Using Tax Data on Filers and Non-Filers to Examine Trends in Employment and Earnings in the Low-Income Population

John Green
Johns Hopkins University

Robert Moffitt
Johns Hopkins University

Daniel Rodriguez
Internal Revenue Service

April 15, 2026

No financial support has been received for this project. The authors have no professional or personal conflicts of interest related to this research.

Using Tax Data on Filers and Non-Filers to Examine Trends in Employment and Earnings in the Low-Income Population

Overview

This project was awarded by the IRS JSRP program, with Moffitt submitting the awarded proposal. He enlisted graduate student John Green and IRS staffer Daniel Rodriguez to join.

The project is using tax data on filers and non-filers from 1999 to 2024 to estimate trends in employment and earnings in the low-income population over that period. We are focusing on the prime-age group 25-55. While the vast majority of work on this topic has used the CPS, and a smaller body of work has used the LEHD, this project will use tax data for this purpose. Tax data have been used to examine trends in income inequality by Piketty et al. (2018) and Auten and Splinter (2024), but those authors, while imputing many income sources to the low-income population, were more focused on the top end of the distribution. In addition, this project will not attempt to estimate family level employment and income, because that would require forming families, or at least filing units, for non-filers. Piketty et al. and Auten-Splinter did so in an ad hoc way, but a better attempt was made by Larrimore et al. (2021). But our project is solely focused on individual employment and earnings.

Despite the prima facie argument for the accuracy of tax data relative to the CPS, using tax data for this purpose faces a number of challenges. The major issue is that, for a study of the low-income population, it is important to estimate the number of those with zero earnings and who are hence not employed. While W-2 data can be used to detect the presence of wage and salary earnings of non-filers (for filers, earnings are reported on 1040s), off-the-books earnings are not reported. And it is well known in the literature that self-employment earnings are much worse reported to the IRS than wage and salary income, and many self-employed do not submit 1099s even if their income is high enough to be above the reporting threshold. Another well-known issue is that 1040s are not always accurate as well, and income is often underreported, as audit studies have shown. There has been work studying the special case of Schedule C income, discussing this issue.

This project's primary aim is to focus on these issues with tax data and to hopefully make progress in establishing accurate employment and earnings data for the low income population with such data. Comparisons to CPS trends will be made.

This project will not be complete by the time of the conference and should only be considered for conference presentations of in-progress work and plans for the future. An estimate of the progress that is likely to be made by the conference date is given below.

Details[1]

The first step is to construct a universe of individuals residing in the U.S. and, to the extent possible, the civilian non-institutional population (at minimum, this is necessary to improve comparability with the CPS, which reports on that population). To this end we have used taxpayer data to identify the individuals of interest, allowing us to use the panel nature of the data in some respects.

We begin by removing anyone who is not alive at any point in 1999-2024. However, moving from this sample to the resident population presents challenges. First, we examine state of residency from the filings (such as the 1040) and information returns (such as the W2 or the 1099-INT). In cases where an individual may have a different state across different records within the same year, we consider the mode to be their true residency. We exclude individual/year combinations with a foreign address or who are living in a territory.

But we also face the issue of what to do with a SSN/year combination where the individual is still living but has no filings or information returns in the tax data. It is possible that these individuals are residents who are non-workers and have no other reason to interact with the IRS, but in many cases, they may no longer be residents. For example, exchange students who come to the US to study may receive an SSN and then return to their home country, at which point we would like to exclude them from our sample. This may be a sizable group. The longitudinal nature of the data allows us to determine if these individuals have an interaction (filing or information return) with the IRS in adjacent years. We consider "donuts" of both 2 and 4 years. This means, for example, that someone with no interactions with the IRS in 2010 may still be considered a resident if they have some interactions in 2009 and 2011, or if they have interactions in either 2008 or 2009 *and* 2010 or 2011. Under the assumption that an individual was a confirmed US resident in a nearby previous year as well as a residency indicate in a nearby later year, we can assume that they are likely to also be resident in the "missing" year t.

To restrict our sample to the civilian noninstitutional population, we are currently working to identify those in the military, where there are several possibilities for identifying them. For the noninstitutional population, we have information on incarceration status for the full universe of residents, but only for a subset of our sample. We plan to impute incarceration status in the other years, and will ultimately exclude the incarcerated from our analysis. Our top age limit of age 55 should exclude most of the older population in nursing homes and other institutions for the aged.

We have compared our counts for 2010 to those in the 2010 Decennial Census. Overall, our procedure allows us to account for about 94% of the resident population reported in the DC using a 2-year donut, and about 96% of the resident population using a 4-year donut. But the DC includes the military and the institutionalized, so we plan to compare to the CPS next.

We will next move on to correct misreporting among filers. For this we will use the long-standing audit program at the IRS, which began with the Taxpayer Compliance Measurement

[1] Due to applicable IRS disclosure guidelines and requirements, certain details have been omitted and are not discussed in this document

Program in the 1980s and transitioned to the National Research Program (NRP) in the 1990s (Slemrod, 2019). The NRP data have information on the results of audits of returns and estimates of unreported income, which is particularly important because estimates of the tax gap—the gap between reported income and total income, including unreported income—is primarily the result of unreported income for filers, not from failure to file (Slemrod, 2007), although the ratio of unreported income to total income increasing in the level of income (Johns and Slemrod, 2010). The NRP have been used to estimate unreported income in past work (Johns and Slemrod, 2010; Auten and Splinter, 2024). We will follow those methods with a focus on earnings rather than unearned income. Past work as also show, as noted previously, that self-employment income, sole proprietor income, nonfarm proprietor income, and related forms of earned income are particularly underreported. We will consequently conduct separate imputations for these different types.

The IRS changed its sampling algorithm for the NRP after 2017 and, consequently, the NRP data for the later years in our study is not ready for analysis. We will be the first to use the post-2016 data and will produce estimates that adjust for the new sampling algorithm.

We expect this work to have been completed by the time of the conference.

The next exercise is to use a matched CPS-IRS data set for the years 2012-2019 currently available at the IRS to conduct estimates of unreported employment and earnings (Wagner and Lane, 2012; Riddles et al., 2017; the data were matched using the Census-Bureau provided PIK, and we will apply inverse probability weighting to adjust for differences in the matched and unmatched samples on observables). It is well-known that, while CPS have serious issues of measurement error and misclassification of the sources of earnings (e.g., wage and salary vs self-employment), underreporting to the IRS also occurs. As a result, matched studies of survey to administrative data sources show differences both ways—some employment appears in tax data (either 1040, Schedule C, or Schedule E) but some employment appears in surveys that does not appear in tax data (Abowd and Stinson, 2013; Abraham et al., 2013; Haltiwanger et al., 2013; Abraham et al., 2021; Bee et al., 2023; and many others). There are many reasons why these differences can legitimately occur (Abowd and Stinson, 2013; Abraham et al., 2021a; Bee et al., 2023) The matched CPS-IRS data have matches from 2012-2019. We will use these data to apply the measurement-error statistical models of Abowd and Stinson (2013) and Bee et al. (2023), which allow measurement error in two different data sources with a priori weights on each, to generate a single best estimate. We will vary the weights to provide a range of estimates to the imputed values to add to the tax data. The estimates will be stratified by wage and salary earnings and self-employment earnings and, within these groups, by age and sex. All analyses will be conducted separately by year and we will impute to tax years 1999-2011 and 2020-2024 with appropriate bounds and sensitivity tests.

Special attention will be paid to Schedule C, which has been a focus of much work, particularly on gig workers (e.g., Abraham et al., 2019, 2021b), which has shown a growing discrepancy between survey and administrative data reports of self-employment over time. Schedule C data are available on the CPS-IRS matched file and we will use prior work on such matches to inform our corrections and imputations to the tax data.

We plan to also consider an algorithm developed by the Urban Institute for its TRIM3 microsimulation model to identify undocumented individuals in the CPS. We will implement their method on the linked CPS-IRS file and construct bounds for its impact on our sample.

We expect these exercises (which precede the final step before developing the analysis file to assess trends in employment and earnings in the low-income population and to compare to the CPS) to be in-progress at the time of the conference.

References

Abraham, K., J. Haltiwanger, K. Sandusky, and J. Spletzer. 2009. "Exploring Differences in Employment Between Household and Establishment Data." Journal of Labor Economics 31 (April 2013, Part 2), 2, S129-S172.

Abraham, K. et al. 2019. "The Rise of the Gig Economy: Fact or Fiction?" AEA Papers and Proceedings 109:357-361.

Abraham, K. et al. 2021a. "Reconciling Survey and Administrative Measures of Self-Employment." Journal of Labor Economics 39 (4): 825-860.

Abraham, K. et al. 2021b. "Measuring the Gig Economy: Current Knowledge and Open Issues." In Measuring and Accounting for Innovation in the 21st Century, eds. C. Corrado et al. Chicago: University of Chicago Press.

Abowd, J. and M. 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.

Auten, G. and D. Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." Journal of Political Economy 132(7): 2179-2227.

Bee, A. et al. 2023. "National Experimental Wellbeing Statistics." Bureau of the Census.

Haltiwanger, J.; H. Hyatt; and J. Spletzer. 2013. "Increasing Earnings Inequality: Reconciling Evidence from Survey and Administrative Data." Journal of Labor Economics 41(S1): S61-S93.

Johns, A. and J.Slemrod. 2010. "The Distribution of Income Tax Noncompliance." National Tax Journal 63(3): 397-418.

Larrimore, J.; J. Mortensen; and D.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.

Piketty, T.; E. Saez; and G.Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." Quarterly Journal of Economics 133(2): 553-609.

Riddles, M. et al. 2017. “Handling Respondent Rounding of Wages Using the IRS and CPS Matched Dataset.” IRS Report.

Slemrod, J. 2007. “Cheating Ourselves: The Economics of Tax Evasion.” Journal of Economic Perspectives 21(1): 25-48.

Slemrod, J. 2019. “Tax Compliance and Enforcement.” Journal of Economic Literature 57(4): 904-954.

Wagner, D. and T. Lane. 2012. “Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications Research Records Software.” Bureau of the Census.