional Hazards Model predicting natural death and/or "dropping out" of the IRS-Census linked universe over time as a function of sex, race, ethnicity, nativity, region, and age. Relative to their male counterparts, females are at a significantly lower risk of dropping out of the IRS universe over time; though this is inconsistent with the overrepresentation of males in all IRS records (Table 1, Column c/d), females are overrepresented among IRS records linked to the 2010 Census. Relative to those reporting their race in

the 2010 Census as White alone, those reporting Black alone, American Indian (AMIN) alone, Some Other Race (SOR) alone, or Multiple Race are at a significantly greater risk of dropping out of the IRS universe, while Asian alone individuals exhibit a significantly lower risk of dropping out. Hispanic and native-born individuals are also at a significantly higher risk of dropping out of the IRS records, relative to their respective Non-Hispanic and foreign-born counterparts. Relative to records with addresses in the West, those in the South are at a significantly higher risk of dropout, while the risk is significantly lower in the Northeast and Midwest. Finally, due to both retirement and increased mortality risk, age is positively associated with the risk of falling out of the IRS universe over time.

Comparing IRS Address Changes and ACS Migration Responses

High survival rates among the population found in IRS records, coupled with known and predictable patterns of selection into and out of the IRS universe, suggest that longitudinally linked IRS records present a viable and vital source of migration data. One potential application of these data involves validating and improving existing Census surveys such as the ACS. The following presents results from an investigation of the potential efficacy of IRS records in this capacity.

A comparison of aggregate interstate migration rates in IRS and ACS data provides a baseline for the person-level, microdata comparison that follows. As shown in Table 4, aggregate interstate migration rates in the IRS 1040s and 1099s are higher than aggregate interstate migration rates in publicly available 1-Year ACS data. For example, interstate migration for the 2009-2010 calendar years was 2.62 percent in our longitudinal IRS data, but 2.27 percent in the public ACS data. Though interstate migration rates trend upward in both datasets through the 2012-2013 migration interval, the rate of increase is larger in the IRS record. The IRS migration rate increased by roughly 28 percent between 2010 and 2013, while the ACS rate increased by only 5 percent over the same period.

A priori, the source of this growing gap between aggregate IRS and ACS migration rates is unclear. As shown in our discussion of selection bias, the IRS records are, by themselves, quite representative of the U.S. population as enumerated in the 2010 Census. One possibility is that limiting the otherwise representative IRS population to those with valid unique address identifiers in two successive tax years in order to measure migration disproportionately selects for populations that are more mobile. We will explore this possibility in future research. Another possibility, which we explore below, is that these aggregate differences reflect the sum of person-level differences in IRS and ACS migration responses.

To compare migration inferred in IRS records to responses in ACS surveys, we utilize person level linkages between IRS 1040 and 1099 records from tax year *y* and restricted-use ACS microdata from year *y+1*. As noted in our analysis of selection bias, this temporal staggering is necessary because

tax records for a given year contain address information for filers as of April of the following year. In this analysis, linking ACS records to IRS records from the previous year ensures that retrospective migration intervals in both data sources overlap as much as possible. However, because the ACS is conducted year round and IRS records are filed at approximately the same time each year, complete overlap in one-year retrospective migration intervals is impossible. To account for this mismatch, we utilize IRS processing dates, variables in the ACS microdata detailing the date the respondent last moved (if applicable), and time stamps showing when ACS responses were received and/or when interviews were conducted. These dates allow us to compare ACS migration responses with the appropriate migration interval in IRS records.

Nearly all 18.5 million individuals assigned unique person identifiers in the ACS microdata for 2010 through 2013 are linked to IRS 1040s or 1099s from at least one tax year, but IRS-ACS linkages in which a valid and appropriately timed IRS migration interval is observed are more difficult to come by. As shown in Table 5, an average of over 97 percent of 2010 to 2013 ACS records are also found in IRS records, and this percentage fluctuates very little in yearly ACS samples. However, only around 43 percent (8 million) of all ACS records are linked to an IRS record with an observed migration interval that overlaps by at least 10 months with the observed ACS migration interval.

Of those approximately 8 million ACS records linked to an IRS record with an appropriate migration interval, about 95 percent contain a direct migration report from the respondent, while the remaining responses are either proxy, assigned from another direct or allocated response, or imputed. Though rates of indirect migration measurement are quite low in this linked IRS-ACS sample, we present the results of our IRS-ACS migration comparison for both all linked records, as well as for the non-proxy, non-allocated, and non-imputed population alone.

Table 6 summarizes the results of our IRS-ACS migration comparison for all ACS respondents, regardless of imputation status for the ACS migration response. We disaggregate results by migration status reported in the ACS (i.e., movers vs. non-movers), then by whether inferred migration status in IRS records matches the ACS response. We refer to non-matches among those reporting a move in the ACS as "False Positives" and non-matches among those reporting not moving in the ACS as "False Negatives". Matches, false positives, and false negatives are also disaggregated by age, race, ethnicity, sex, region, and ACS migration imputation status.

Of the approximately 8 million ACS respondents in our linked 2010-2013 IRS-ACS population, 14.2 percent reported having moved in the reference year preceding their survey; IRS records corroborate the majority (57.0 percent, or over 640,000) of these reports. We classify the remaining 43.0 percent (some 480,000 records) as false positives. Match rates among those reporting migration in the ACS are highest among those under age 18 and between 25 and 44, but quite low among those of college (43.4

percent) and retirement (42.0 percent) age. Variability in migration match rates by racial group is low, though non-Hispanic and female respondent match rates tend to be higher than rates for their Hispanic and male counterparts. Regional variation in rates is also high, with above average IRS-ACS migration matching in the South (58.4 percent) and West (59.2 percent), and below average matching in the Northeast (51.7 percent). Finally, though the size of the linked IRS-ACS sample with an indirect migration report is small, it is clear that proxy, allocated, and imputed responses capture migration much less accurately than direct reports. Migration matches were above average for those ACS respondents reporting a move directly (58.3 percent), below average for those whose migration responses were proxy or assigned from other directly reported information (48.2 and 48.6 percent, respectively), lower still for those whose responses were assigned from other proxy or allocated information (31.7 percent), and lowest for those whose responses were imputed (26.8 percent).

For the vast majority (85.9 percent) of ACS non-movers linked to the IRS administrative records, we also find no evidence of a change in address in the IRS record. We classify the remaining 14.1 percent of ACS non-movers as false negatives, finding evidence of a change of address in the IRS record despite no such report in the ACS. False negative rates are highest among those children age 0 to 2 (22.2 percent) and those of college age (21.1 percent), but lowest among those 45 and older (just over 10 percent). White alone and Asian alone false negative rates are also quite low (12.6 and 12.8 percent, respectively), while rates are quite high for other racial groups (18 percent and above) and for Non-Hispanic respondents (24.6 percent). Geographically, false negatives are disproportionately concentrated in the West (17.1 percent), but are low in the Northeast and Midwest (11.4 and 12.2 percent, respectively). Finally, despite substantial variability in match rates among ACS movers by imputation status, we find less variability among those reporting no move. While rates of false negatives are highest among those with imputed responses (18.8 percent), rates among those with proxy responses (12.4 percent) are actually lower than those with direct, unedited responses (14.0 percent).

Table 7 restricts the analysis to those only those ACS respondents with direct, unedited migration responses; as noted above, this includes the vast majority (roughly 95 percent) of the linked IRS-ACS records. Restricting the sample in this way increases migration match rates marginally for ACS movers (from 57.0 to 58.3 percent) and non-movers (from 85.9 to 86.0 percent) alike. Broadly speaking, patterns of matching are similar to those shown in Table 6 and discussed previously. As such, our findings with respect to imputation status suggest that the IRS migration methodology and matching strategy outlined and demonstrated here may prove particularly useful in improving responses that would otherwise be proxy, allocated, imputed.

With respect to the growing gaps in aggregate IRS and publicly available ACS data identified in Table 4, it remains unclear to what extent migration mismatches between IRS and ACS records are

responsible. Limitations on the restricted ACS microdata used in this analysis prevent the weighting and aggregation of raw responses and, therefore, the construction of a "corrected" or "hybrid" IRS-ACS migration estimate. Nevertheless, our analysis finds that as many as 14.1 percent of all those reporting no move in the IRS-ACS linked records may have, according to their IRS records, actually moved. As such, accounting for these "false negatives" in the ACS microdata could have a profound impact on our understanding of migration behaviors in the United States. We reiterate, however, that these results are preliminary and our research is ongoing. In particular, these results are somewhat tenuous in that, using our current methodology for matching IRS and ACS microdata, fewer than half of all ACS respondents with a non-missing migration response can be appropriately linked to an IRS migration interval.

Conclusion:

This paper is the first that we are aware of to utilize linked IRS, decennial census, and ACS survey microdata for the study of migration in the United States. The results reported here suggest that selection and survival bias are, by and large, negligible in the combined IRS 1040, 1099, and W2 administrative records. However, when we restrict our analysis to only those records with unique address identifiers and/or link IRS microdata with other data sources using unique person identifiers, issues with respect to representativeness in the data emerge. Nevertheless, by comparing matched 2009 IRS and 2010 Census person records, we were able to discern the direction and magnitude of these sorts of biases. We also find that, given differences in the coverage of individuals in IRS and decennial census data, IRS administrative records may prove an important source of data on populations difficult to reach with conventional Census Bureau and survey operations.

To demonstrate this point, we construct a longitudinal migration dataset from the universe of individuals identified in IRS 1040s, 1099s, and W2s in tax years 2000 through 2015, then link records for ACS respondents from 2010 through 2013, allowing us to compare migration responses at the individual level. Our results validate the vast majority of all ACS migration responses (85.9 percent of non-movers and 57.0 percent of movers), but suggest that IRS migration data may be particularly efficacious in gauging migration among ACS respondents whose migration status would otherwise be proxy, allocated, or imputed. We stress, however, that these results are preliminary and that our research is ongoing. In future work, we hope to expand the scope of our IRS-ACS linkage to include the entirety of the ACS responses from 2005 to present, as well as to hone our methodology to increase the share of IRS-ACS linkages for which a valid IRS migration interval exists.

We also stress that inconsistencies in IRS and ACS migration measures do not necessarily suggest that one data source is inherently better than the other at capturing migration. Rather, these inconsistencies, to a large extent, reflect fundamental differences in how households and migration events

are defined. ACS respondents are asked to reference their place of usual residence when answering migration related questions, while IRS addresses more closely delineate ties of financial responsibility. These important differences help explain, among other things, the relatively high rates of inconsistency in migration measures for those of college age (18-24). An out-of-state college student may be considered a non-migrant in the ACS because her parents report their home as her place of usual residence, but be counted as a migrant in the IRS records because her work study earnings suggested a move from home to school. There is no "correct" migration measure in such a scenario, though policy makers or researchers may judge one or the other more suitable for answering specific questions. As such, we prefer to think of migration measures drawn from the IRS administrative records as an important complement to Census Bureau survey measures.

Nevertheless, given minimal selection bias into and quantifiable survival bias out of the IRS administrative records, the longitudinal migration dataset constructed, vetted, and demonstrated in this study holds great potential for broad application in migration research. Indeed, the consistency, coverage, and longevity of the IRS administrative record takes a step in the direction of what Franklin and Plane (2006: 233) refer to as the "ideal ultimate migration database."

We highlight three important research directions in which the IRS migration dataset we describe here may be particularly useful, directions which we plan to pursue in the near future. First, these data facilitate innovative methodological approaches to studies of neighborhood change, particularly with respect to segregation and gentrification. Though a handful of nationally representative panel surveys allow insight into the interactions between neighborhood change and individual/household mobility decisions, the IRS administrative record provides a near population level longitudinal look at these processes and facilitates an analysis of how these processes work in different places and at various geographic scales. Second, the IRS record provides largely unprecedented insight into several difficult to measure populations, including the non-citizen foreign-born, U.S. citizens living (and paying taxes) abroad, and millions of individuals from various sociodemographic strata who, for whatever reason, are not enumerated in conventional Census Bureau surveys. Understanding the mobility patterns of these typically overlooked populations is important, particularly as the U.S. population continues to diversify. Finally, though researchers have recently used IRS 1040 records to shed light on the relationships between geography and earnings for individuals over the life course (Chetty et al 2014; Chetty and Hendren Forthcoming), supplementing the 1040s with information from 1099s and W2s holds the potential to answer long-standing questions in the migration literature. For example, though it is a truism among migration researchers that the economic motive is dominant (Ravenstein 1885), such that individuals move if and when it is in their economic interest to do so, it remains relatively unclear whether this truism holds true for all individuals constituting larger households. Individual earnings

reports (W2s), coupled with aggregate household earnings (1040s) and linked longitudinally, promise to provide unprecedented insight into the intra-household dynamics of migration returns.

References:

Akee, Randall, Maggie R. Jones, and Sonya R. Porter. 2017. "Adding Insult to Injury: Racial Disparity in an Era of Increasing Income Inequality." CARRA Working Paper #2017-01. Available from https://census.gov/library/working-papers/2017/adrm/carra-wp-2017-01.html.

Allison, Paul D. 2014. Event History and Survival Analysis: Regression for Longitudinal Event Data. Los Angeles: SAGE Publications.

Bond, Brittany, J. David Brown, Adela Luque, and Amy O'Hara. 2014. "The Nature of the Bias When Studying Only Linkable Person Records: Evidence from the American Community Survey." CARRA Working Paper #2014-08. Available from https://census.gov/library/working-papers/2014/adrm/carra-wp-2014-08.html.

Chetty, Raj and Nathaniel Hendren. Forthcoming. "The Effects of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects." *Quarterly Journal of Economics.*

Chetty, Raj and Nathaniel Hendren. Forthcoming. "The Effects of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." *Quarterly Journal of Economics.*

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *The Quarterly Journal of Economics* 129(4): 1553-1623.

Conway, Karen Smith and Jonathan C. Rork. 2016. "How has Elderly Migration Changed in the 21st Century? What the Data Can and Can't Tell Us." *Demography* 53(4): 1011-1025.

Franklin, Rachel S. and David A. Plane. 2006. "Pandora's Box: The Potential and Peril of Migration Data from the American Community Survey." *International Regional Science Review* 29(3): 231-246.

Johnson, David S., Catherine Massey, and Amy O'Hara. 2015. "The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility." *The Annals of the American Academy of Political and Social Science* 657(1): 247-64.

Ramakrishnan, Karthick, and Farah Z. Ahmad. 2014. "State of Asian Americans and Pacific Islander Series: A Multifaceted Portrait of a Growing Population." *Center for American Progress Technical Report*.

Rastogi, Sonya R., and Amy O'Hara. 2012. "2010 Census Match Study Report." 2010 Census Planning Memo No. 247. Available from https://www.census.gov/library/publications/2012/dec/2010_cpex_247.html.

Ravenstein, Ernst G. 1885. "The Laws of Migration." *Journal of the Statistical Society of London* 48(2): 167-235.

Raymer, James and Andrei Rogers. 2007. "The American Community Survey's Interstate Migration Data: Strategies for Smoothing Irregular Age Patterns." University of Colorado-Boulder Population Program Working Paper #POP2007-08. Available from http://www.colorado.edu/ibs/pubs/pop/pop2007-0008.pdf

Rogers, Andrei, Bryan Jones, and Wanran Ma. 2008. "Repairing the Migration Data Reported by the American Community Survey." University of Colorado-Boulder Population Program Working Paper #POP2008-01. Available from http://www.colorado.edu/IBS/pubs/pop/pop2008-0001.pdf.

Wagner, Deborah and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) record linkage software." CARRA Working Paper Series #2014-01. Washington, DC. Available from https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html.

Figures and Tables:

Figure 1: Longitudinal Data Structure Linking Individual Records from the Internal Revenue Service, Decennial Census, American Community Survey, and Social Security Administration

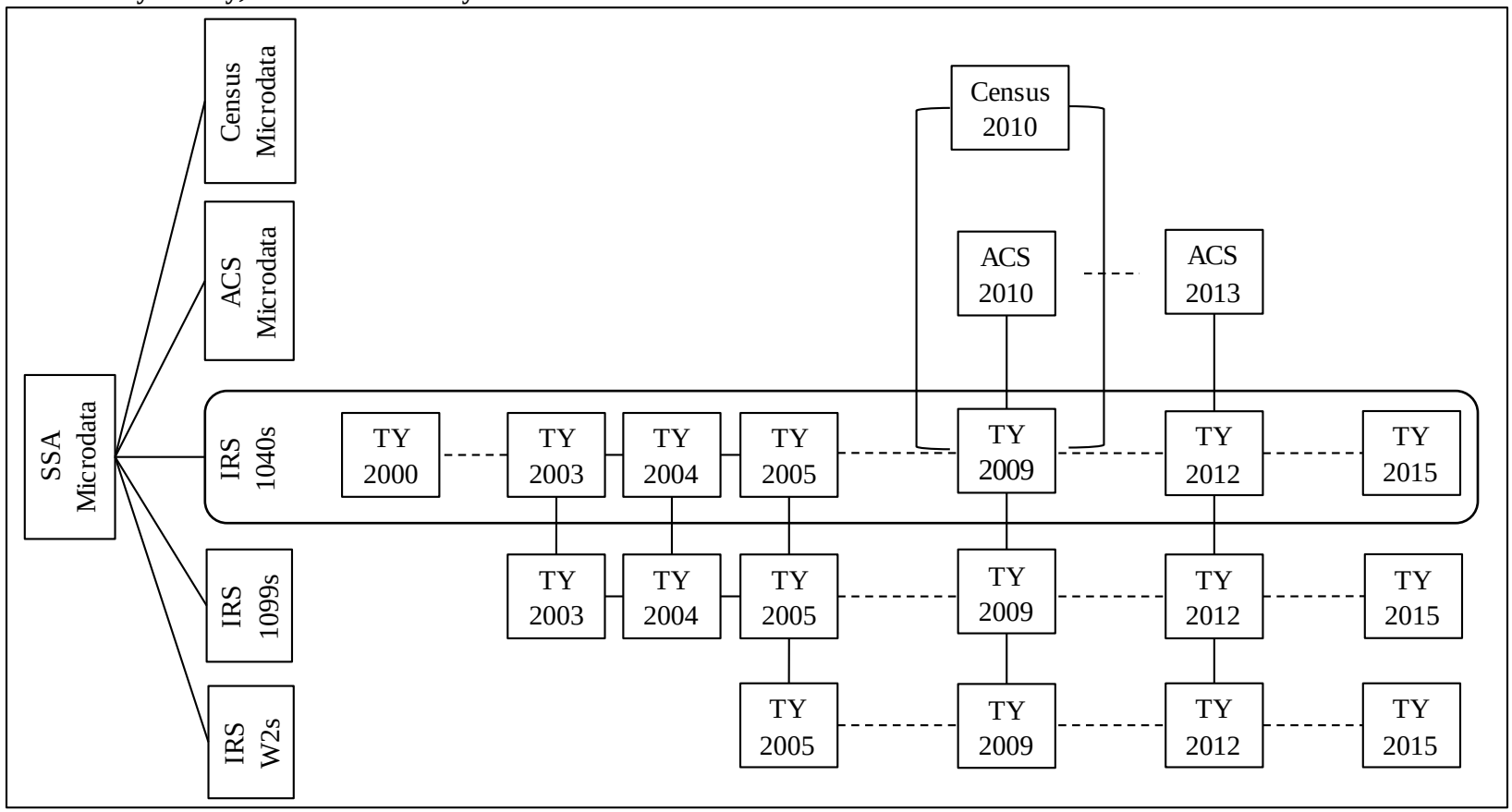

Table 1: Selection Bias in Internal Revenue Service Administrative Records: Coverage of 2010 Census Records by the 2009 IRS Administrative Records and Sociodemographic Distribution of 2010 Census, 2009 IRS, and Linked Census-IRS Records

	Census Records (a)	Percent (b)	IRS Records (c)	Percent (d)	Linked Census-IRS (e)	Percent (f)	Linked Census-IRS with Address IDs (g)	Percent (h)	Share of Census Records Linked to IRS (i)	Share of Census Records Linked to IRS with Address IDs (j)
Total	276,700	100.0	304,500	100.0	261,100	100.0	244,900	100.0	94.4	88.5
Age										
0-2	10,820	3.9	11,590	3.8	8,943	3.4	8,214	3.4	82.7	75.9
3-4	7,394	2.7	7,846	2.6	6,704	2.6	6,181	2.5	90.7	83.6
5-17	48,650	17.6	51,690	17.0	45,300	17.4	41,920	17.1	93.1	86.2
18-24	25,990	9.4	30,470	10.0	24,830	9.5	23,590	9.6	95.5	90.8
25-44	70,860	25.6	79,210	26.0	66,970	25.7	63,640	26.0	94.5	89.8
45-64	75,280	27.2	80,420	26.4	71,750	27.5	67,220	27.5	95.3	89.3
65+	37,670	13.6	43,320	14.2	36,630	14.0	34,130	13.9	97.3	90.6
Hispanic origin										
Hispanic	39,040	14.1	N/A	N/A	34,640	11.4	32,130	13.1	88.7	82.3
Non-Hispanic	237,600	85.9	N/A	N/A	226,500	74.4	212,800	86.9	95.3	89.5
Race										
White alone	205,100	74.1	N/A	N/A	196,200	75.2	184,800	75.5	95.7	90.1
Black alone	34,200	12.4	N/A	N/A	31,240	12.0	28,930	11.8	91.4	84.6
AMIN alone	2,484	0.9	N/A	N/A	2,245	0.9	1,708	0.7	90.4	68.8
Asian alone	12,940	4.7	N/A	N/A	12,000	4.6	11,540	4.7	92.7	89.2
NHPI alone	448	0.2	N/A	N/A	410	0.2	370	0.2	91.6	82.7
SOR alone	13,500	4.9	N/A	N/A	11,660	4.5	10,770	4.4	86.4	79.8
Multiple-race	7,985	2.9	N/A	N/A	7,334	2.8	6,791	2.8	91.9	85.1

Continued on next page.

	Census Records (a)	Percent (b)	IRS Records (c)	Percent (d)	Linked Census-IRS (e)	Percent (f)	Linked Census-IRS with Address IDs (g)	Percent (h)	Share of Census Records Linked to IRS (i)	Share of Census Records Linked to IRS with Address IDs (j)
Sex										
Male	134,800	48.7	150,200	49.3	126,700	48.5	118,800	48.5	94.0	88.1
Female	141,800	51.3	154,300	50.7	134,400	51.5	126,100	51.5	94.8	88.9
Region										
Northeast	50,200	18.1	50,920	16.7	47,640	18.2	44,770	18.3	94.9	89.2
Midwest	62,110	22.5	61,630	20.2	59,290	22.7	56,570	23.1	95.5	91.1
South	101,800	36.8	101,000	33.2	96,010	36.8	89,690	36.6	94.3	88.1
West	62,520	22.6	62,320	20.5	58,190	22.3	53,870	22.0	93.1	86.2
Missing	N/A	N/A	28,670	9.4	N/A	N/A	N/A	N/A	N/A	N/A
Nativity										
Native-Born	245,300	88.7	268,300	88.1	232,200	88.9	217,400	88.8	94.7	88.7
Foreign-Born	31,390	11.3	36,210	11.9	28,910	11.1	27,460	11.2	92.1	87.5

Sources: 2010 Census records and IRS 1040, 1099, and W2 administrative records from tax year 2009 linked to SSA records.
Note: Per U.S. Census Bureau disclosure avoidance protocol, all counts are shown in 1,000s of records and rounded to four significant digits.
Note: "AMIN" = American Indian; "NHPI" = Native Hawaiian/Pacific Islander; "SOR" = Some Other Race.
Column Notes:

a) Records from the 2010 Census assigned unique, anonymous person identifiers. All records have address identifiers.
b) Column Percentages.
c) 1040, 1099, and W2 administrative records assigned and unduplicated by unique, anonymous person identifiers. Not all IRS records were assigned an address identifier.
d) Column Percentages.
e) Matched 2010 Census and IRS (1040, 1099, and W2) records (by person identifiers).
f) Column Percentages.
g) Matched 2010 Census and IRS records (by person identifiers) with an address identifier. Note that address identifiers are only found on 1040 and 1099 administrative records.
h) Column Percentages.
i) Row Percentages (i=e/a). Specifically, this column shows the share of 2010 Census records "covered" by IRS (1040, 1099, and W2) administrative records.
j) Row Percentages (j=g/a). Specifically, this column shows the share of 2010 Census records "covered" by IRS (1040 and 1099) administrative records with address identifiers.

Table 2: Survival Trends for the Population of Individuals in Internal Revenue Service Administrative Records, 2000 to 2015

	Percent of Sample Surviving to…															
	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Sample																
2000	100.0	95.1	93.0	96.5	95.8	94.9	94.3	94.0	92.9	92.0	91.3	90.7	89.8	89.2	88.4	87.7
2001		100.0	94.9	97.3	96.5	95.5	95.0	94.7	93.6	92.7	92.0	91.3	90.5	89.9	89.1	88.4
2002			100.0	98.3	97.3	96.3	95.7	95.4	94.4	93.4	92.7	92.1	91.2	90.6	89.9	89.2
2003				100.0	97.0	95.2	94.2	93.5	92.0	90.9	90.0	89.2	88.2	87.3	86.4	85.7
2004					100.0	96.6	95.4	94.6	93.1	91.9	91.0	90.1	89.1	88.3	87.3	86.5
2005						100.0	97.1	95.9	94.4	93.1	92.2	91.2	90.2	89.4	88.4	87.6
2006							100.0	97.2	95.4	94.1	93.1	92.1	91.0	90.2	89.2	88.4
2007								100.0	96.5	94.9	93.8	92.8	91.7	90.8	89.9	89.1
2008									100.0	96.7	95.3	94.1	93.0	92.1	91.1	90.3
2009										100.0	96.9	95.5	94.2	93.2	92.1	91.3
2010											100.0	96.9	95.3	94.2	93.1	92.3
2011												100.0	96.8	95.4	94.2	93.3
2012													100.0	97.0	95.6	94.5
2013														100.0	97.0	95.7
2014															100.0	97.2

Source: One-percent sample of IRS 1040, 1099, and W2 administrative records from tax years 2000 through 2015 linked to SSA records.

Table 3: Cox Proportional Hazard Model Results Predicting Death or Dropout in the Population of Individuals Found in Tax Year 2000 Internal Revenue Service Administrative Records between 2000 and 2015

	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Age (years)	0.035	0.000	36510.0	<.001	1.035
Male (ref)					
Female	-0.171	0.007	634.2	<.001	0.843
White Alone (ref)					
Black Alone	0.405	0.011	1381.0	<.001	1.499
AMIN Alone	0.543	0.034	247.5	<.001	1.720
Asian Alone	-0.127	0.024	27.5	<.001	0.881
NHPI Alone	0.265	0.099	7.2	0.007	1.303
SOR Alone	0.273	0.023	140.9	<.001	1.314
Multiple Race	0.409	0.024	286.0	<.001	1.506
Non-Hispanic (ref)					
Hispanic	0.194	0.015	175.9	<.001	1.214
Foreign Born (ref)					
Native Born	0.330	0.014	554.5	<.001	1.391
West (ref)					
Northeast	-0.093	0.011	70.6	<.001	0.911
Midwest	-0.058	0.010	30.7	<.001	0.944
South	0.040	0.009	18.0	<.001	1.041
N	1,839				
N (Failures)	86.98				
N (Right Censored)	1,752				
Likelihood Ratio	39,820				
AIC (Null)	2,507,000				
AIC (w/ Covariates)	2,467,000				

Source: One-percent sample of IRS 1040, 1099, and W2 administrative records from tax years 2000 through 2015 linked to SSA and 2010 Census records.

Note: Per U.S. Census Bureau disclosure avoidance protocol, all N (shown in 1,000s of observations), Goodness of Fit, Parameter, SE, Chi-Square, p, and Hazard Ratio values are rounded to 4 significant digits.

Note: "AMIN" = American Indian; "NHPI" = Native Hawaiian/Pacific Islander; "SOR" = Some Other Race.

Table 4: Comparing Aggregate Interstate Migration in Internal Revenue Service Administrative Records and American Community Survey Data, 2010 through 2013

	Retrospective Domestic Interstate Migration in IRS Records			Retrospective Domestic Interstate Migration in Public ACS Estimates					
Year	Movers	Non-Movers	Rate	Movers	MoE	Non-Movers	MoE	Rate	MoE
2010	6,682	248,557	2.62	6,743	69.39	297,135	321.2	2.27	0.02
2011	7,333	249,060	2.86	6,987	77.86	299,087	345.1	2.34	0.03
2012	7,628	255,173	2.90	7,070	58.60	301,309	308.4	2.35	0.02
2013	8,871	254,927	3.36	7,207	70.98	303,322	280.7	2.38	0.02

Sources: IRS 1040 and 1099 administrative records from tax years 2008 through 2012 linked to SSA records; Public ACS 1-Year Estimates (Table B07007) from 2010 through 2013.

Note: All counts are shown in 1,000s of records.

Note: Per U.S. Census Bureau disclosure avoidance protocol, all IRS counts are rounded to four significant digits.

Note: The denominator in the IRS migration rate is the sum of movers and nonmovers, and does not include records from y2 that could not be found in y1.

Note: Counts and rates for IRS records are appropriately adjusted to allow direct comparisons between IRS and ACS records for a given year.

Note: ACS non-movers do not include those moving from abroad.

Note: ACS records are available at https://factfinder.census.gov/. Margins of Error in ACS estimates reflect a 90 percent confidence interval. See https://www.census.gov/programs-surveys/acs/guidance/training-presentations/acs-moe.html for detailed instructions on calculating Margins of Error for sums and ratios of tabulated ACS estimates.

Table 5: Comparing Individual Migration in American Community Survey Records and Internal Revenue Service Administrative Records: Shares of ACS Records Matched to IRS Records with Complete Migration Information, 2010 through 2013

Year	ACS Records	Linked IRS Records	Share of ACS Records Linked	Linked with Valid Migration Measure	Share of ACS Records Linked with Valid IRS Migration Measure
2010	4,164	4,049	97.22	1,674	40.20
2011	4,529	4,404	97.25	1,902	41.99
2012	5,077	4,938	97.26	2,326	45.81
2013	4,785	4,654	97.27	2,077	43.42

Sources: Linked IRS 1040 and 1099 administrative records (2008-2013), Restricted-Use 1-Year ACS Microdata Records (2010 through 2013), and SSA records.

Note: Per U.S. Census Bureau disclosure avoidance protocol, all counts are shown in 1,000s of records and rounded to four significant digits.

Note: ACS records column counts only those with non-missing, migration variable.

Note: Only those ACS records with migration originating and ending in the 50 states are retained.

Table 6: Comparing Individual Migration in American Community Survey Records and Internal Revenue Service Administrative Records: IRS-ACS Migration Matches, False Positives, and False Negatives by Various Sociodemographic Characteristics, 2010 through 2013

	Moved in ACS					Did not Move in ACS				
	Matched	Percent	False Positives	Percent	Total	Matched	Percent	False Negatives	Percent	Total
Total	643.7	57.0	485.8	43.0	1,130.0	5,886.0	85.9	964.2	14.1	6,850.0
Age										
0-2	17.0	69.2	7.6	30.8	24.6	51.0	77.8	14.5	22.2	65.6
3-4	24.6	68.4	11.4	31.6	35.9	118.8	79.4	30.9	20.7	149.7
5-17	101.2	64.0	56.9	36.0	158.1	964.1	83.4	191.5	16.6	1,156.0
18-24	113.6	43.4	148.1	56.6	261.6	429.3	78.9	114.9	21.1	544.2
25-44	248.1	65.0	133.4	35.0	381.5	1,246.0	82.7	261.2	17.3	1,507.0
45-64	100.5	57.2	75.2	42.8	175.7	1,870.0	89.7	214.4	10.3	2,085.0
65+	38.7	42.0	53.4	58.0	92.2	1,206.0	89.8	136.7	10.2	1,343.0
Race										
White alone	473.5	56.4	366.2	43.6	839.7	4,610.0	87.4	662.0	12.6	5,272.0
Black alone	88.5	59.4	60.4	40.6	148.9	626.5	78.6	170.5	21.4	797.0
AMIN alone	7.3	57.6	5.4	42.4	12.7	72.2	77.2	21.4	22.9	93.5
Asian alone	26.5	56.0	20.8	44.0	47.3	234.3	87.2	34.3	12.8	268.5
NHPI alone	1.4	61.7	0.9	38.3	2.3	8.8	77.5	2.6	22.5	11.3
SOR alone	21.2	58.0	15.3	42.0	36.5	192.4	82.0	42.1	18.0	234.5
Multiple-race	25.3	60.1	16.9	40.0	42.2	141.8	81.9	31.4	18.1	173.2
Hispanic origin										
Non-Hispanic	108.1	61.5	67.7	38.5	175.8	389.2	75.4	127.2	24.6	516.4
Hispanic	535.6	56.2	418.1	43.8	953.7	5,497.0	86.8	836.9	13.2	6,334.0

Continued on next page.

	Moved in ACS					Did not Move in ACS				
	Matched	Percent	False Positives	Percent	Total	Matched	Percent	False Negatives	Percent	Total
Sex										
Female	328.4	57.4	243.7	42.6	572.1	3,054.0	86.1	492.1	13.9	3,546.0
Male	315.3	56.6	242.1	43.4	557.4	2,832.0	85.7	472.1	14.3	3,304.0
Region										
Northeast	81.9	51.7	76.4	48.3	158.2	1,098.0	88.6	141.7	11.4	1,240.0
Midwest	157.4	55.7	125.1	44.3	282.5	1,498.0	87.8	208.8	12.2	1,707.0
South	227.3	58.4	162.2	41.7	389.5	1,934.0	85.3	333.7	14.7	2,267.0
West	177.2	59.2	122.1	40.8	299.3	1,355.0	82.9	279.9	17.1	1,635.0
ACS Imputation Status										
None	605.9	58.3	432.9	41.7	1,039.0	5,637.0	86.0	916.9	14.0	6,554.0
Proxy	9.7	48.2	10.4	51.8	20.0	114.9	87.6	16.3	12.4	131.2
Assigned	20.0	48.6	21.1	51.4	41.1	-	N/A	-	N/A	-
Assigned from Allocated	1.5	31.7	3.3	68.4	4.9	-	N/A	-	N/A	-
Hot Deck Imputed	6.6	26.8	18.1	73.2	24.8	133.7	81.2	31.0	18.8	164.7

Sources: Linked IRS 1040 and 1099 administrative records (2008-2013), Restricted-Use 1-Year ACS microdata (2010 through 2013), and SSA records.
Note: Per U.S. Census Bureau disclosure avoidance protocol, all counts are shown in 1,000s of records and rounded to four significant digits.
Note: Only linked cases from Table 5 with valid ACS and IRS migration values are summarized here.
Note: "AMIN" = American Indian; "NHPI" = Native Hawaiian/Pacific Islander; "SOR" = Some Other Race.

Table 7: Comparing Individual Migration in Non-Proxy, Non-Allocated, and Non-Imputed American Community Survey Records and Internal Revenue Service Administrative Records: IRS-ACS Migration Matches, False Positives, and False Negatives by Various Sociodemographic Characteristics, 2010 through 2013

	Moved in ACS					Did not Move in ACS				
	Matched	Percent	False Positives	Percent	Total	Matched	Percent	False Negatives	Percent	Total
Total	605.9	58.3	432.9	41.7	1,039.0	5,637.0	86.0	916.9	14.0	6,554.0
Age										
0-2	15.6	69.9	6.7	30.1	22.3	47.4	78.1	13.3	21.9	60.7
3-4	22.9	69.1	10.3	30.9	33.2	112.2	79.5	29.0	20.5	141.1
5-17	94.7	65.3	50.4	34.7	145.0	918.7	83.5	181.6	16.5	1,100.0
18-24	108.5	43.9	138.4	56.1	246.9	410.7	79.0	108.9	21.0	519.6
25-44	237.3	65.9	122.9	34.1	360.2	1,210.0	82.9	250.5	17.2	1,460.0
45-64	92.4	60.0	61.6	40.0	154.0	1,802.0	89.8	204.6	10.2	2,006.0
65+	34.4	44.7	42.6	55.3	77.0	1,137.0	89.8	129.0	10.2	1,266.0
Race										
White alone	448.6	57.7	329.4	43.6	778.0	4,425.0	87.5	631.5	43.6	5,057.0
Black alone	81.2	61.0	51.9	40.6	133.0	592.6	78.7	160.6	40.6	753.3
AMIN alone	6.9	58.6	4.9	42.4	11.8	69.7	77.1	20.7	42.4	90.4
Asian alone	24.4	58.2	17.5	44.0	42.0	221.9	87.4	31.9	44.0	253.8
NHPI alone	1.3	62.8	0.8	38.3	2.1	8.3	77.4	2.4	38.3	10.7
SOR alone	19.7	59.7	13.3	42.0	32.9	183.6	82.1	40.0	42.0	223.5
Multiple-race	23.8	61.1	15.2	40.0	39.0	135.7	82.0	29.8	40.0	165.4
Hispanic origin										
Non-Hispanic	101.6	62.2	61.6	38.5	163.3	366.4	75.4	119.7	38.5	486.2
Hispanic	504.2	57.6	371.2	43.8	875.5	5,271.0	86.9	797.1	43.8	6,068.0

Continued on next page.

	Moved in ACS					Did not Move in ACS				
	Matched	Percent	False Positives	Percent	Total	Matched	Percent	False Negatives	Percent	Total
Sex										
Female	309.0	58.8	243.7	42.6	572.1	3,054.0	86.2	492.1	42.6	3,546.0
Male	296.8	57.9	242.1	43.4	557.4	2,832.0	85.8	472.1	43.4	3,304.0
Region										
Northeast	76.7	53.3	76.4	48.3	158.2	1,098.0	88.6	141.7	48.3	1,240.0
Midwest	148.6	56.8	125.1	44.3	282.5	1,498.0	87.8	208.8	44.3	1,707.0
South	214.0	59.8	162.2	41.7	389.5	1,934.0	85.4	333.7	41.7	2,267.0
West	166.5	60.4	122.1	40.8	299.3	1,355.0	83.0	279.9	40.8	1,635.0

Sources: Linked IRS 1040 and 1099 administrative records (2008-2013), Restricted-Use 1-Year ACS microdata (2010 through 2013), and SSA records.
Note: Per U.S. Census Bureau disclosure avoidance protocol, all counts are shown in 1,000s of records and rounded to four significant digits.
Note: Only linked cases from Table 5 with valid, non-imputed, non-allocated, non-proxy ACS and IRS migration values are included in this table.
Note: "AMIN" = American Indian; "NHPI" = Native Hawaiian/Pacific Islander; "SOR" = Some Other Race.