ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

After 40 Years, How Representative Are Labor Market Outcomes in the NLSY79?

Authors	Alexander Bick, Adam Blandin, and Richard Rogerson
Working Paper Number	2024-010B
Revision Date	April 2024
Citable Link	https://doi.org/10.20955/wp.2024.010
Suggested Citation	Bick, A., Blandin, A., Rogerson, R., 2024; After 40 Years, How Representative Are Labor Market Outcomes in the NLSY79?, Federal Reserve Bank of St. Louis Working Paper 2024-010. URL https://doi.org/10.20955/wp.2024.010

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

After 40 Years, How Representative Are Labor Market Outcomes in the NLSY79?[*]

Alexander Bick
Federal Reserve Bank of St. Louis, CEPR

Adam Blandin
Vanderbilt University

Richard Rogerson
Princeton University, NBER

April 16, 2024

Abstract

In 1979, the National Longitudinal Study of Youth 1979 (NLSY79) began following a group of US residents born between 1957 and 1964. It has continued to re-interview these same individuals for more than four decades. Despite this long sampling period, attrition remains modest. This paper shows that after 40 years of data collection, the remaining NLYS79 sample continues to be broadly representative of their national cohorts with regard to key labor market outcomes. For NLSY79 age cohorts, life-cycle profiles of employment, hours worked, and earnings are comparable to those in the Current Population Survey. Moreover, average lifetime earnings over the age range 25 to 55 closely align with the same measure in Social Security Administration data. Our results suggest that the NLSY79 can continue to provide useful data for economists and other social scientists studying life-cycle and lifetime labor market outcomes, including earnings inequality.

JEL Codes: E24, J22, J31
Keywords: Lifetime earnings and hours worked, life-cycle earnings and hours worked

[*] Contact: alexander.bick@stls.frb.org, adam.blandin@vanderbilt.edu, rdr@princeton.edu. We thank Kevin Bloodworth II, Elizabeth Harding, and Siyu Shi for research assistance. The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

1 Introduction

The National Longitudinal Study of Youths 1979 (NLSY79) is a long-running panel dataset for the US. It began in 1979 by interviewing a group of US residents aged 14 to 22 (born 1957 to 1964) and has continued to re-interview these same individuals for more than four decades. The NLSY79 collects information on a wide range of topics, including demographics, family structure, labor market outcomes, health, and criminal activity. This rich information, combined with a long panel, have made the NLSY79 a valuable data source for economists and other social scientists, especially those interested in lifetime or life-cycle patterns. For example, between 2010 and 2023 the NLSY79 has been used in at least 34 articles published in the "Top 5" economics journals.[1]

Work on inequality has long understood that it is important to distinguish between transitory and persistent components of inequality. Motivated by this, recent work on inequality has leveraged administrative data to document features of lifetime inequality; see e.g. the work by Guvenen et al. (2022) who use data by the Social Security Administration (SSA). A key advantage of adminstrative data is the large sample size that if offers. But there are also some disadvantages to relying on adminstrative data sets: access to such data is extremely limited, especially in the US, and some variables of interest are typically not present. For example, because the SSA data does not include information on hours worked, it cannot distinguish between inequality in earnings and inequality in wage rates.

The long panel of the NLSY offers the possibility of a publicly available data set that can now be used to study lifetime inequality for a specific set of cohorts.[2] Moreover, because it provides information on both earnings and hours, it can distinguish between inequality in earnings and inequality in wage rates. For example, in Bick et al. (2024) we use the sample constructed in this paper to document the relationship between lifetime hours worked, hourly wages, and earnings.

However, the NLSY has two major disadvantages relative to administrative data like the SSA: the sample size is much smaller and participation is voluntary. Voluntary participation gives rise to two related practical issues: attrition and missing observations. While the initial sample was designed to be nationally representative, non-random attrition over time may have introduced bias into the sample. Further, even individuals who remain in the sample may have missing observations for one or more variables in one or more surveys. Throwing out all individuals with any missing values will severely limit sample size. Maintaining sample size is of particular importance when studying second moment properties of the data like inequality. Our interest in lifetime earnings

[1] We based this on searching for the term "NLSY79" on the webpages of the journals American Economic Review, Econometrica, Journal of Political Economy, and Review of Economic Studies as well as portals providing access to these journals such as JSTOR.

[2] We are not aware of any other US survey with comparable information that covers as many individuals for as long. For example, in Panel Study of Income Dynamics, the sample that can be followed for a similar time period is less than 60% of the corresponding NLSY79 sample.

inequality thus requires out to impute such missings. This is compounded by a particular feature of the NLS79. The earnings measure we rely on, which is comparable to the earnings reported in the Current Population Survey's Annual Social and Economic Supplement and SSA data, is from 1994 onwards only available for odd years.

The goal of this paper is twofold. First, we assess the extent to which demographics and labor market outcomes in the NLSY79 remain nationally representative through the 2020 interview, four decades after it began. Second, we assess the extent to which a modest amount of imputation for missing values facilitates a reasonably large sample size to study inequality in lifetime labor market outcomes for both earnings and hours.

We begin our analysis by constructing several subsamples in the NLSY79. In a preliminary step, we impute missing observations with a simple weighted linear interpolation of nearby observations. We only impute values if we observe at least one direct report within five years of the year of interest. Our first sample, which we refer to as the Cross-Sectional Sample, includes all observations with either a direct report or an imputed value. This sample is not balanced: at age 21 the sample size is 9858, while at age 55 it is 7154. To be included in the Lifetime Sample, we require at least one direct report at age 55 or older and an observation at each possible age 21 to 55 (either a direct report or an imputation). This balanced sample comprises 6330 individuals.

For each of our two samples we can also define subsamples in which we only use only direct reports, i.e., we exclude person-year observations that rely on imputation. Comparing the direct reports subsamples allows us to assess whether the observations with imputations appear to be selected relative to the direct reports. We find that direct report subsamples display very similar properties to the overall sample, suggesting that the observations with imputed values do not feature an important amount of selection relative to the direct reports.

To assess the non-random nature of attrition, we ask whether the NLSY79 is representative at each age over the life-cycle. We first establish that attrition is approximately random along each of several dimensions of observables: gender, race and educational attainment. This still leaves open the possibility that there is selection on unobservables. To assess this we compare life cycle outcomes for employment, hours and earnings in our Cross-Sectional Sample to outcomes for the same birth cohorts in the Current Population Survey's Annual Social and Economic Supplement (ASEC), also often referred to as the March CPS. We find that life-cycle means and standard deviations for these outcomes in all of our samples track those in the ASEC quite closely, though we note our sample does imply somewhat higher values for weekly hours of work and employment rates. We also compute the distribution of average lifetime earnings, defined as the average annual earnings of individuals between ages 25 and 55, and compare it to results in Guvenen et al. (2022) based on Social Security Records. Again, we find a close match.

Based on these comparisons we conclude that, four decades on, labor market outcomes in the

NLSY79 remain broadly representative of individuals in their birth cohort. This is particularly true for measures of cross-sectional and lifetime earnings inequality, which closely align with both the ASEC and Social Security data. Our results suggest that the NLSY79 can continue to provide useful data for economists and social scientists studying life-cycle and lifetime labor market outcomes, including earnings inequality. To facilitate these analyses, the replication package of this paper (to be posted soon) will provide code for constructing our sample and dataset, as well as a basic dataset with imputed weeks worked, hours worked, and earnings for each individual in the NLSY79. These codes and dataset can be easily merged with any existing analysis, providing researchers confidence that their data work is based on a representative sample.

Our paper is closely related to two existing papers that analyze attrition in the NLSY79. MaCurdy et al. (1998) finds that attrition up through the 1991 interview appears to be close to random with respect to labor market characteristics. They also show that labor market outcomes in the NLSY79 align closely with the ASEC up through the 1991 interview.[3] Our paper extends these ASEC comparisons to include 30 additional years, through the 2020 interview. We also compare the distribution of average lifetime earnings in the NLSY79 to Social Security data, which is a novel comparison. More recently, Aughinbaugh et al. (2017) find that attrition through the 2014 interview is close to random, but do not compare outcomes in the NLSY79 to other datasets.[4] Our paper is also similar in spirit to Heathcote et al. (2010) and Heathcote et al. (2023) who compare cross-sectional income, consumption, and wealth inequality across various datasets in the US and also benchmark them to the degree possible with NIPA and flow of funds data.

An outline of the paper follows. Section 2 introduces the NLSY79 and our baseline sample, and discuss how we measure our key variables of interest. Section 3 describes our imputation strategy for missing observations and our sample selection criteria. Section 4 compares demographics and labor market outcomes in the NLSY79 to analogues in the ASEC and Social Security data. Section 5 concludes.

[3] MaCurdy and Timmins (2001) show that attrition does not affect the analysis of wage dymamics for men in the NLSY79 using data through 1991, but has some impact on wage dynamics for women.

[4] Aughinbaugh and Gardecki (2007) document evidence of non-random attrition in NLSY97, but argue that it is difficult to assess the impact of attrition because their analysis only covers the first 8 waves with many sample members not yet or just recently having completed their education.

2 Data

2.1 Basic Sample

This section describes the basic NLSY79 sample that we use for our analysis. The NLSY79 began in 1979 by interviewing 12,686 US residents between the ages of 14 and 22 who were born between 1957 and 1964. The initial sample was designed to be nationally representative of the noninstitutionalized civilian population corresponding to those birth cohorts and also included several supplemental over-samples. The initial respondents were re-interviewed annually each year through 1994, and then every other year since then. The most recent available data come from the 2020 interview wave, for which the interviews were conducted throughout 2020 and the first half of 2021.

Figure 1a summarizes survey attrition between the 1979 and 2020 interview waves. The black line (circles) shows the share of all initial respondents who participated in the survey in a given year. There are two sharp reductions in survey participation. First, after 1984 the vast majority of a supplemental military over-sample were discontinued. Second, after 1990 the econonomically disadvantaged non-Black/non-Hispanic youths supplemental over-sample was discontinued. For the remainder of the paper, we exclude these two discontinued supplements, which leaves us with 9961 respondents.[5] The blue line (diamonds) in Figure 1a shows the share of these respondents who participated in the survey in a given year. Among these respondents, 65.6% participated in the 2020 survey. If one further excludes respondents who had died by 2020, the participation rate in the 2020 survey increases to 74.5%, see the red line (squares). The high retention rate may be attributable to efforts by the NLSY79 to re-interview all initial respondents, even if they have missed several recent surveys.

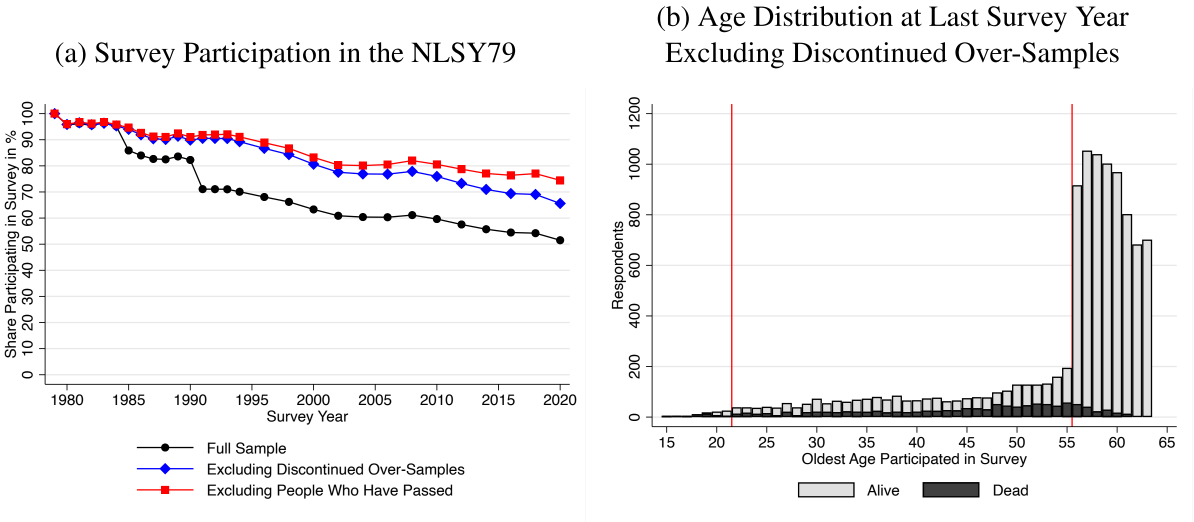
Among the 9961 respondents discussed above, our analysis restricts attention to the 9867 respondents with at least one observation between ages 21 - 55. We will refer to this sample as our "Initial Sample".

2.2 Measurement

Since we are primarily interested in life-cycle facts, our results are typically presented by age. We define age as calendar year minus calendar year of birth.

[5] 201 respondents randomly selected from the military sample remained in the survey, but only 198 of them actually continued to participate. We keep those individuals in our sample.

Figure 1: Survey Participation and Age Distribution



(a) Survey Participation in the NLSY79

(b) Age Distribution at Last Survey Year Excluding Discontinued Over-Samples

Notes: The discontinued over-samples refer to the military over-sample and the econonomically disadvantaged non-Black/non-Hispanic youths supplemental over-sample. The left red vertical lines in indicate the age range at time of the survey for our final sample.

For each individual-age, we construct the number of reported weeks worked and weeks with missing employment status. In each interview, the NLSY79 asks individuals how many jobs they have held since their previous interview. (Even if the most recent interview was several years ago, the survey asks about all jobs since then.) For each of these jobs, the survey records the week-year pair in which the job started and ended. Based on these dates, the NLYS79 creates a weekly employment variable for each individual-year. If an individual had a temporary employment gap during their tenure at a specific employer, this is reflected in the weekly employment arrays. In weeks without employment, an individual's employment status can either be non-employed or unknown if no information was provided for that week.

For each individual-age, we also construct (i) average reported weekly hours worked and (ii) the number of workweeks with missing hours information. In each interview, the NLSY79 collects the usual weekly hours worked by the individual for each of their jobs. (Hence, there is no variation in the report of usual weekly hours worked within a job between two interviews.) Combined with the data on job start and end dates, this provides a weekly array of hours worked. If someone holds multiple jobs simultaneously, the weekly hours measure is the sum of usual hours worked in all jobs. We impose a cap on weekly hours at 98 hours, which corresponds to a 14-hour workday for seven days per week. A given week's hours are missing if either the employment status is missing or if the individual did not provide their hours worked for a particular week worked.

Finally, for each individual-age we rely on the respondent's reported annual earnings for the calendar year prior to an interview.[6] In particular, the NLYS79 collects this information for two

[6] The NLYS79 also collects information to construct the weekly earnings for each job held since the last interview.

different types of earnings: (i) income from wages, salary, commissions, or tips from all jobs before deductions for taxes or anything else last year, and (ii) income received from an own farm/business last year. Going forward, we abbreviate (i) as "income from wages and salary." Both earnings variables are restricted to nonnegative values. Following Guvenen et al. (2022) we deflate earnings with the personal consumption expenditure index normalized to one in 2013. As with the other variables, earnings may be missing even in years in which a respondent worked and took part in the survey.

2.3 Data Cleaning

For a small number of observations we can eliminate missing information on earnings using information on weeks worked, or vice versa. Among person-year observation with zero weeks worked in a year and the employment status known for all weeks of the year, 1.4% miss wage and salary income and 0.2% miss business/farm income. We assign zero for those missing earnings. In the other direction, among person-year observations with an earnings report of zero, 3.0% feature zero weeks worked while missing the employment status missing for some but not all weeks of the year, and for 5.8% the employment status is missing for all weeks of the year. In either case, we assign non-employment for the missing weeks.

Occasionally, a respondent will report positive earnings but zero annual hours worked, with no missing employment or hours information; in our initial sample, this is the case for 1.7% of observations with positive income.[7] Similarly, 5.81% of observations with positive weeks worked report zero earnings. In both cases, we have three options: either a) take the "positive" report at face value, but not the "zero" report and therefore set the "zero" report to missing, b) take the "zero" report at face value, but not the "positive" report and therefore set the "positive" report to zero, or c) set both the "zero" and "positive" reports to missing. Since we do not have any indication which report is more reliable, we proceed with the last option (c). However, the actual choice among those three options is not of much relevance as the final sample size and the key labor market outcomes of interest do not vary much across them.

We use the annual earnings measure because it is more comparable with annual earnings in the ASEC and SSA data. We leave it for future research to analyze the job-level usual earnings measure.

[7] This can be broken down further by the two different income sources: a) 1.6% of observations with positive income from wages and salary but no reported farm/business income, b) 10.3% of observations with no reported income from wages and salary but positive farm/business income, and c) 1.0% of observations with both positive income from wages and salary and positive farm/business income.

3 Imputation of Missings and Sample Selection

As discussed in the previous Section, the NLSY79 contains many missing values, either because an individual did not participate in a given interview or because the individual participated in the interview but did not provide some information. We deal with missing values in one of two ways: either we drop an individual, or we impute a value using information from nearby interviews from that same individual. Loosely, our guiding principle is to drop individuals with long stretches of missings that span several years, but otherwise to impute missings when nearby information is available. The remainder of this section provides further details on the procedures we adopt.

3.1 How We Impute Missings

In each year t and week k, the employment status $e_{i,t,k}$ of individual i is either not employed (0), employed (1) or missing (-1). For each week in which an individual is employed ($e_{i,t,k} = 1$), hours worked $h_{i,t,k}$ are either positive or missing (-1). Earnings in contrast are only available on the annual level $y_{i,t}$ and can be either positive or missing (-1) for years with positive weeks worked. We now describe how we impute missing values for each of these three variables.

3.1.1 Imputing Weeks Worked and Weekly Hours When Some But Not All Weeks Missing

In the discussion below, the term "observation" indicates an individual-year pair. To facilitate the discussion, we introduce the indicator function $\mathbb{I}_{v=x}$, which takes the value one if v equals x, and zero otherwise.

Imputing Weeks Worked When an observation has at least one week with missing employment status, we impute the share of weeks worked among those missing weeks. To do so, we first define individual i's share of weeks worked among non-missing weeks for each year t, $s_{i,t}$:[8]

$$s_{i,t} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1}}{\sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}\neq -1}}, \tag{1}$$

[8] A few years have 53 weeks, in which case we adjust the summations in (1) and all other equations in this section. Note that this share only exists for years with at least one week of non-missing employment status.

Next, define $\bar{s}_{i,t}$, to be a three-observation weighted moving average of $s_{i,t}$, using the most recent prior year $\underline{t}$ and subsequent year $\bar{t}$ with at least one week of non-missing the employment status:

$$\bar{s}_{i,t} = \frac{\sum_{j=\underline{t},t,\bar{t}} q^e_{i,j} \times s_{i,j}}{\sum_{j=\underline{t},t,\bar{t}} q^e_{i,j}}. \tag{2}$$

Here, $q^e_{i,j}$ are weights for each year $j \in \{\underline{t}, t, \bar{t}\}$ used in the moving average, which are given by:

$$q^e_{i,\underline{t}} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,\underline{t},k} \neq -1}}{t - \underline{t}}, \quad q^e_{i,t} = \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k} \neq -1}, \quad q^e_{i,\bar{t}} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,\bar{t},k} \neq -1}}{\bar{t} - t}.$$

In the case where an individual has no missing employment status in either $t-1$ or in $t+1$, then $\underline{t} = t-1$, $\bar{t} = t+1$, and the formulas above reduce to a simple equal weighted three year moving average. Our procedure generalizes this to potentially use observations that are more than one year backward or forward if necessary, and to more heavily weight nearby observations that are likely to be more informative. Specifically, we construct imputation weights such that observations have more weight if they a) are more recent, and b) have fewer weeks with missing employment status.[9]

Finally, we use the moving average $\bar{s}_{i,t}$ to impute weeks worked for year t:

$$\widehat{wks}_{i,t} = \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1} + \text{round}\left(\bar{s}_{i,t} \times \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=-1}\right). \tag{3}$$

We round the second summand in Equation (3) to the nearest integer and set hours worked for those weeks to missing. Note that the second summand is zero for observations with the employment status known for all weeks of the year: in this case there is no need to impute weeks worked.

Imputing Weekly Hours Worked When an observation has one or more weeks worked with missing hours information, we impute hours worked, $\hat{h}_{i,t}$, for those weeks. We use exactly the same procedure as the one above for weeks worked, except applied to weekly hours rather than weeks worked. The detailed formulas are in Appendix A.1.

3.1.2 Imputing Values for Full-Year Missings

Imputing Weeks Worked and Weekly Hours We apply a very similar imputation procedure when the weekly employment status or weekly hours are missing for all weeks (worked) in year t. The only difference is that in this case we cannot use information from year t, since the full year is

[9] If there is no (recent) prior or subsequent year available among the ages 21-55, we set $q^e_{i,\underline{t}} = 0$ or $q^e_{i,\bar{t}} = 0$, respectively.

missing. In particular, we linearly interpolate these full-year missing using the most recent prior $\underline{t}$ and subsequent year $\bar{t}$ determined in the previous section:

$$v_{i,t} = v_{i,\underline{t}} + \frac{v_{i,\bar{t}} - v_{i,\underline{t}}}{\bar{t} - \underline{t}} \times (t - \underline{t}) \; \forall \; v = \{\widehat{wks}, \widehat{h}\}. \tag{4}$$

If there is no (recent) prior or subsequent year available, we set $v_{i,t} = v_{i,\underline{t}}$ or $v_{i,t} = v_{i,\bar{t}}$, respectively.

Imputing Earnings To impute missing earnings, we first construct an hourly wage for each observation with a direct earnings report:

$$w_{i,t} = \frac{y_{i,t}}{\widehat{wks}_{i,t} \times \widehat{h}_{i,t}}. \tag{5}$$

We use wages from Equation (5) to impute an hourly wage for the missing observation using the linear interpolation procedure in (4). We then impute missing earnings by multiplying the imputed wage by annual hours worked:

$$\widehat{y}_{i,t} = \widehat{w}_{i,t} \times \widehat{wks}_{i,t} \times \widehat{h}_{i,t}.$$

Prior to imputing wages, we make two adjustments to reported wages. First, we impose a floor on hourly wages. Specifically, we set the floor equal to half the federal minimum hourly wage, and set any wages below this floor equal to it. This affects 4.3% of person-year observations with non-missing earnings. Among respondents who work at least one year, 36.2% have at least one year affected by this adjustment. Appendix A.2 provides more details and discussion.

Second, we flag and adjust unreasonably high wage observations. Specifically, for all observations with hourly wages in the top 0.1% of the wage distribution, we set weeks worked and weekly hours to missing, and impute them using the linear interpolation procedure in (4). We then impute wages using the original earnings and imputed annual hours via (3.1.2). This procedure treats extremely high wage outliers as being produced by misreported annual hours that are too low. While both earnings and hours are potentially mismeasured, these extremely high hourly wages are not driven by extremely high earnings but by extremely low hours.[10] We provide additional details on this topic in Appendix A.3.

[10] For example, among these top wage outliers the average annual earnings are $153423 compared to $238284 in the top 99% to 99.9% of wage earners, and $36509 below the top 99% of wage earners. Meanwhile, among these top wage outliers the average annual hours worked are 222 compared to 1566 in the top 99% to 99.9% of wage earners, and 1966 below the top 99% of wage earners.

3.2 Sample Selection

The previous section imputed missing values by linearly interpolating between the nearest prior and subsequent years with non-missing information. For individuals who miss several interviews in a row, or who decline to provide information for several interviews in a row, these "nearest" prior and subsequent years may be quite distant from the year to be imputed. In these cases, concerns over the accuracy of the imputation procedure may be particularly acute.

To address these concerns, we impose a maximum distance in years, $\overline{T}$, that can be used for imputation. If an individual has a year t in which the employment status is missing for the entire year, we keep them in our sample only if they have at least one observation with a direct report of the number weeks worked, which could be zero or positive, within $[t-\overline{T}, t+\overline{T}]$. We follow the same procedure for cases of missing usual weekly hours for all weeks worked in a year and missing annual earnings: we keep them in our sample only if they have at least one observation with a direct report of weekly hours or earnings within $[t-\overline{T}, t+\overline{T}]$. Appendix B provides a formal description of each criteria.

Lifetime Sample Our Initial Sample (see Section 2.1) of 9867 respondents is reduced to 9075 when removing individuals who died before turning 55, and is reduced further to 7171 when we condition on being interviewed at least once after age 55. This group of respondents forms the potential pool of our Lifetime Sample, which will be further restricted by our choice of $\bar{T}$.

How $\bar{T}$ Affects the Lifetime Sample The lower panel of Table 1a shows how many of those individuals remain in the Lifetime Sample after applying the selection criteria for different values of $\bar{T}$. The choice of $\bar{T}$ has a modest effect on the sample size when applied to missing employment status or missing weekly hours. In contrast, the sample size quickly increases with $\bar{T}$ when applied to missing earnings: in particular, increasing $\bar{T}$ from 1 year to 3 years increases the sample size from 30.4% to 57.5% of the initial sample. This is partly driven by the switch to a biennial survey after 1994: for example, an individual who misses earnings in a single reference year t after 1994 will not have a positive earnings report in either year $t-1$ or $t+1$ that can be used for imputation when $\bar{T}=1$.

Table 1b shows key labor market variables after applying different choices of $\bar{T}$. Mean employment (defined as working at least 520 hours that year) is decreasing in $\bar{T}$. One explanation for this is that individuals with fewer years of employment have fewer observations available to impute missing earnings; our selection procedure is therefore more likely to drop these people when $\bar{T}$ is low.[11] A second possible explanation is that individuals with lower attachment to the labor market

[11] For example, consider hypothetical individuals A and B, who both participate in all surveys. Suppose A is

Table 1: Full Year Missings — Sample Selection Criteria, Sample Size, and Outcome Variables

(a) Sample Selection Criteria and Number of Respondents

Selection Criterion for	$\bar{T}=1$	$\bar{T}=2$	$\bar{T}=3$	$\bar{T}=4$	$\bar{T}=5$	$\bar{T}=6$	$\bar{T}=7$	$\bar{T}=8$	$\bar{T}=9$
Initial Sample					9867				
Alive by 55					9075				
Interviewed after 55					7171				
Lifetime Sample									
Employment Status	7032	7115	7133	7142	7147	7149	7153	7159	7161
Weekly Hours	6622	6978	7078	7113	7125	7133	7141	7149	7151
Annual Earnings	3002	4892	5678	6064	6330	6482	6625	6706	6778
% of Initial Sample	30.4%	49.6%	57.5%	61.5%	64.2%	65.7%	67.1%	68.0%	68.7%

(b) Labor Market Outcomes in the Lifetime Sample

Outcome Variable	$\bar{T}=1$	$\bar{T}=2$	$\bar{T}=3$	$\bar{T}=4$	$\bar{T}=5$	$\bar{T}=6$	$\bar{T}=7$	$\bar{T}=8$	$\bar{T}=9$
Employment Rate	83.1%	82.5%	81.7%	81.3%	81.0%	80.8%	80.6%	80.5%	80.4%
Avg. Annual Hours	2146	2115	2103	2098	2096	2095	2095	2095	2093
Avg. Annual Earnings	51571	48210	47263	46793	46594	46607	46553	46505	46504
Observations	105070	171220	198730	212240	221550	226870	231875	234710	237230

(c) Labor Market Outcomes in the Cross-Sectional Sample

Outcome Variable	$\bar{T}=1$	$\bar{T}=2$	$\bar{T}=3$	$\bar{T}=4$	$\bar{T}=5$	$\bar{T}=6$	$\bar{T}=7$	$\bar{T}=8$	$\bar{T}=9$
Employment Rate	83.0%	82.4%	81.6%	81.1%	80.8%	80.6%	80.4%	80.3%	80.2%
Avg. Annual Hours	2149	2119	2106	2102	2100	2099	2099	2098	2097
Observations	306918	307727	308034	308173	308253	308313	308362	308398	308423
Avg. Annual Earnings	51636	48279	47333	46866	46669	46684	46631	46583	46583
Observations	221418	230514	234273	236430	237872	238861	239614	240157	240577

Notes: In Tables 1b and 1c, employment is defined as working at least 520 hours per year. Annual hours worked and annual earnings are conditional on being employed according to this definition.

might be more likely to miss several survey rounds and therefore have more consecutive missing observations. Because our sample selects positively for employment for lower values of $\bar{T}$, it is not surprising that annual hours and earnings are also higher for lower values of $\bar{T}$.

How $\bar{T}$ Affects the Cross-Sectional Sample Table 1c summarizes how our Cross-Sectional Sample changes as we vary $\bar{T}$. In contrast to our Lifetime Sample, which drops any individuals with a missing observation that cannot be imputed given $\bar{T}$, our Cross-Sectional Sample only drops the person-year observations that cannot be imputed. If a particular variable is missing, we

employed every year and B is employed every year except $t-1$ and $t+1$. Finally, suppose both A and B have missing earnings in year t. In this scenario, A will remain in the sample for any value of $\bar{T} \geq 1$ and B would remain in the sample for $\bar{T} \geq 2$. However, B would be dropped with $\bar{T} = 1$ because earnings in year t could not be imputed since the individual did not work in $t-1$ or $t+1$.

only drop the observation of that variable but not other variables: for example, an observation may have missing earnings but include weeks worked and usual weekly hours. As a consequence, in Table 1c the number of person-year observations varies between different variables for a given value of $\bar{T}$; in particular, annual earnings have fewer observations than employment and annual hours. Like in the Lifetime Sample, mean employment, annual hours, and annual earnings in the Cross-Sectional Sample are decreasing in $\bar{T}$.

Our Baseline Value of $\bar{T}$ In all the remaining analyses, we set $\bar{T}$ =5. Increasing $\bar{T}$ beyond five changes the sample size and the outcome variables in the Lifetime Sample by much less than increasing $\bar{T}$ for values below five. From our perspective this strikes a reasonable balance between maximizing sample size and minimizing measurement error in our imputation procedure.

The next three subsections provide information on the distribution of missing observations. We begin in Section 3.2.1 by studying missing observations among the 7171 respondents with at least one interview after age 55. We then successively apply the criterion for $\bar{T}$ =5 starting with the employment status, followed by weekly hours worked, and earnings.

3.2.1 Missing Weekly Employment Status

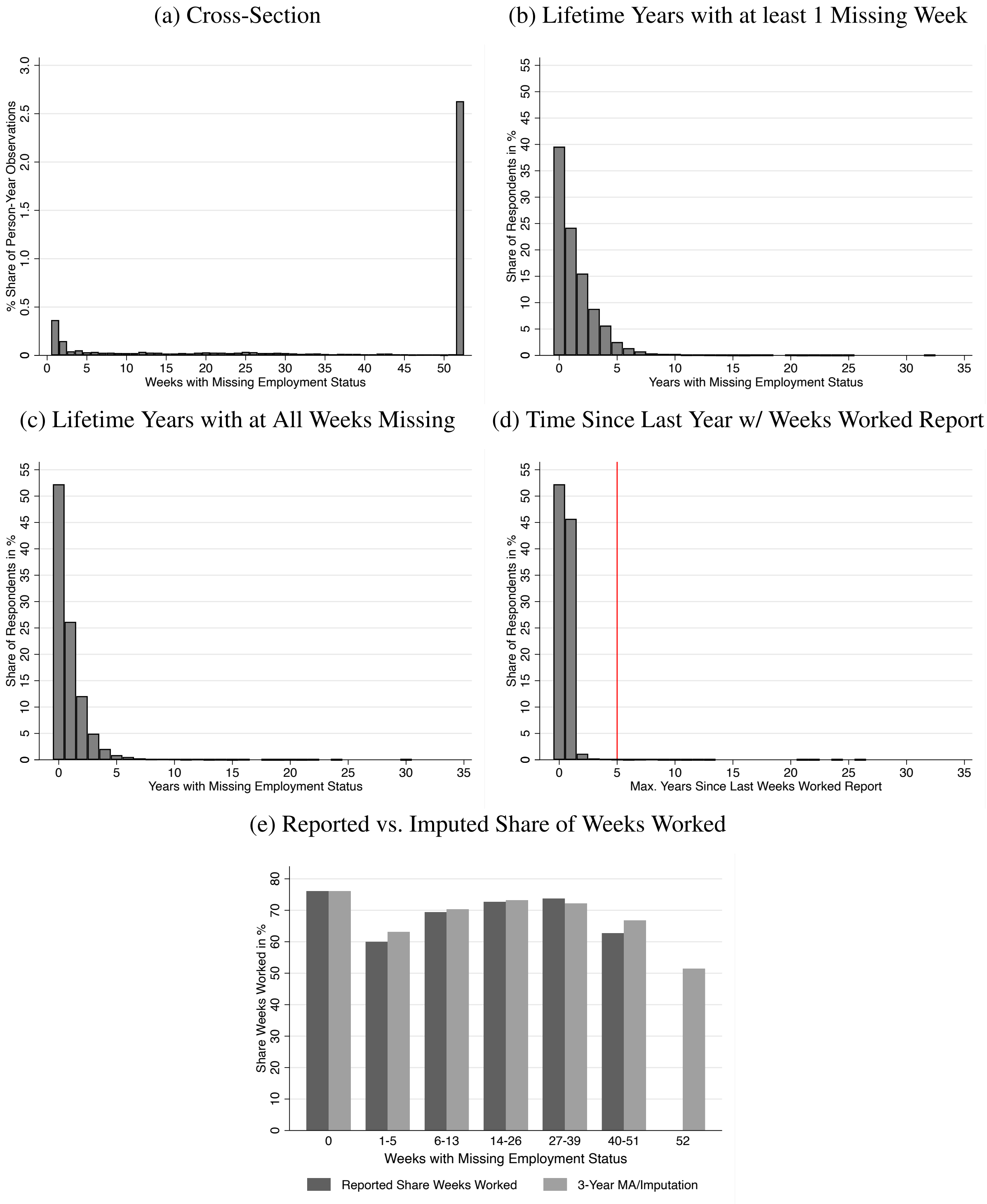
Of the total 250985 person-year observations in our initial sample, 4.2% have at least one week with missing employment status.[12] Figure 2a shows the distribution of weeks with missing employment status among person-year observations with at least one week of missing employment status. They are heavily concentrated on missing the employment status for all weeks of a year; a smaller but still notable number miss one or two weeks.

Figure 2b takes a lifetime perspective and shows the distribution of years with at least one week of missing employment status across respondents. 39.6% respondents do not have a single week with missing employment status, while 24.2% have just one year with at least one missing week. Figure 2c shows the distribution of "full-year" missings, in which the weekly employment status is missing for the entire year. For 26.2% this happens never, while 26.2% have just one full-year missing.

Figure 2d displays the largest gap between a year with the employment status missing for the entire year and the nearest year (either prior or subsequent) with a direct report of weeks worked. The red vertical line indicates our cutoff of $\bar{T} = 5$, to the right of which individuals are dropped. This leaves us with a sample of 7147 individuals.

[12] Figure C.1 shows that years with the employment status missing for at least some weeks are much more prevalent if an interview was missed that year.

Figure 2: Missing Employment Status



(a) Cross-Section

(b) Lifetime Years with at least 1 Missing Week

(c) Lifetime Years with at All Weeks Missing

(d) Time Since Last Year w/ Weeks Worked Report

(e) Reported vs. Imputed Share of Weeks Worked

Notes: In Figures 2a to 2d, we exclude the respective group with zero missings. For illustrative purposes, in Figures 2a and 2e years with 53 weeks worked in which the employment status is missing for 53 weeks are included in the 52 weeks bin.

Figure 2e compares (i) the mean share of weeks worked among weeks with non-missing employment (dark gray) with (ii) the mean imputed share of weeks worked among weeks with missing employment (light gray). We show this separately for different bins of missing weeks worked. The main takeaway is that the reported and imputed share of weeks worked are similar. When the employment status is missing for all weeks (represented by the 52 hours bin), there is no reported value of weeks worked to compare with. In this case, the low mean imputed employment rate implies that individuals who miss a full year of weeks have lower mean employment rates in the prior and subsequent years.

3.2.2 Missing Weekly Hours

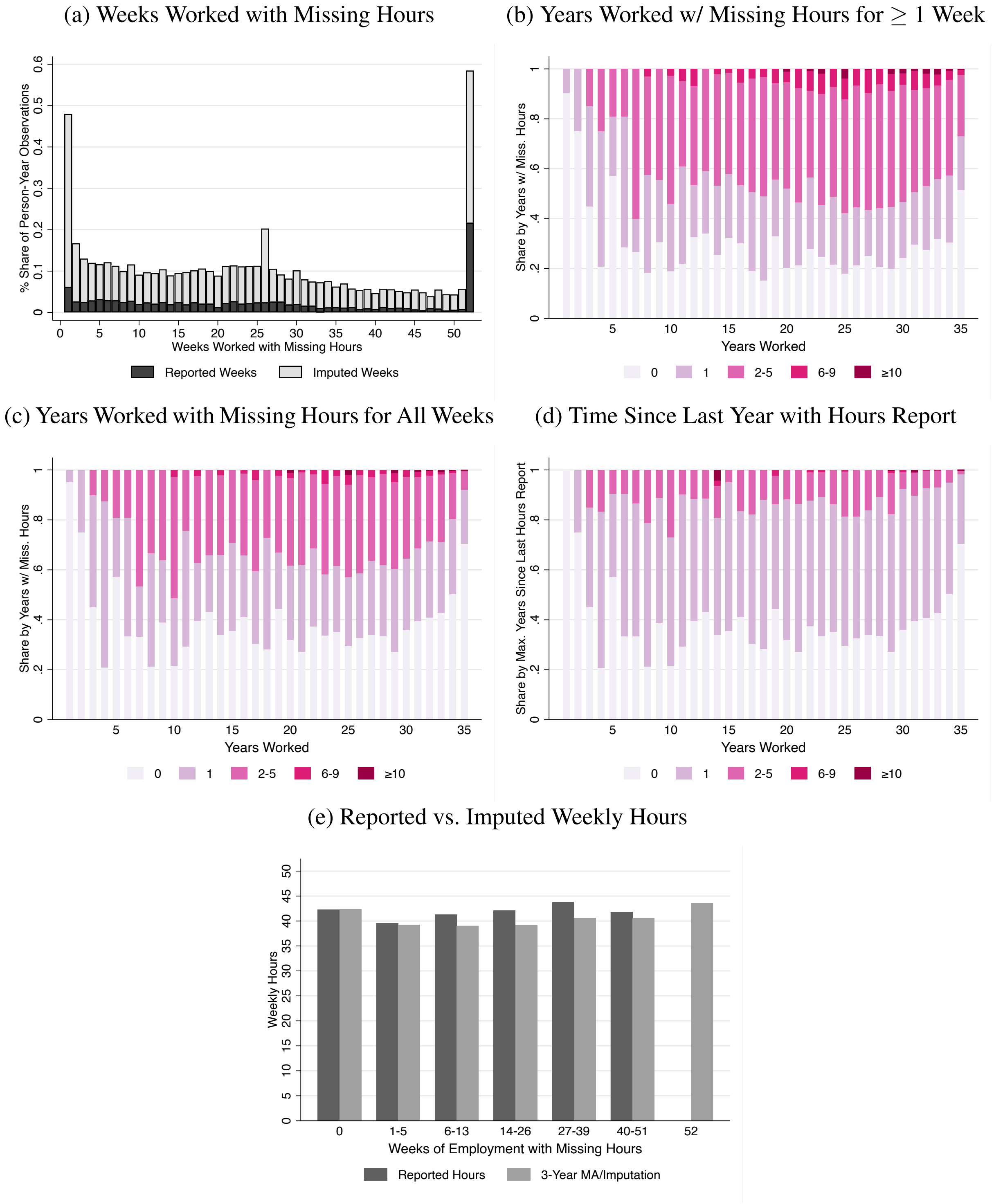
Conditional on working in a given week, usual weekly hours may be missing. Among the 7147 remaining individuals, we have 199296 person-year observations with positive weeks worked. Among those, 5.5% have at least one week worked with missing hours. Figure 3a shows the distribution of weeks worked with missing hours. Observations where the individual is employed but hours are not reported are in light gray; observations where the employment status is missing and weeks worked have been imputed (and thus weekly hours are missing) are in dark gray. Cases with missing hours are heavily concentrated on missing either one or all workweeks in a year.

Figure 3b takes a lifetime perspective. We group workers by their years worked (the horizontal axis), then within each group show the distribution of years with at least one week of missing hours. For example, among respondents who worked during all 35 years, 51.4% do not have a single year with some missing hours, and 21.6% have only one year with some missing hours. Figure 3c is constructed the same way except that the vertical axis now represents years worked with no hours information at all. The main takeaway from both of these figures is that a large majority of individuals have zero or only a few years with missing weekly hours.

Figure 3d shows the distribution of the largest gap between a year without any hours information and the nearest year worked (either prior or subsequent) with a direct hours report. Anyone with more than five years between two such observations is dropped. This leaves us with a sample of 7125 individuals.

Figure 3e compares mean hours worked among non-missing weeks (dark gray) with mean imputed hours worked among missing weeks (light gray). We show this separately for different bins of weeks worked with missing hours. The main takeaway is that the mean of the directly reported hours is similar to mean of the imputed hours.

Figure 3: Missing Weekly Hours



(a) Weeks Worked with Missing Hours

(b) Years Worked w/ Missing Hours for $\geq$ 1 Week

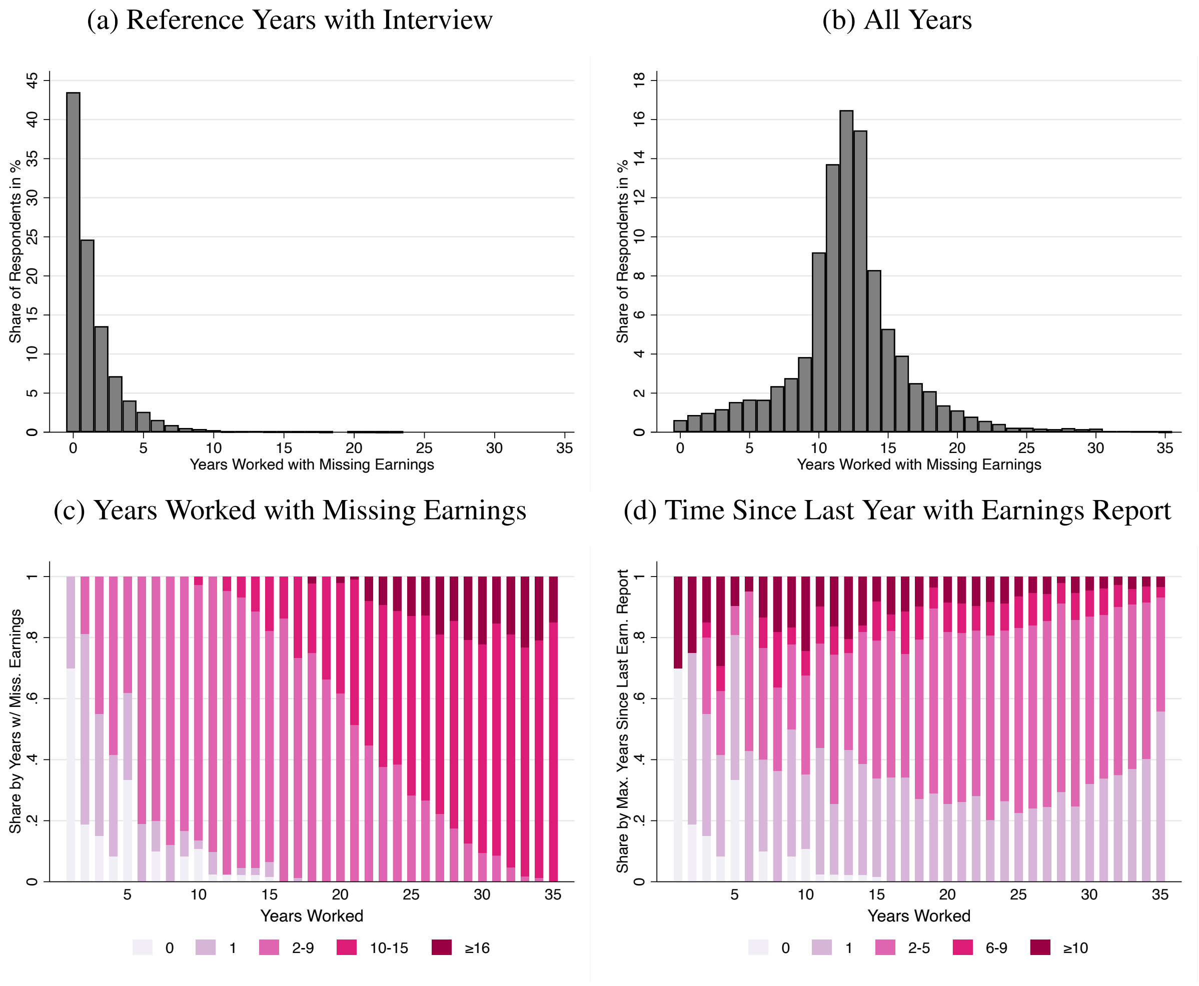
(c) Years Worked with Missing Hours for All Weeks

(d) Time Since Last Year with Hours Report

(e) Reported vs. Imputed Weekly Hours

Notes: A year of work is defined here as a year with positive weeks worked. In Figures 3a to 3d, we exclude the respective group with zero missings. For illustrative purposes, in Figures 3a and 3e years with 53 weeks worked in which hours are missing for 53 weeks are included in the 52 weeks bin.

Figure 4: Years Worked with Missing Earnings


(a) Reference Years with Interview

(b) All Years

(c) Years Worked with Missing Earnings

(d) Time Since Last Year with Earnings Report

Notes: A year of work is defined here as a year with positive weeks worked.

3.2.3 Missing Earnings

Most of our analysis will focus on the sum of both types of earnings in the NLSY79, i.e. income from wages and salary, and income from an own farm/business. We treat those combined earnings as missing only if earnings for both sources are missing. If one is missing but the other is not, we set total earnings equal to the non-missing source of earnings.

In contrast to weeks worked and hours worked, which are in principle collected retrospectively for all jobs since the last interview, earnings are only available for the reference year with an interview. Among the 7125 remaining individuals, we have 134444 person-year observations in a reference year in which an individual worked and took the interview. Among those, 7.3% have missing earnings.

Table 2: In-Sample Fit of Imputed Wages and Earnings

	Total Income		Wages and Salary	
	Report	Imputation	Report	Imputation
Hourly Wage	20.20	20.19	19.71	19.63
Annual Earnings	40198.7	42349.3	39046.1	40842.8
Wage-Annual Hours Correlation	0.01	0.14	0.01	0.16

From a lifetime perspective, Figure 4a shows that 43.6% of respondents do not miss earnings in any interview reference year and very few miss earnings in more than five reference years. Figure 4b shows the distribution of years with missing earnings, including non-interview years. This includes (i) years where an individual was interviewed but did not report earnings, (ii) interview years where the individual did not participate in the interview, and (iii) non-interview years.[13] Focusing on the last case, recall that the NLSY79 was conducted every other year after 1994. Depending on their birth cohort, individuals who worked every year automatically have missing earnings in at least 10-13 years simply due to the biennial switch in 1994. The mass of years worked with missing earnings is concentrated in precisely this range: specifically, 54.3% of our remaining respondents have between 10-13 years of missing earnings. Note that some individuals have fewer years of missing earnings because they did not work in all non-interview years.

Figure 4c shows how the distribution of years with missing earnings varies with years worked. For example, as described in the previous paragraph, every worker who works all 35 years has at least 10 years of missing earnings due to the biennial switch. At the same time, the figure also shows that few of these workers have more than 15 years of missing earnings.

Figure 4d reports the largest gap between a year worked with missing earnings and the nearest year worked (either prior or subsequent) with an earnings report conditional on the number of years worked. Anyone with more than five years between two such observations is dropped. This leaves us with a Lifetime Sample of 6330 individuals.

Table 2 compares earnings-related statistics in direct reports to those from observations with imputed earnings. Hourly wages are equal to annual earnings divided by annual hours. The mean of reported hourly wages almost exactly equals the mean of imputed hourly wages. The mean of imputed annual earnings exceeds that of reported earnings by 5.3%. This discrepancy is attributable to a wage-hours correlation that is somewhat higher in imputed data compared to direct reports.

The last two columns of Table 2 repeat the exercise for income from wages and salary only. We show these results because this is relevant measure for comparison of lifetime earnings with Social Security earnings histories from Section 4.3. The conclusions are essentially the same as

[13] Figure C.2a shows the prevalence of missing earnings in reference years distinguishing (i) and (ii).

for total earnings, partly because income from wages and salary is the by far dominating source of total earnings; see Appendix A.4 for details.

4 After 40 Years, How Representative is the NLSY79?

This section evaluates the representativeness of demographic characteristics and labor market outcomes in the NLSY79. Our primary point of comparison is the Annual Social and Economic Supplement of the Current Population Survey (ASEC), which we obtain from the IPUMS CPS database by Flood et al. (2023). Crucially, the ASEC collects annual-level information about labor market outcomes that correspond to measures in the NLSY79. In particular, weeks worked include weeks with (i) only a few hours or (ii) paid time off, and annual earnings are reported separately for (i) wages and salary and (ii) own farm/business income. We construct a subsample of the ASEC respondents who were born in the same years as the NLSY79 participants, i.e. between 1957 – 1964.[14] We further condition on having lived in the US in 1979 for those who were not born in the US. (This criteria is only fully observable in the ASEC beginning in 1994, so prior to 1994 we do not impose this; Appendix D.1 shows that after 1994 this criteria does not substantially change our results.)

Going forward, all statistics reported for the NLSY79 use the initial 1979 cross-sectional sample weights. In Appendix Table D.1 we show that using custom weights via the "Weight IDs" option on `https://nlsinfo.org/weights/nlsy79` has a minimal impact on our results. Results from the ASEC use the person sample weights.

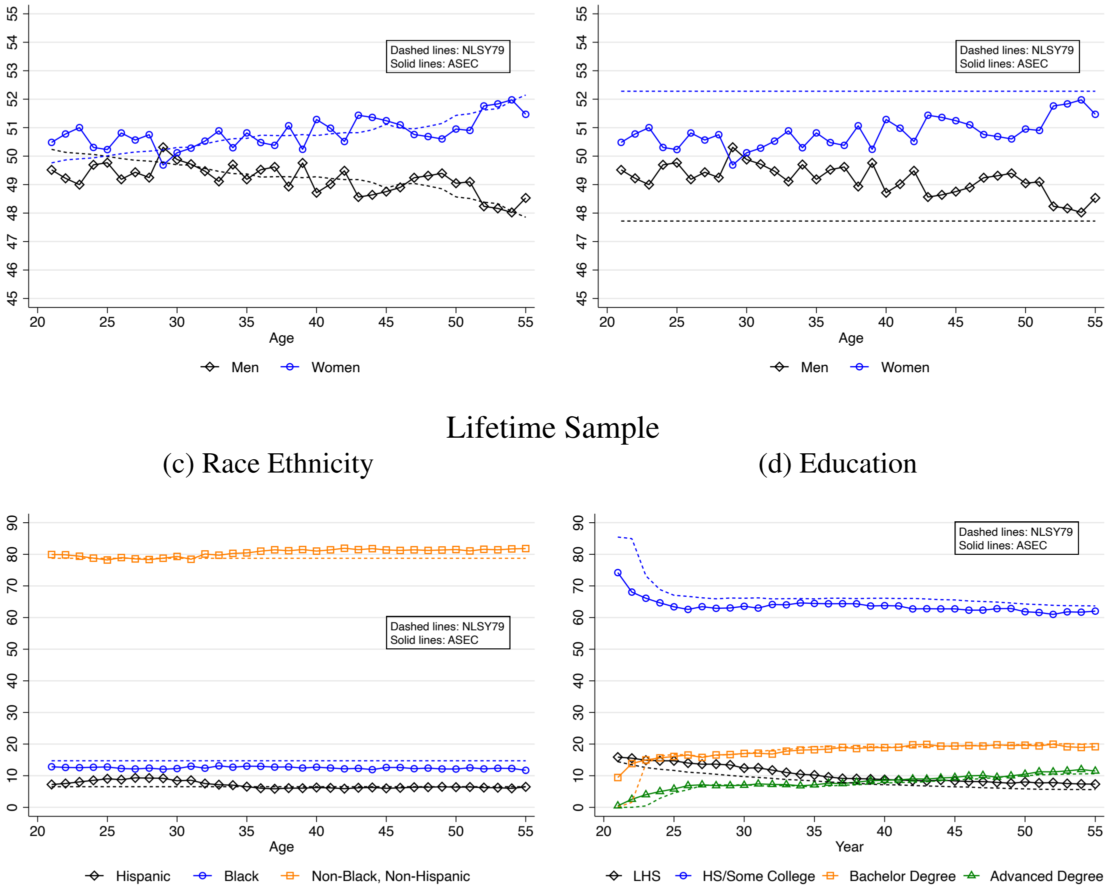
4.1 Demographics

Figure 5 compares the distribution of demographic characteristics in the NLSY79 (dashed lines) and ASEC (solid lines). Figure 5a shows a very similar share of men and women by age in the ASEC and the Cross-Sectional Sample of the NLSY79. In both datasets the share of women increases with age due to differential mortality. Figure 5b displays the same data from the ASEC now alongside the NLSY79 Lifetime Sample (the constant sex distribution is due to the balanced panel). We can see that women are slightly over-represented in the Lifetime Sample.

Figure 5c compares racial and ethnic shares by age in the ASEC and the NLSY79 Lifetime Sample. The two data sources have very similar shares of Hispanic, Black, and non-Hispanic/non-Black respondents; the largest discrepancy is a slight over-representation of Black respondents and slight under-representation of non-Hispanic/non-Black respondents in the NLSY79. Figure

[14] As in the NLS79, we calculate the birth year as the survey year minus age at the time of the survey.

Figure 5: Demographic Composition in the ASEC and NLSY79



Gender

(a) Cross-Sectional Sample

(b) Lifetime Sample

Lifetime Sample

(c) Race Ethnicity

(d) Education

Notes: In 1992 the CPS changed how education was recorded. In Figure 5d through 1991 we classify individuals by their highest grade completed (LHS=completed at most 11 grase, HS/Some College=completed 12-15 grades, Bachelor Degree=completed 16 grades, Advanced Degree=completed 17 or more grades) and from 1992 onwards by their highest degree completed.

5c compares education shares. During the early 20s the share having completed a bachelors or advanced degree is lower in the NLSY79 than in the ASEC. By age 35 the education shares have largely converged, though the NLSY79 has a slightly higher share of highschool/some college individuals and the ASEC has a slightly higher share of individuals who did not graduate from high school.

To sum up, demographics in both NLSY79 samples closely align with the ASEC; modest exceptions include slightly higher shares of female and Black respondents.

4.2 Labor Market Outcomes

Figure 6 compares the means of labor market outcomes in the NLSY79 and the ASEC. For each NLSY79 sample we plot two lines: one including imputed values and one using only directly reported data.[15] In almost all cases imputations have a minimal impact on our results. Therefore, all our discussion will focus on the series with imputations, unless otherwise noted. In Table 3a, we report the average difference over all ages between the four NLSY79 samples and the ASEC.

Figure 6a displays employment rates, defined as as having worked at least 520 hours in a year. Our first key takeaway is that the employment rate in the NLSY79 Cross-Sectional Sample is higher than in the ASEC. At age 21, the employment rate in the Cross-Sectional Sample is 4.4 percentage points higher than in the ASEC. The gap narrows until the mid 40s and then opens up again. On average the gap is 3.0 percentage points and by age 55 it is 4.1 percentage points. We emphasize that this disparity is not attributable to non-random sample attrition in the later years of the survey since it is apparent even at very early ages. Our second key takeaway is that the Lifetime Sample is further positively selected with regard to employment. On average, the employment rate including imputations in the Lifetime Sample is 1.8 percentage points higher than in the Cross-Sectional Sample. This difference also does not appear to be driven by survey attrition because at age 55 the gap is still 1.6 percentage points. Instead, the difference is attributable to a lower employment rate among individuals who remain in the survey through age 55 but are not in the Lifetime Sample because they have at least one missing observation that cannot be imputed. Our third key takeaway is that these patterns continue to hold if we alternately define employment status as having worked any positive number of annual hours rather than a minimum value of 520 hours; see Figure 6b. In summary, we find that the NLSY79 is positively selected based on employment, and that this is further exacerbated in our lifetime sample. However, we emphasize that these discrepancies are quantitatively modest and do not affect the shape of the employment profile over the life-cycle.

Figure 6c shows that, conditional on working at least 520 hours per year, average weeks worked are virtually identical across both NLSY79 samples and the ASEC. In contrast, hours worked per work week are on average 1.4 (3.5%) higher in the Lifetime Sample than in the ASEC, see Figure 6d. The largest gap is 2.0 hours (4.7%).

Figures 6e and 6f displays life-cycle profiles for mean annual earnings and hourly wages.[16]

[15] For the directly reported data, we make the following adjustments. We set weeks with missing employment status to non-employment. In situations where hours are available for only a subset of weeks worked, we us this reported value for the remaining weeks. We continue to replace wages below half the minimum wage with half the minimum wage and recalculate annual earnings as annual hours times half the minimum wage. This latter adjustment is also implemented in the ASEC.

[16] We do not harmonize topcodes across the two datasets because the topcoding thresholds and strategies are very similar both at a given point in time and over time, see `https://nlsinfo.org/content/cohorts/nlsy79/topical-guide/income/income` and `https://cps.ipums.org/cps/topcodes_tables.shtml`.

Table 3: Differences in Labor Market Variables between NLSY79 and CPS ASEC

(a) Average Differences for Figure 6 (Mean of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Employment Rate (Annual Hours $\geq$ 520)	4.7pp	4.8pp	3.0pp	3.0pp
Employment Rate (Annual Hours $>$ 0)	4.2pp	4.0pp	2.6pp	2.4pp
Weeks Worked	-0.3%	-0.1%	-0.5%	-0.2%
Weekly Hours	3.5%	3.6%	3.9%	4.0%
Annual Earnings	7.2%	6.8%	6.9%	6.0%
Hourly wage	3.7%	3.8%	3.5%	3.4%

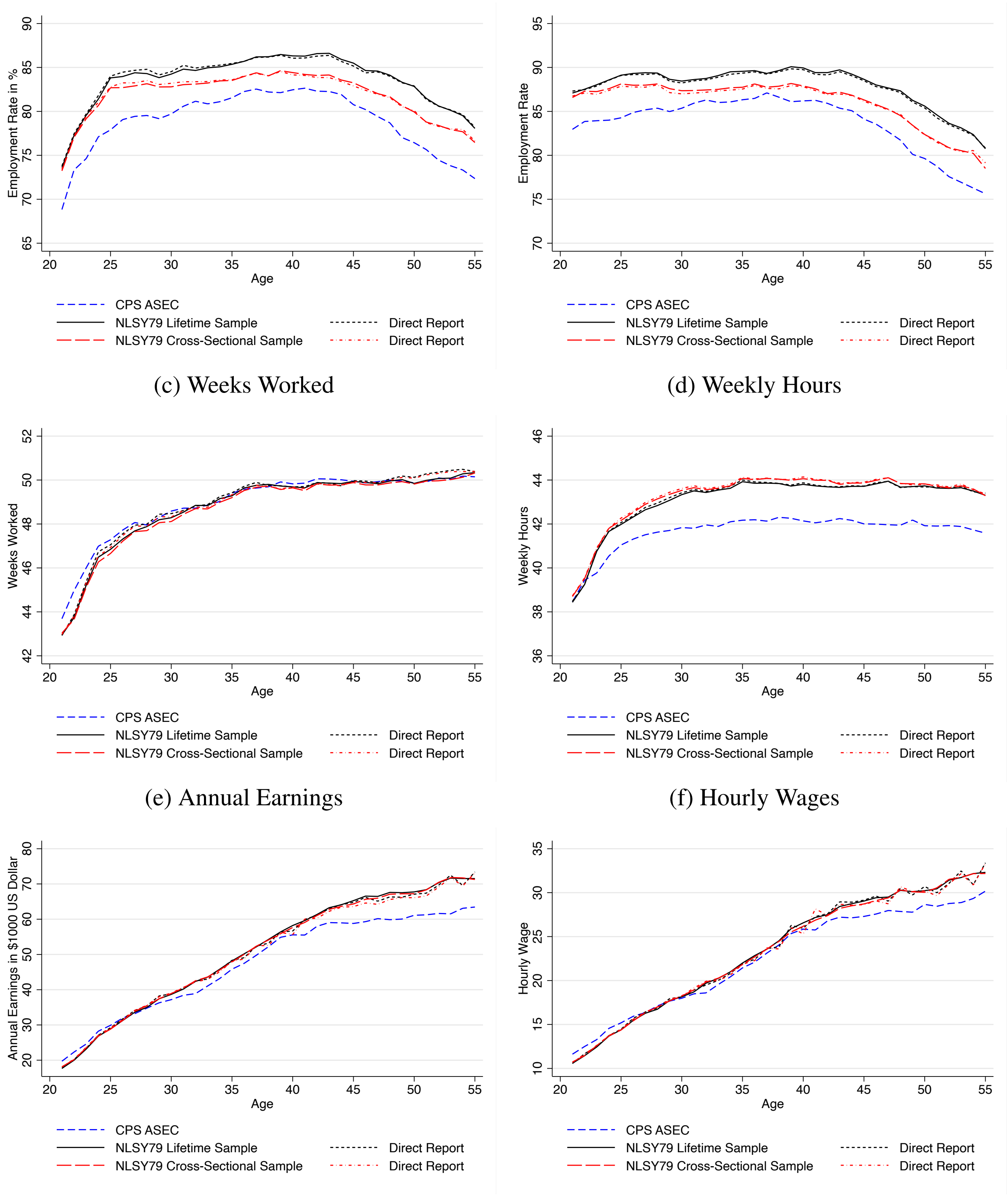
(b) Average Differences for Figure 7 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	3.9%	2.3%	6.2%	4.2%
Weekly Hours	16.8%	16.7%	19.5%	19.4%
Annual Earnings	2.5%	1.6%	5.4%	2.1%
Hourly wage	-1.5%	6.8%	2.1%	11.7%

(c) Average Differences for Figure 8 (Correlations)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Hourly Wages & Annual Hours Worked	-0.02	-0.03	-0.02	-0.03
Weekly Hours & Weeks Worked	-0.03	-0.04	-0.04	-0.04

Figure 6: Means of Labor Market Variables



(a) Employment Rate (Annual Hours ≥ 520)

(b) Employment Rate (Annual Hours > 0)

(c) Weeks Worked

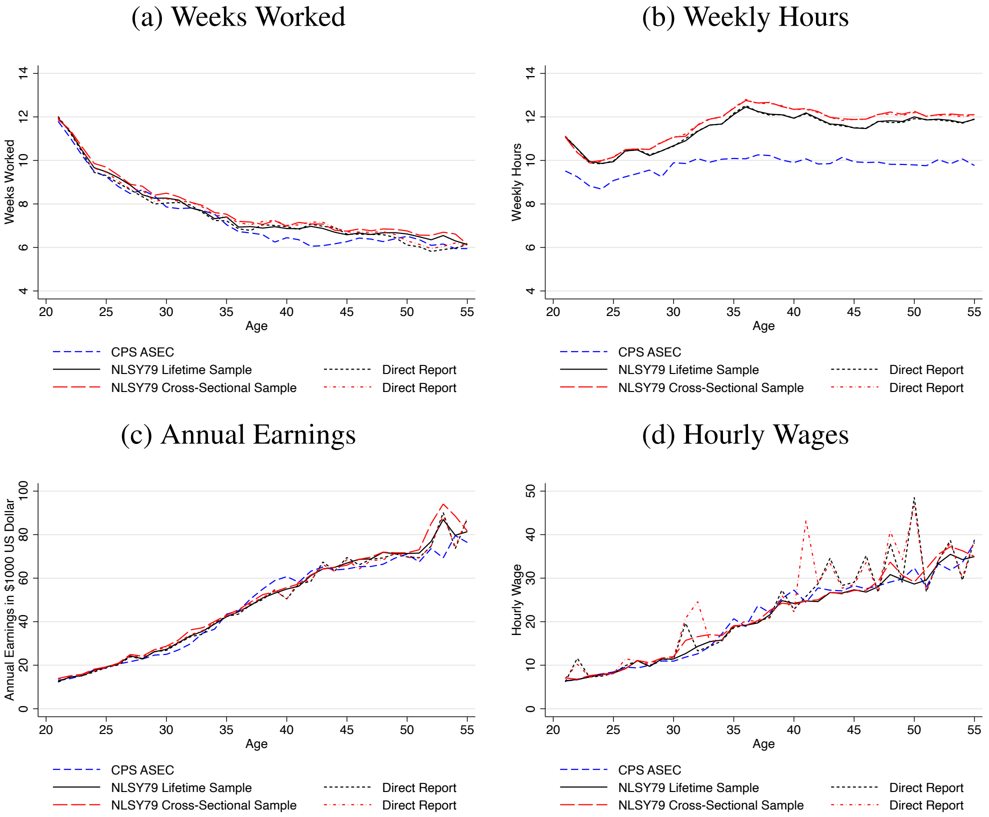
(d) Weekly Hours

(e) Annual Earnings

(f) Hourly Wages

Notes: In Figures 6c to 6f we condition on working at least 520 hours per year.

Figure 7: Standard Deviations of Labor Market Variables


(a) Weeks Worked

(b) Weekly Hours

(c) Annual Earnings

(d) Hourly Wages

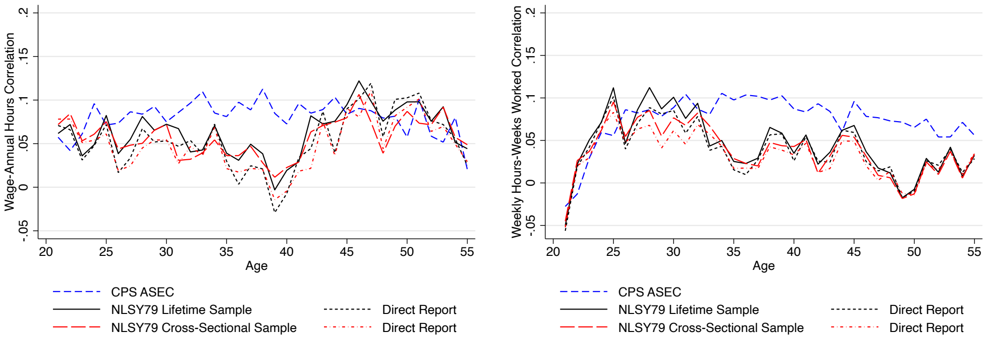
Notes: Only person-year observations for which annual hours are at least 520 are included.

(Appendix Figures D.4a and D.4b report the corresponding statistics for income from wages and salary only.) Prior to age 40, mean earnings in the two datasets track each other quite closely, with a slightly higher growth rate in the NLSY79. This is in line with the findings in MaCurdy et al. (1998) who analyzed the first 13 years of the NLSY79 data. After age 40 the two series begin to diverge because earnings growth slows more sharply in the ASEC. By age 55, mean total earnings in the NLSY79 Lifetime Sample are 12.7% higher than in the ASEC. This is partly driven by the higher hours per week in Figure 6d. But Figure 6f also shows a similar pattern in mean hourly wages: by age 55, mean hourly wages are 7.1% higher in the NLSY79 than in the ASEC.

Figure 7 displays standard deviations of weeks worked, weekly hours worked, annual earnings, and hourly wages, all conditional on working at least 520 hours in a year. In Table 3b, we report the average difference over all ages between the four NLSY79 samples and the ASEC. Figure 7a shows that the standard deviation of weeks worked is very similar across all samples and datsets. Figure 7b shows that the standard deviation of weekly hours is higher in the NLSY79 than in the ASEC. Both these patterns are similar to the corresponding mean profiles in Figure 6.

Figures 7c and 7d shows that the standard deviations of annual earnings and hourly wages in

Figure 8: Correlation of Labor Market Variables



(a) Hourly Wages & Annual Hours Worked

(b) Weekly Hours & Weeks Worked

Notes: Only person-year observations for which annual hours are at least 520 are included.

the NLSY79 closely align with those in the ASEC. In particular, while the means in the NLSY79 were somewhat higher than in the ASEC at later ages, this is not reflected in different standard deviations. These figures are one place in which imputations have a noticeable effect on the results. Specifically, the standard deviation of hourly wages for direct reports are quite noisy, which is attributable to individuals with missing employment status for a larger number of weeks in a given year with earnings in a typical range. Since we set those weeks with missing employment status to non-employment for the direct repots, the implied hourly wages are very high and erratic.[17]

Finally, Figure 8 shows that the NLS79 also closely tracks the correlation between key labor market variables in the ASEC, with Table 3c reporting the average difference over all ages between the four NLSY79 samples and the ASEC. Figure 8a shows the correlation between two components of annual earnings: hourly wages and annual hours. The two datasets produce similar average correlations (0.06 in the NLSY79 vs. 0.08 in the ASEC) and in both there is little systematic variation over the life-cycle.[18] Figure 8b shows the correlation between two components of annual hours: weekly hours and weeks worked. Again, the two datasets produce similar average correlations (0.04 in the NLSY79 vs. 0.07 in the ASEC). Both datasets also have the correlation increasing from slightly negative to slightly positive throughout the early 20's, then gradually declining for the remainder of the life-cycle.

In summary, the major discrepancies in labor market outcomes between the NLSY79 and the ASEC include an employment rate that is 4.7 percentage points higher on average, weekly hours that are 3.5% higher on average, mean hourly wages that are 2.6% on average, and annual earnings that are 5.5% higher on average (all these gaps refer to the Lifetime Sample, for which the gaps relative to the ASEC are largest). These discrepancies however do not affect the shape of the life-

[17] In fact, for expositional purposes, we cap in the NLSY79 the maximum direct reported wage at twice the maximum of the imputed wages. Otherwise, the last spike at age 50 would be about twice as large for both direct reports.

[18] In Bick et al. (2022), we show that the non-linear relationship between hourly wages and weekly hours worked in the NLSY79 and ASEC (and other standard data sets) are very similiar.

Table 4: Differences in Earning Ratios b/t the NLSY79 Lifetime Sample and the ASEC and SSA

	Cross-Section		Lifetime
	Pooled (Figure 9b)	Average 25-55 (Figures 9c-9f)	(Figure 10b)
P90/P50	6.2%	3.7%	0.2%
P50/P10	4.7%	2.9%	-0.2%
P90/P10	11.3%	6.7%	0.1%
P75/P25	4.8%	2.4%	0.0%

cycle profiles. By contrast, we find no meaningful differences in mean weeks worked conditional on employment, in the standard deviation of weeks, hourly wages, and annual earnings, or in the correlation of these variables with each other. Overall, we conclude that after 40 years labor market outcomes in the NLSY79 remain fairly representative of the 1957-1964 birth cohorts when compared with the ASEC.

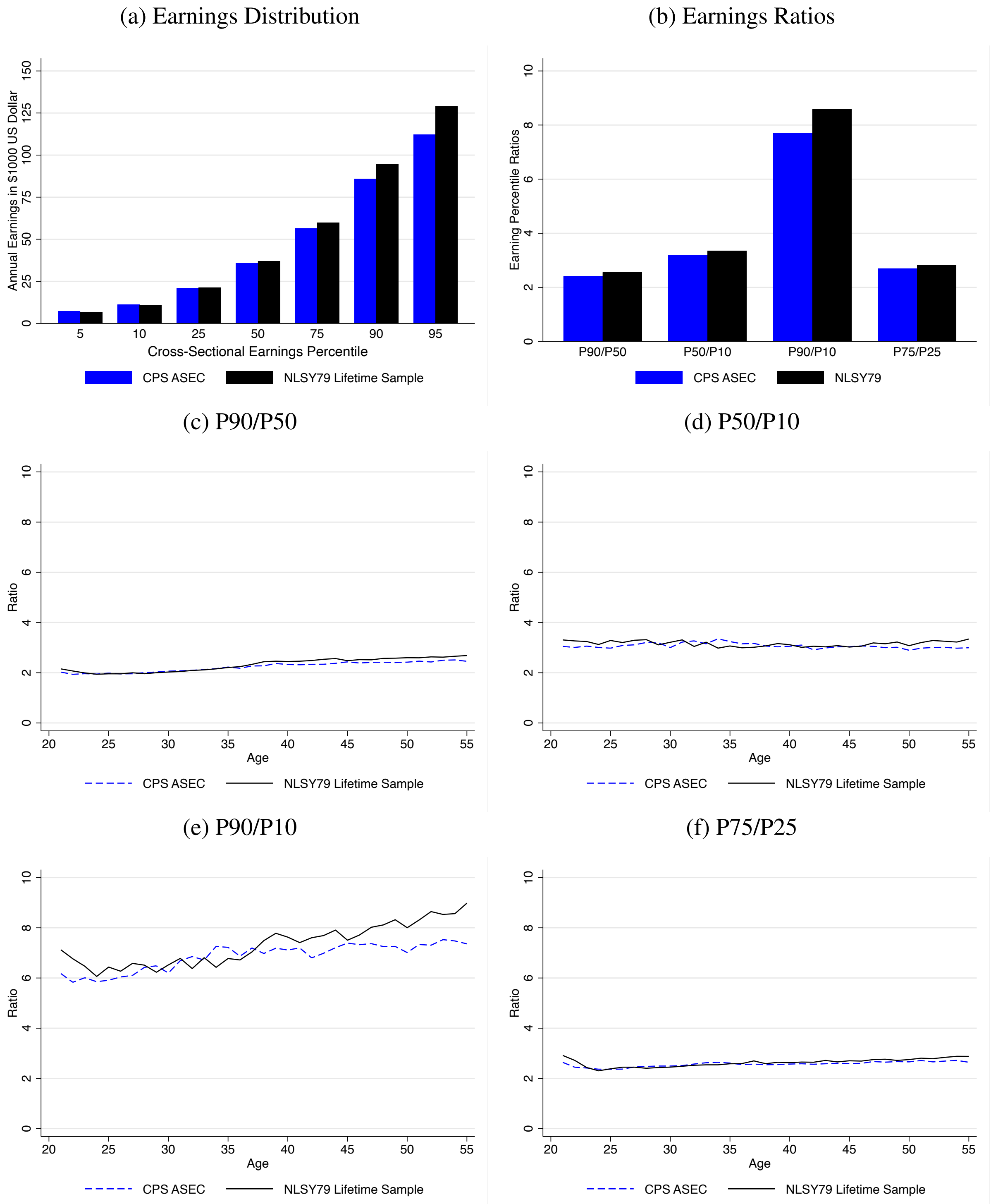
4.3 Earnings Inequality

This section documents a richer set of facts on the earnings distribution in the NLSY79. Given the similarity across different NLSY79 samples, our discussion focuses on the Lifetime Sample with imputed values.

Figure 9a documents various percentiles of the earnings distribution in the NLSY79 and ASEC. The gap between the NLSY79 and the ASEC is increasing in the earnings percentile. Hence, the upper part of the earning distribution explains why mean earnings in the NLSY79 exceed those in the ASEC documented in Figure 6e. Figure 9b displays four earnings percentile ratios commonly used in the earnings inequality literature: 90/50, 50/10, 90/10, 75/25; Table 4 summarizes the differences in the earnings ratios. All four measures are higher in the NLSY79, with the largest discrepancy for the P90/P10 ratio, see the first column of Table 4 for the exact magnitudes. Figures 9c to 9f show that these patterns hold not only in the cross-section but also over the life-cycle. The only exception again is the P90/P10 ratio, which starts to grow faster in the NLSY79 from the late 30s onwards, which is also around the time when mean earnings in the NLSY79 continue to grow much faster than in the ASEC (Figure 6e). We note that the difference between the NLSY79 and ASEC for the average of the P90/P10 over the life-cycle is substantially smaller than when the data are pooled.

These cross-sectional patterns do not necessarily imply that the NLSY79 has a representative distribution of average lifetime earnings. In particular, the cross-sectional comparisons do not evaluate whether the persistence of earnings is nationally representative. To assess whether this is the case, we compare outcomes in the NLSY79 with facts documented by Guvenen et al. (2022).

Figure 9: Cross-Sectional Earnings


(a) Earnings Distribution

(b) Earnings Ratios

(c) P90/P50

(d) P50/P10

(e) P90/P10

(f) P75/P25

Notes: We first identify the individual at the respective percentile of the cross-sectional earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower cross-sectional earnings and the five individuals with the closest highest cross-sectional earnings. Only person-year observations for which annual hours are at least 520 are included.

That study uses Social Security Administration (SSA) data on W2 wage and salary earnings to document lifetime earnings for several US cohorts. An advantage of using SSA data is that they likely contain less measurement error than survey data. Lifetime earnings are defined as average earnings from age 25-55. The two youngest cohorts in their sample with data covering this age range were born in 1957 and 1958, the two oldest cohorts in the NLSY79. We compare their results for these two cohorts to results from our lifetime sample; to maximize our sample size we include all birth cohorts 1957-1964.

We construct our measure of lifetime earnings following the same selection criteria imposed by Guvenen et al. (2022). An individual is included in the sample if he or she: (i) survived until at least age 55; (ii) had earnings that are larger than a year-specific threshold-level, denoted by Y_t, in at least 15 years between the ages of 25 and 55; and (iii) had total lifetime earnings of at least $\sum_{t=25}^{t=55} Y_t$, where Y_t, is the earnings level that corresponds to working at least 520 hours at one-half of the legal minimum wage in the year that individual was of age t. As in Guvenen et al. (2022) we only consider earnings from salary and wages from years when an individual was employed in "commerce and industry," a group of sectors that was continuously covered by the SSA. Also following Guvenen et al. (2022), we construct lifetime earnings as the average over those earnings without discounting.[19] The NLSY79 provides information of the industry of the most recent job until 1993, and from 1994 onwards we use information from the first job reported for that year. For years of employment with missing industry information, we use the last reported industry. Out of our final Lifetime Sample of 6330 individuals, 3964 satisfy these criteria.

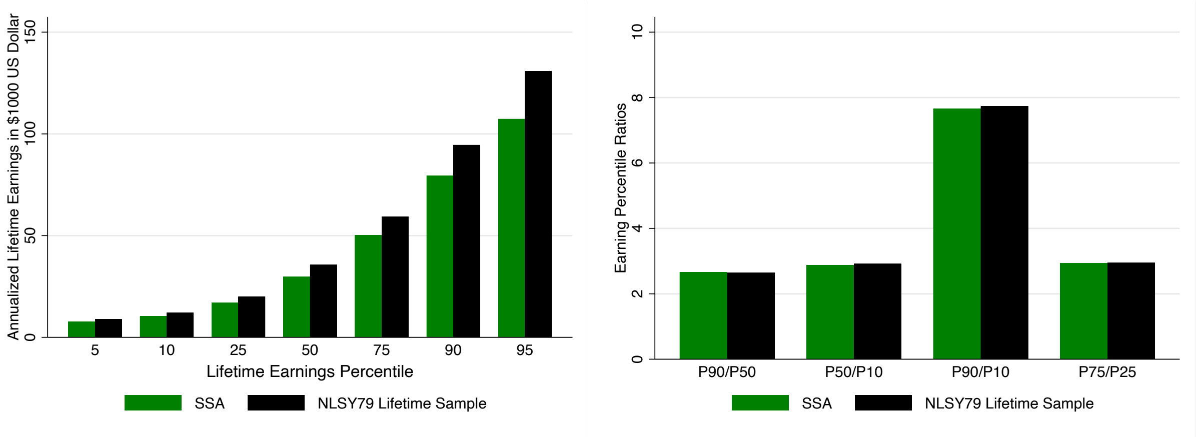
Figure 10a displays lifetime earnings percentiles in the NLSY79 and in Guvenen et al. (2022). Lifetime earnings are on average 14.8% higher in the NLSY79, with the magnitude differing across the distribution.[20] Figure 10b compares the earnings ratios in both data sets. These earnings ratios are remarkably similar, see also the last column of Table 4.[21] We conclude that the mean of lifetime earnings in the NLSY79 is modestly higher than in the SSA, but that inequality in lifetime earnings is very similar.

[19] Guvenen et al. (2022) define "commerce and industry" workers to include all indudstries except for agriculture, forestry and fishing, hospitals, educational services, social service, religious organizations and nonclassified membership organizations, private households, and public administration. We apply this same definition.

[20] Part of this difference in the mean can be attributed to SSA data covering only the two oldest NLSY79 cohorts. For these two cohorts mean lifetime earnings in the NLSY79 are 11.0% larger than in the SSA data, but 16.0% higher for the six younger cohorts.

[21] Appendix Figure D.6 adds to the picture lifetime earnings when including wage and salary earnings from years employed in industries other than "commerce and industry" and adds farm/business income. These additions have only a modest impact on the lifetime earnings distribution and lifetime earnings ratios.

Figure 10: Lifetime Earnings



(a) Lifetime Earnings Distribution

(b) Lifetime Earnings Ratios

Notes: For the NLSY, we identify the individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower lifetime earnings and the five individuals with the closest highest lifetime earnings. The Social Security Administration (SSA) data are from Guvenen et al. (2022).

4.4 Comparions by Sex

Figure 11 displays our key results separately by sex. We document the full set of outcome variables for the Lifetime Sample and Cross-Sectional Sample in Appendix D.5. Overall, we find similar results when we make comparisons separately by sex, though occasionally the discrepancies are larger for one sex or another.

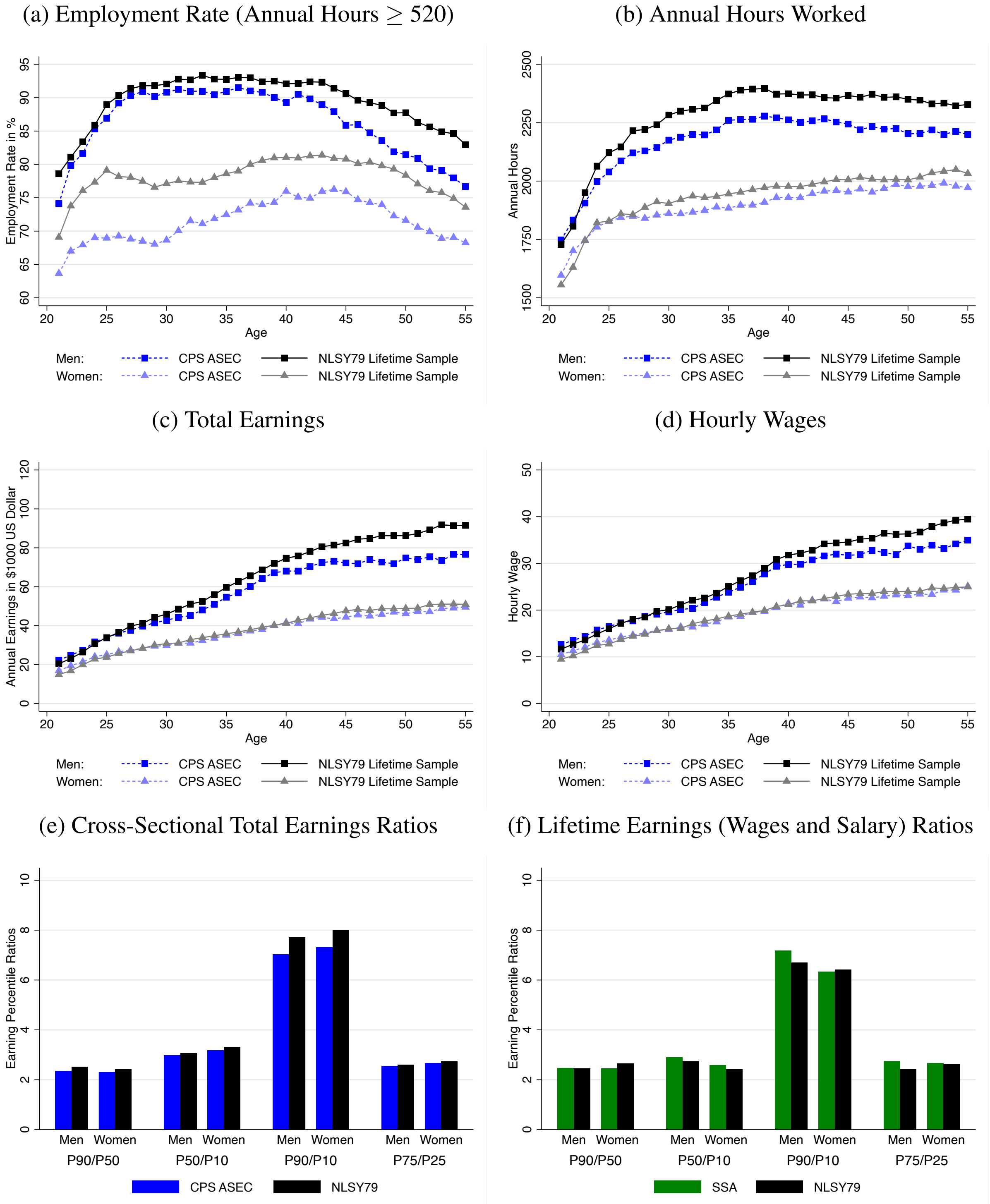
Figure 11a shows larger average employment rate gaps for women (on average by 6.7 percentage points) than men (on average by 3.1 percentage points). For men an employment gap emerges starting in the early 40s, when employment begins to decrease more quickly in the ASEC. By age 55, the employment rate in the NLSY79 exceeds the one in the ASEC by 5.3 percentage points for women and by 6.3 percentage points for men.

Figure 11b show that, conditional on being employed, men work on average 103.6 (4.8%) more hours per year in the NLSY79. For women the gap is with 38.3 (2.0%), somewhat smaller than for men. For both men and women, the gaps are again entirely driven by higher weekly hours, while weeks worked in both datasets are almost identical, see Appendix Figures D.7c and D.7d.

Figures 11c and 11d show that the aggregate gap in mean earnings and wages is almost entirely driven by higher mean earnings and wages of men. By age 55, male earnings in the NLSY79 exceed those in the ASEC by 11.9%, while this difference is only 2.5% for women.[22]

[22] Goldin et al. (2023) show that the gender gap in earnings for college graduates in the NLSY79 closely aligns with the Longitudinal Employer Household Dynamics, US linked administrative employer-employee database.

Figure 11: Key Results by Gender


(a) Employment Rate (Annual Hours ≥ 520)

(b) Annual Hours Worked

(c) Total Earnings

(d) Hourly Wages

(e) Cross-Sectional Total Earnings Ratios

(f) Lifetime Earnings (Wages and Salary) Ratios

Notes: In Figures 11b to 11e we condition on working at least 520 hours per year. In Figure 11e, we identify the individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the cross-sectional earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower lifetime earnings and the five individuals with the closest highest lifetime earnings. In Figure 11f, we follow the same approach for the NLSY79, while the Social Security Administration (SSA) data are from Guvenen et al. (2022).

Finally, Figures 11e and 11f show that earnings percentile ratios also align fairly closely with the ASEC and the SSA data for both men and women. Cross-sectional earnings inequality is slightly higher in the NLSY79 than in the ASEC, both for men and women. Lifetime earnings is slightly lower in the NLSY79 than in the SSA data, except for among women in the top half of the earnings distribution.

5 Conclusion

For more than 40 years, the NLSY79 has re-interviewed a sample of individuals born between 1957-1964. When excluding two supplemental samples which were (almost entirely) discontinued early in the survey, 74.5% of surviving initial respondents participated in the most recent available survey, conducted in 2020-2021. In this paper, we document that demographics and labor market outcomes in the NLSY79 remain broadly representative of their birth cohorts, both from a cross-sectional and from a lifetime perspective.

Based on these findings, we conclude that the NLSY79 continues to be a valuable dataset for studying life-cycle and lifetime labor market outcomes in the US. To facilitate these analyses, the replication package of this paper provides code for constructing our sample and dataset, as well as a basic dataset with imputed weeks worked, hours worked, and earnings for each individual in the NLSY79. These codes and dataset can be easily merged with any existing analysis, providing researchers confidence that their data work is based on a representative sample.

References

AUGHINBAUGH, A. AND R. M. GARDECKI (2007): "Attrition and Its Implications in the National Longitudinal Survey of Youth 1997," Working paper, Bureau of Labor Statistics and Ohio State University.

AUGHINBAUGH, A., C. R. PIERRET, AND D. S. ROTHSTEIN (2017): "Attrition and Its Implications in the National Longitudinal Survey of Youth 1979," Statistical survey paper, Bureau of Labor Statistics.

BICK, A., A. BLANDIN, AND R. ROGERSON (2022): "Hours and Wages," *Quarterly Journal of Economics*, 137, 1901–62.

——— (2024): "Hours Worked and Lifetime Earnings Inequality," Working paper.

FLOOD, S., M. KING, R. RODGERS, S. RUGGLES, J. R. WARREN, D. BACKMAN, A. CHEN, G. COOPER, S. RICHARDS, M. SCHOUWEILER, AND M. WESTBERRY (2023): "Integrated Public Use Microdata Series, Current Population Survey: Version 11.0," [dataset], Minneapolis, MN: IPUMS.

GOLDIN, C., S. PEKKALA KERR, AND C. OLIVETTI (2023): "The Parental Pay Gap over the Life Cycle: Children, Jobs, and Labor Supply," Working paper, Harvard University, Wellesley College, and Dartmouth College.

GUVENEN, F., G. KAPLAN, J. SONG, AND J. WEIDNER (2022): "Lifetime Earnings in the United States over Six Decades," *American Economic Journal: Applied Economics*, 14, 446–79.

HEATHCOTE, J., F. PERRI, AND G. L. VIOLANTE (2010): "Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006," *Review of Economic Dynamics*, 13, 15–51, special issue: Cross-Sectional Facts for Macroeconomists.

HEATHCOTE, J., F. PERRI, G. L. VIOLANTE, AND L. ZHANG (2023): "Unequal we stand: nequality Dynamics in the United States 1967-2021," *Review of Economic Dynamics*, forthcoming.

MACURDY, T., T. MROZ, AND R. M. GRITZ (1998): "An Evaluation of the National Longitudinal Survey on Youth," *The Journal of Human Resources*, 33, 345–436.

MACURDY, T. AND C. TIMMINS (2001): "Bounding the Influence of Attrition on Intertemporal Wage Variation in the NLSY," Working paper, Duke University and Stanford University.

A Imputations

A.1 Weekly Hours

When an observation has at least one week worked with missing hours, we impute average yearly hours worked as follows. We construct a three observation weighted moving average of weekly hours using the most recent prior year $\underline{t}$ and subsequent year $\bar{t}$ with positive weeks worked and an hours report for at least some weeks worked, where we denote the weights for each year as $q^h_{i,j} \forall \; j = \underline{t}, t, \bar{t}$:

$$\bar{h}_{i,t} = \frac{\sum_{j=\underline{t},t,f} q^h_{i,j} \times h_{i,j}}{\sum_{j=\underline{t},t,\bar{t}} q^h_{i,j}}. \tag{A.1}$$

Here, $q^h_{i,j}$ denote the weights for each year $j \in \{\underline{t}, t, \bar{t}\}$ used in the moving average. The formula for the weights is[23]

$$q^h_{i,\underline{t}} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,\underline{t},k}=1, h_{i,\underline{t},k}>0}}{t-\underline{t}}, \quad q^h_{i,t} = \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}>0}, \quad q^h_{i,\bar{t}} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,f,\bar{t}}=1, h_{i,\bar{t},k}>0}}{\bar{t}-t}.$$

We construct weights such that observations have more weight if they a) are more recent, and b) have more weeks worked with an hours report.[24] We use this moving average $\bar{h}_{i,t}$ to impute average weekly hours for year t:

$$\widehat{h}_{i,t} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}>0} \times h_{i,t,k} + \left(\widehat{wks}_{i,t} - \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}>0}\right) \times \bar{h}_{i,t}}{\widehat{wks}_{i,t}}. \tag{A.2}$$

Note that the second summand is zero for observations with no weeks worked with missing hours: there is no need to impute weekly hours for these observations.

If an hours report is available for all weeks worked, Equation (A.2) are average reported hours worked for this year.

[23] A few years have 53 weeks, in which case we adjust the summations in all expressions in this section.

[24] If there is no (recent) prior or subsequent year either because the individual is of age 21 or it is the respondent's last year in our sample, respectively, the individual did not work in any prior or any subsequent year, or all prior or subsequent years of work no hours report, we set $q^h_{i,\underline{t}} = 0$ or $q^h_{i,\bar{t}} = 0$, respectively. Note that for the younger NLSY79 cohorts we might also have observations before age 21 and for the older cohorts after age 55. However, in order to treat all cohorts symmetrically we do not use this information.

A.2 Wages Below Half the Minimum Wage

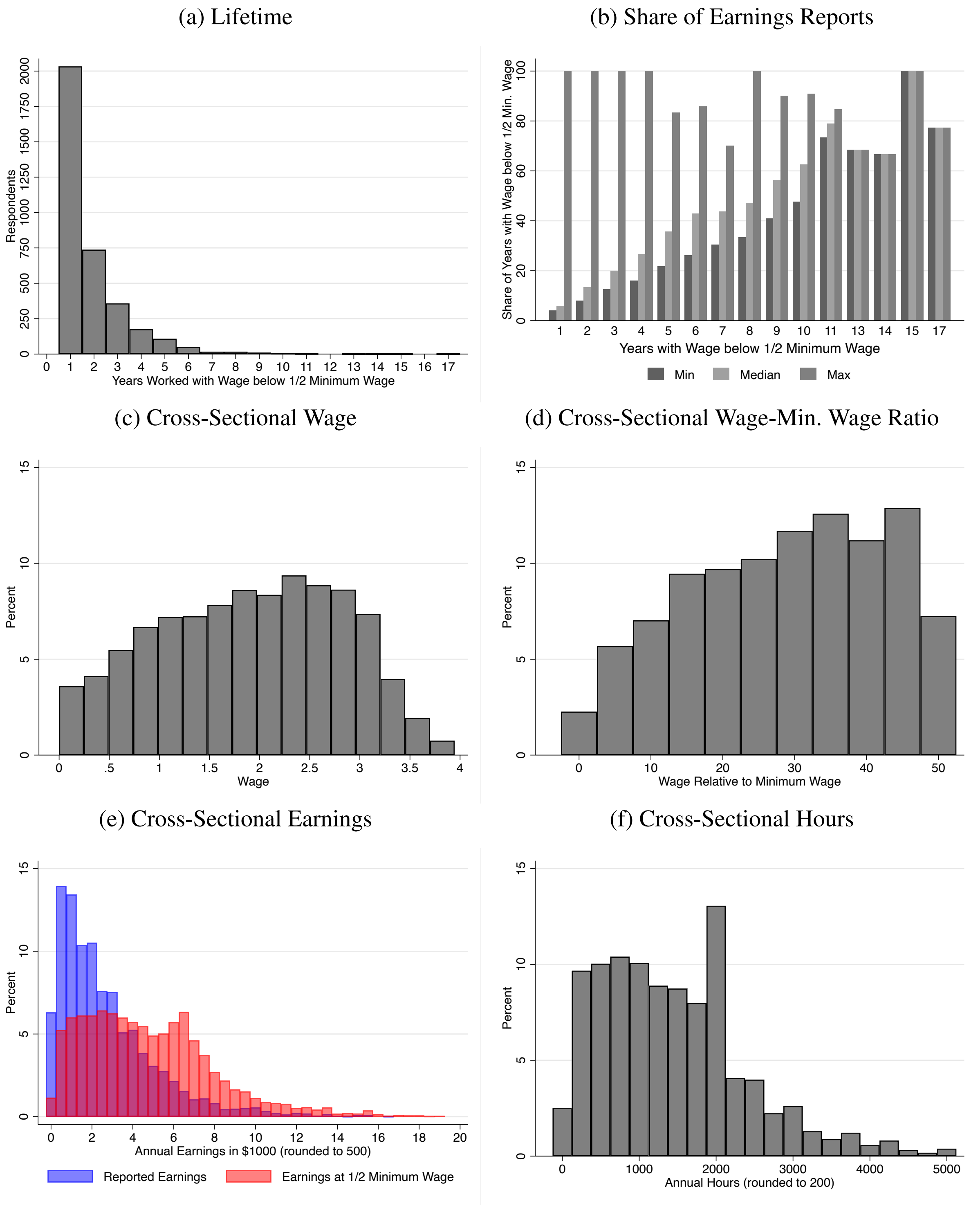
Figure A.1 provides more details on the prevalence of wages below half the minimum wage. We only use observations from reference years with a positive earnings report among the respondents in our final sample.

Figure A.1a plots the lifetime distribution of years worked with a wage below half the minimum wage in reference years with a positive earnings report conditional on having at least one such years. This is the case for 36.2% of individuals in our final sample who have worked at least one year. Figure A.1b provides some information on what share of years with an earnings report feature a wage below half the minimum wage conditional on the number of years with an earnings report with a wage below half the minimum wage.

Figure A.1c shows the distribution of wages below half the respective minimum wage and Figure A.1d the same wages relative to the respective minimum wage.

Figure A.1e contrasts the distribution of reported earnings with a wage below half the minimum wage with the distribution of earnings we assign to those individuals based on their hours worked (see Figure A.1f) and half the minimum wage in the respective year. Average reported earnings are $2688 compared to $4791 after the adjusment.

Figure A.1: Distributional Facts on Wages below Half the Minimum Wage



(a) Lifetime

(b) Share of Earnings Reports

(c) Cross-Sectional Wage

(d) Cross-Sectional Wage-Min. Wage Ratio

(e) Cross-Sectional Earnings

(f) Cross-Sectional Hours

A.3 Top Earnings and Wages

Figure A.2 provides more information on the top 1% of the cross-sectional wage distribution implied by Equation (5). To put those numbers into perspective, we also provide information on the top 1% of the cross-sectional distribution of directly reported earnings.

In each panel of Figure A.2, we show an outcome variable for people who are in the 99th to 99.9th percentile of the earnings distribution (first three bars), in the top 0.1% of the earnings distribution (next three bars), in the 99th to 99.9th percentile of the wage distribution (next three bars), and in the top 0.1% of the wage distribution (last three bars). The "Current" (dark grey) bar represents individuals who are currently in the respective earnings groups, while the "Before" (blue) and "After" (red) bars show the outcome variable for these individuals in the year before and after, independent in which earnings group they were "Before" or "After". The year before or after refers to the most recent respective year with a positive earnings report. The "Before" and "After" sample can be smaller than the "Current" sample if the current year is the first or last one with a positive earnings observation. The set of individuals used for each bar is identical across all six subfigures of Figure A.2.

Figure A.2a shows that individuals in the top 0.01% of the cross-sectional wage distribution not only have a drastically larger current wage relative to the three other groups, but also relative to the years before and after.

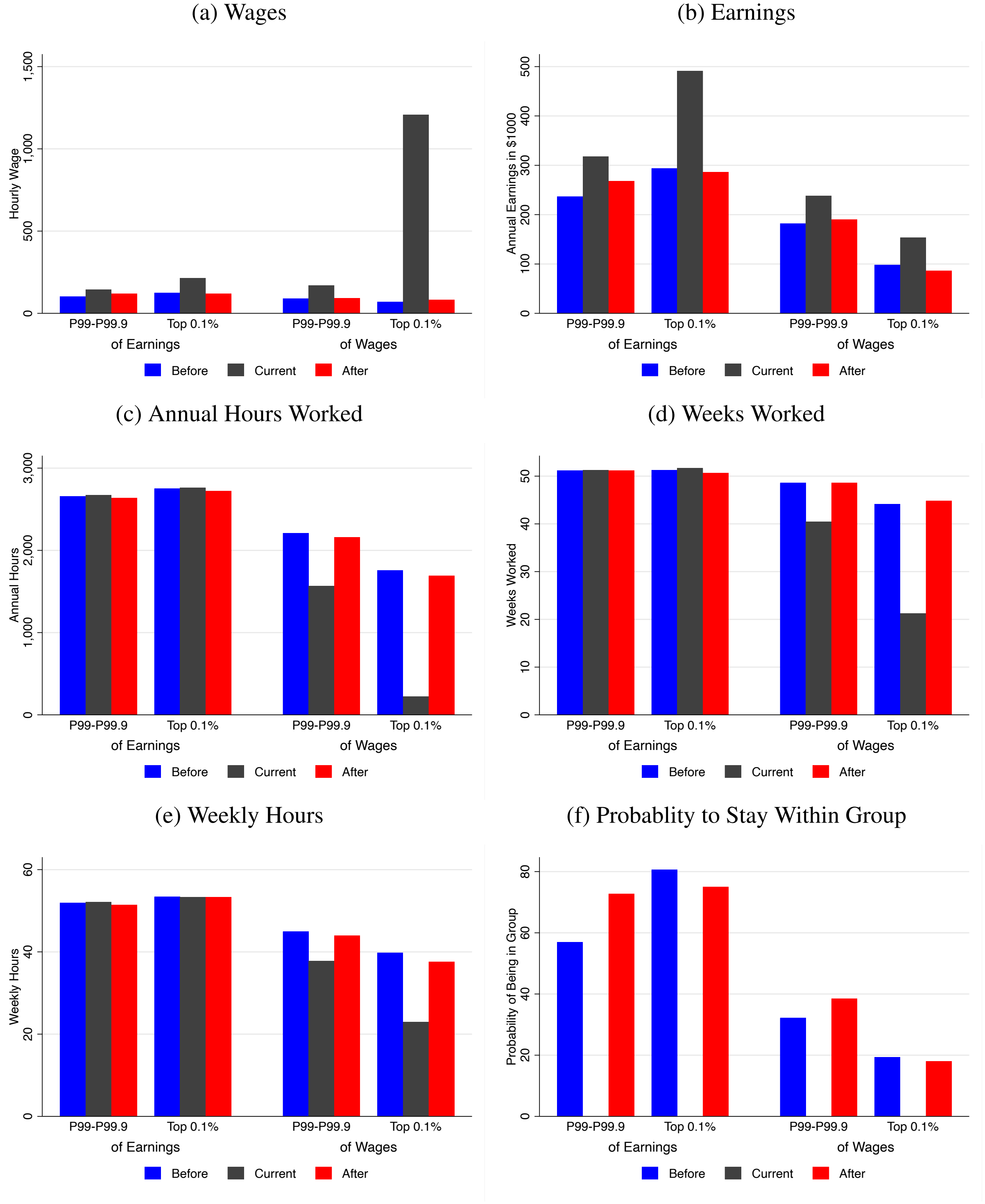
Figure A.2b repeats this exercise but now reports average earnings rather than average wages. Here the patterns are more similar across all four groups. It does stand out though that the top wage earners have lower earnings than the top earners, and that the top 0.1% of wage earners have even lower earnings than the top 99% to 99.9% of wage earners.

Figure A.2c rationalizes these patterns. Hours for those with top wages are lower than hours for top earners. Moreover, hours for top earners change fairly little, which is not the case for those in the top of the wage distribution. This is particularly true for the top 0.01% of wages, for which hours essentially collapse – and therefore mechanically wages increase so drastically – and immediately recover again. Figures A.2d and A.2e show that this drop annual hours comes both from a reduction in weeks worked and weekly hours worked, with the former effect being larger.

Finally, Figure A.2f shows the probability of being in the respective earnings group also in the year before and after. Top 0.01% earners are much more likely to have been and remain in that group compared to those with top 0.01% wages.

Table A.1 has some extra information. The first two rows in each panel of Table A.1 state the corresponding numbers for the top 1% of wage earners. The third row in each panel shows the updated outcome for those previously in the top 0.1% of the wage distribution after imputing weeks worked and weekly hours work for those years. Annual hours are still lower relative to before and after, but the reduction is now more in line with the top 99% to 99.9% of wage earners, and thus are wages. The two bottom panels further show that almost all person-year observations in the top 1% of the wage distribution are direct reports without virtually any weeks of missing employment status for those years.

Figure A.2: Top Earnings and Wages



(a) Wages

(b) Earnings

(c) Annual Hours Worked

(d) Weeks Worked

(e) Weekly Hours

(f) Probablity to Stay Within Group

Table A.1: Initial Top 1% of Wages

	Before	Current	After
Annual Earnings (in 1000)			
Top 99%-99.9%	182	238	190
Top 0.1%	98	153	87
Annual Hours			
Top 99%-99.9%	2211	1566	2161
Top 0.1%	1759	222	1690
Top 0.1% - Adjusted	1761	1199	1695
Weeks Worked			
Top 99%-99.9%	49	40	49
Top 0.1%	44	21	45
Top 0.1% - Adjusted	44	39	45
Weekly Hours			
Top 99%-99.9%	45	38	44
Top 0.1%	40	23	38
Top 0.1% - Adjusted	40	31	38
Average Hourly Wage			
Top 99%-99.9%	91	170	93
Top 0.1%	71	1207	82
Top 0.1% - Adjusted	70	202	97
Weeks with Missing Employment Status			
Top 99%-99.9%	0	0	0
Top 0.1%	0	1	1
Weeks Worked with Missing Hours			
Top 99%-99.9%	0	0	0
Top 0.1%	0	0	1

A.4 Income from Wages and Salary

Our point of comparison for lifetime earnings will Guvenen et al. (2022), who use earnings reported in W2 forms to the Social Security Administration and thus by definition only covers this source of income. In our sample, in the cross-section of employed in reference years 95.3% person-year observations feature earnings only from wages and salary, 1.2% only from farm/business income and 3.5% from both sources. From a life-time perpsective, almost everyone who works for at least one year receives earnings from salaries, wages, etc. in at least one year of their working life, while 30.3% have at least one year with earnings from farm/business income and 26.7% have at least one year with both types of earnings.

We only use earnings observations as input into the imputation for the years in which individuals received earnings from wages and salary exclusively to ensure that they are not a mix of the two type sources of earnings. We also require that for any missing earnings from wages and salary, we have at least one earings report from wages and salary with in the previous or next for years. Only 3% do not satisfy this criteria. For the years, for which we impute earnings we assume that all hours worked are from work as an employee even though some of the hours might come from self-employed work. We think however that this assumption is not particularly problematic given the relatively small share of individuals receiving earnings from farm/business income (whether solely or jointly with income from wages and salary) documented in the previous paragraph. In addition for those earning both types of income, income from wages and salary is on average the dominant source. The share of income wages and salary from total earnings for those having both sources of earnings is 48.7% at the 10th percentile, 51.9% at the 25th percentile, 76.9% at the 50th percentile, and 92.6% at the 75th percentile.

B Sample Selection

Employment Status For the employment status, we define the following indicator variable

$$m^e_{i,t} = \begin{cases} 0 & \text{if } \sum_{k=1}^{K_t} \mathbb{I}_{e_{i,t,k}=-1} = K_t \\ 1 & \text{otherwise,} \end{cases} \tag{B.1}$$

i.e., $m^e_{i,t}$ equals zero when the employment status is missing for all weeks in year t. We keep an individual if for any $m^e_{i,t} = 0$, we have at least one observation within the past $\bar{T}$ or next $\bar{T}$ years with $m^e_{i,j} = 1,\ j \in \{t-\bar{T}, t+\bar{T}\}$. Put differently, we require that for any $m^e_{i,t} = 0$, $\sum_{j=t-\bar{T}}^{J=t+\bar{T}} m^e_{i,j} \geq 1$.

Weekly Hours For weekly hours worked, the indicator takes the form:

$$m^h_{i,t} = \begin{cases} 0 & \text{if } \sum_{k=1}^{K_t} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}=-1} = \widehat{wks}_{i,t} \ \& \ \widehat{wks}_{i,t} > 0 \\ 0 & \text{if } \widehat{wks}_{i,t} = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{B.2}$$

i.e., $m^h_{i,t}$ equals zero if hours worked are missing for all weeks worked (first line), or if an individual has not worked at all this year (second line). We treat years of non-employment equal to years of all missing weekly hours because neither of the two carries direct information on hours worked conditional on working for that year. Using the same threshold $\bar{T}$ as for weeks worked, we keep an individual if for any $m^h_{i,t} = 0$ and $\widehat{wks}_{i,t} > 0$, we have at least one observation within the past $\bar{T}$ or next $\bar{T}$ years in which which $m^h_{i,j} = 1,\ j \in \{t-\bar{T}, t+\bar{T}\}$. Put differently, we require that for any $m^h_{i,t} = 0$ and $\widehat{wks}_{i,t} > 0$, $\sum_{j=t-\bar{T}}^{J=t+\bar{T}} m^h_{i,j} \geq 1$.

Earnings For annual earnings, the indicator takes the form:

$$m^y_{i,t} = \begin{cases} 0 & \text{if } y_{i,t} = -1 \ \& \ \widehat{wks}_{i,t} > 0 \\ 0 & \text{if } \widehat{wks}_{i,t} = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{B.3}$$

i.e., $m^y_{i,t}$ equals zero if earnings are missing (first line), or if an individual has not worked at all this year (second line). By the same logic as for weekly hours, we treat years of non-employment equal to years of missing employment because neither of the two carries direct information on annual earnings. Using the same threshold $\bar{T}$ as for weeks worked, we keep an individual if for any $m^y_{i,t} = 0$ and $\widehat{wks}_{i,t} > 0$, we have at least one observation within the past $\bar{T}$ or next $\bar{T}$ years in which which $m^y_{i,j} = 1,\ j \in \{t-\bar{T}, t+\bar{T}\}$. Put differently, we require that for any $m^y_{i,t} = 0$ and

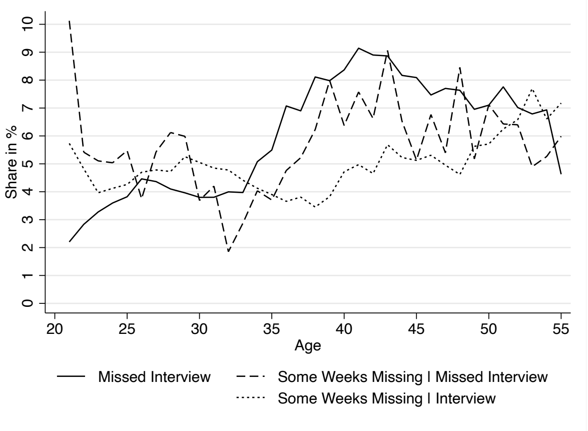
$\widehat{wks}_{i,t} > 0, \sum_{j=t-\bar{T}}^{J=t+\bar{T}} m_{i,j}^{y} \geq 1.$

C More on Missings

C.1 Weeks with Missing Employment Status

The solid line in Figure C.1 shows for reference years, i.e. years prior to which an interview round took place, the share of individuals in our sample of 7171 individuals who missed that interview. It increases from around 2% to a peak of 9% around in the early 40s. The sharp drop at age 55 is a result of the selection criteria requiring at least one interveiw after age 55. The two youngest cohorts have to participate in that interview to be included in our sample. Also note that the stark increase in the mid 30s coincides with the switch to the NLSY79 being conducted only every other year. The two other lines in this figure report the probability of having some missing weeks conditional on having missed the interview (long dashs) and conditional on having participated in the interview (short dashs). The probability of having weeks with missing employment status is not surprisingly much larger for those who also missed an interview.

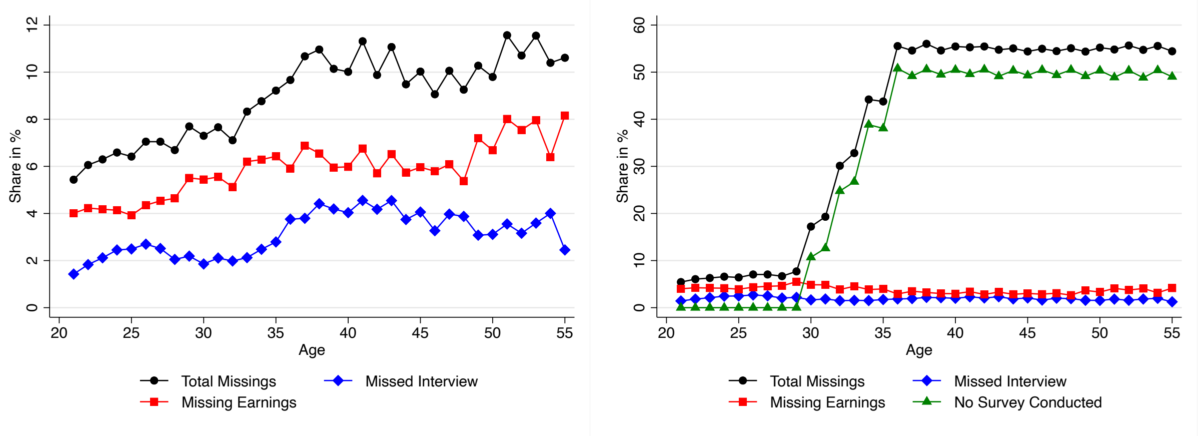
Figure C.1: Interview Status and Prevalence of Weeks with Missing Employment Status



C.2 Missing Earnings

Figure C.2a documents the prevalence of missing earnings during references years. The top line with circles refers shows the total share of missing earnings. The two remaining lines breakdown total missings between a missed interview (blue line with diamonds) and actual missings earnings, i.e. an individual taking part in the survey but not reporting their earnings (red line with squares), with the latter being somewhat more important. Figure C.2b now includes also missing earnings in non-reference years. The raise between age 29 and 36 reflects the different ages at which the switch to the NLYS79 becoming biennial took place. As a result of this switch about 50% of earnings are missing by construction in the second half of the life-cycle.

Figure C.2: Missing Earnings



(a) Missing Earnings in Reference Years

(b) Missing Earnings in All Years

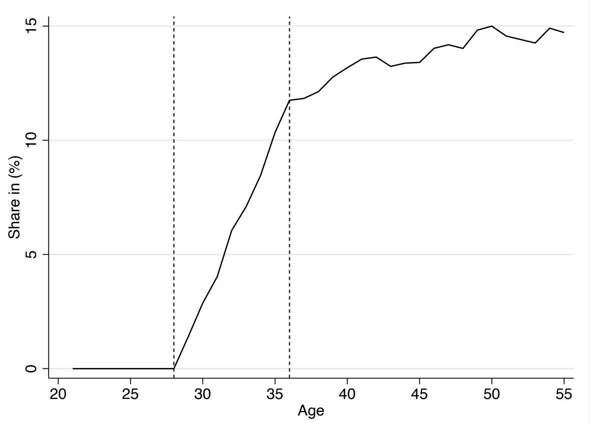
D Representativeness of the NLSY79

D.1 The Effect of Immigration in the ASEC

The NLSY79 cohorts only include inviduals who lived in the US in 1979. The ASEC includes anyone living in the US independent of whether someone was born in the US or immigrated to the US. The information to identify when someone moved to the US (independent of whether they were already an US citizen by then or not) is only available from 1994 onwards in the ASEC. The oldest NLSY79 cohort was 37 in 1994 and the youngest 30. Hence, only once that youngest cohort reaches age 37, the corresponding ASEC sample is fully comparable to the NLSY79. Figure D.1 shows the share of individuals identified as not living in the US in 1979. Here we shifted everything by one year as all our labor market variables refer to the previous year.

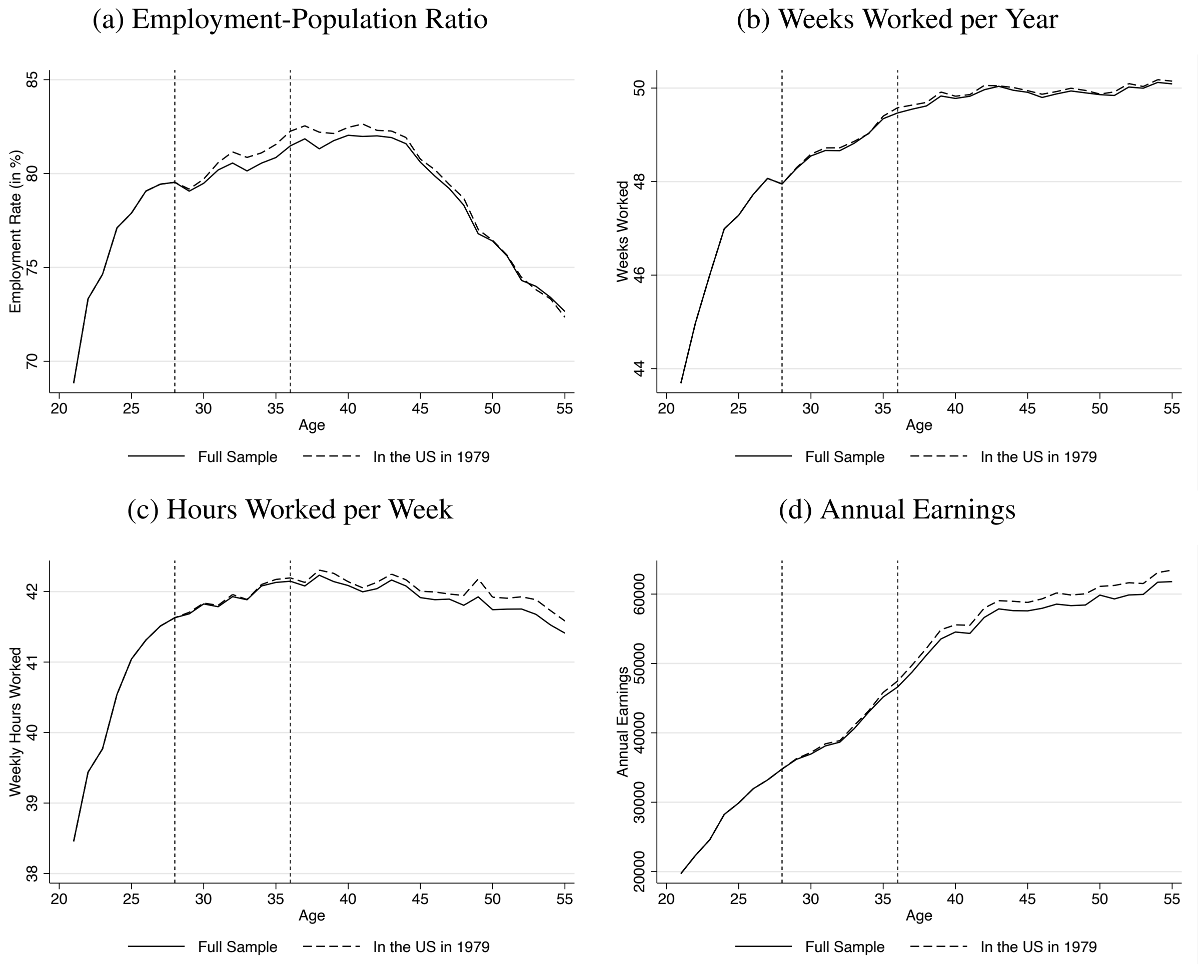
Given this difference in sample composition, Figure D.2 compares our main variables of interest for everyone born between 1957 and 1964 to only those who also were already in the US in 1979. In each figure, for all ages left to the first vertical line the year of immigration status is not available for anyone and for all right to the second vertical line it is available. We define someone as employed if they worked at least 520 hours per year (Figure D.2a), and report weeks worked (Figure D.2b), hours worked per week (Figure D.2c), and total annual earnigns (Figure D.2d) conditional on being employed according to this definition. We conclude that the gaps are relatively small between two samples. To achieve maximum consistency with the NLSY79, we opt for the sample which excludes individuals not living in the US prior to 1979 to the degree possible.

Figure D.1: Share of the 1957-1964 Cohorts Not Living in the US before 1979 in the ASEC



Notes: The first dashed line indicates the oldest age for which the immigration status is not known for any cohort and the second dashed line indicates the youngest age for which the immigration status is known for all cohorts.

Figure D.2: The Role of Immigration


(a) Employment-Population Ratio

(b) Weeks Worked per Year

(c) Hours Worked per Week

(d) Annual Earnings

Notes: The first dashed line indicates the oldest age for which the immigration status is not known for any cohort and the second dashed line indicates the youngest age for which the immigration status is known for all cohorts.

D.2 Custom Weights

In our main analysis we use the initial weights for weighting. These were constructed including the military and economically disadvantaged non-Black/non-Hispanic youths subsamples as well as individuals without an interview after age 21. We drop all of those individuals such that the initial weights are no longer necessarilly representative of the 1957-1964 cohorts residing in the US in 1979. In addition, our Lifetime Sample drops even more respondents because they do not have an interview after age 55, either because they have passed or stopped participating in the survey. The "Weight IDs" option on `https://nlsinfo.org/weights/nlsy79` allows to construct custom weights for the set of individuals in either of the two samples separately. Table D.1 shows that for our main variables of interest, the difference between using the initial or the custom weights is minimal.

Table D.1: The Effect on Custom Weights

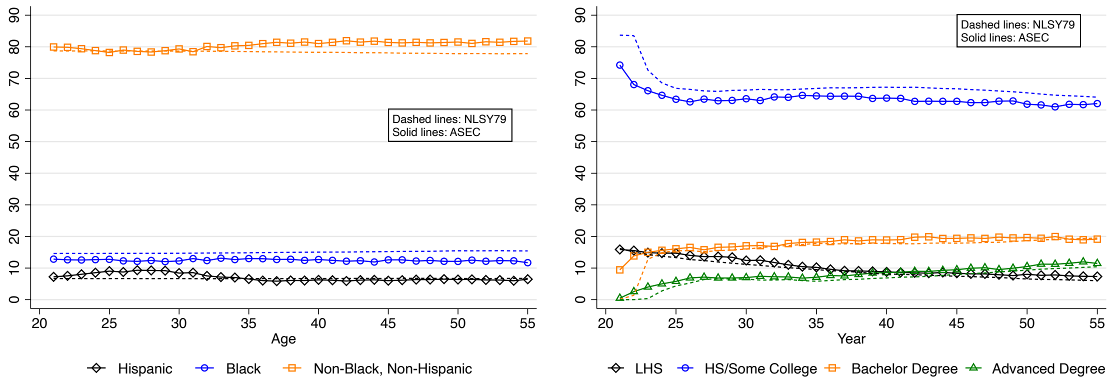
	Cross-Sectional Sample			Lifetime Sample		
	Initial Wgts	Custom Wgts	Diff	Initial Wgts	Custom Wgts	Diff
Employment Rate	81.0	80.6	-0.4pp	83.2	83.2	0.0pp
Annual Hours	2113.9	2113.8	0.0%	2118.0	2124.7	0.3%
Annual Earnings	49808.7	49511.4	-0.6%	51529.9	51717.6	0.4%
Hourly Wage	22.9	22.8	-0.6%	23.6	23.6	0.0%

Notes: The table reports the average over all individuals in the respective sample covering ages 21 to 55. Employment is defined as working at least 520 hours per year. All other variables are conditional on being employed.

D.3 Comparisons with the CPS

Figure D.3: Demographic Composition in the CPS ASEC and NLSY79 Cross-Sectional Sample

Gender

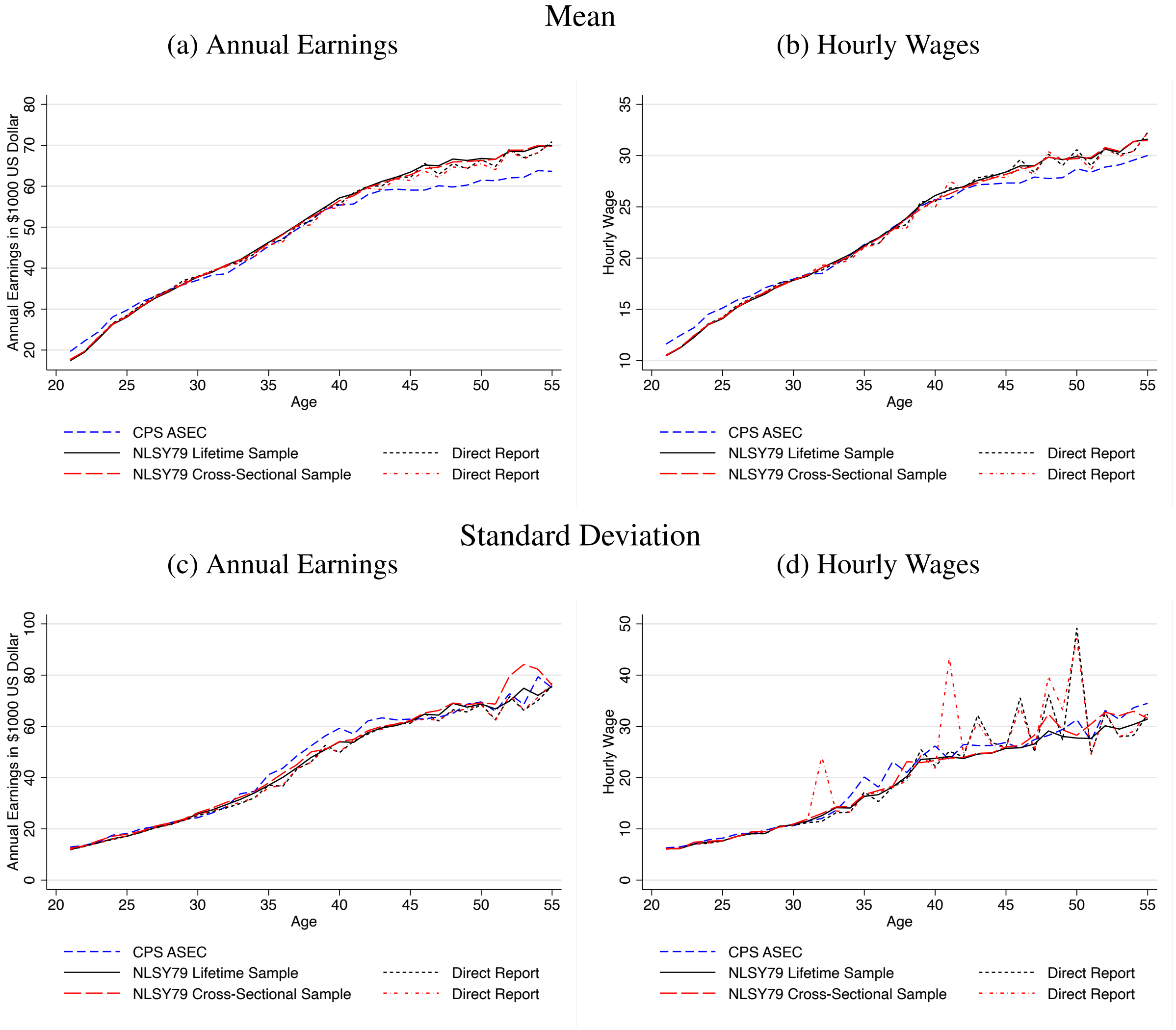


(a) Race Ethnicity

(b) Education

Notes: In 1992 the CPS changed how education was recorded. In Figure 5d through 1991 we classify individuals by their highest grade completed (LHS=completed at most 11 grase, HS/Some College=completed 12-15 grades, Bachelor Degree=completed 16 grades, Advanced Degree=completed 17 or more grades) and from 1992 onwards by their highest degree completed.

Figure D.4: Income from Wages and Salary



Mean

(a) Annual Earnings

(b) Hourly Wages

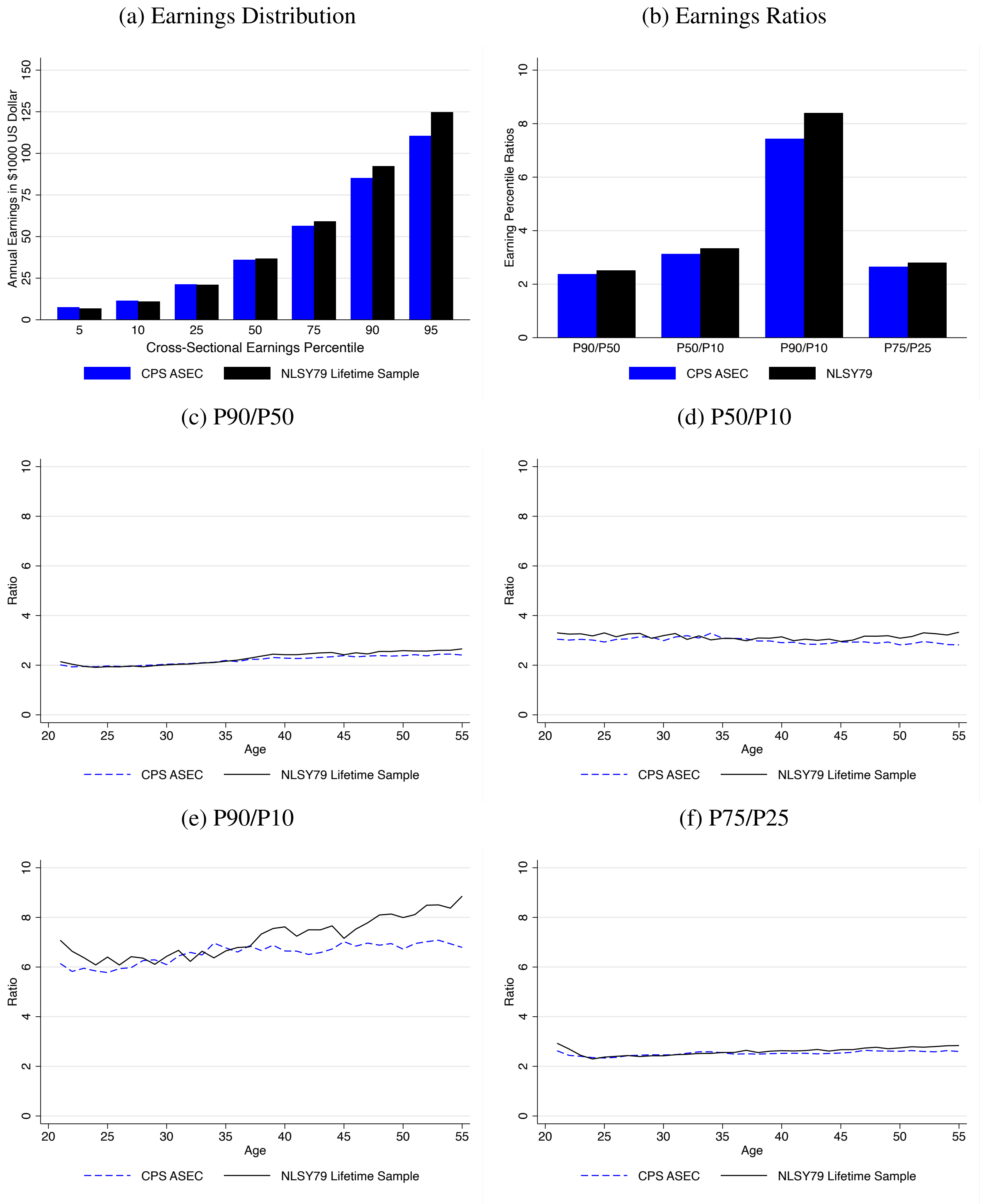
Standard Deviation

(c) Annual Earnings

(d) Hourly Wages

Notes: Only person-year observations for which annual hours are at least 520 are included.

Figure D.5: Cross-Sectional Earnings from Wages and Salary


(a) Earnings Distribution

(b) Earnings Ratios

(c) P90/P50

(d) P50/P10

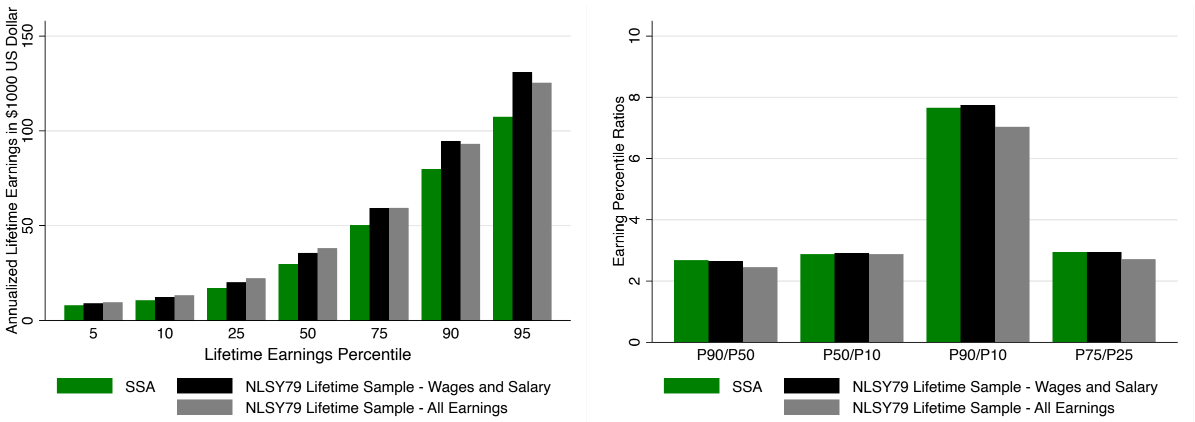
(e) P90/P10

(f) P75/P25

Notes: We first identify the individual at the respective percentile of the cross-sectional earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower cross-sectional earnings and the five individuals with the closest highest cross-sectional earnings. Only person-year observations for which annual hours are at least 520 are included.

D.4 Comparisons with the SSA Data

Figure D.6: Lifetime Earnings



(a) Lifetime Earnings Distribution

(b) Lifetime Earnings Ratios

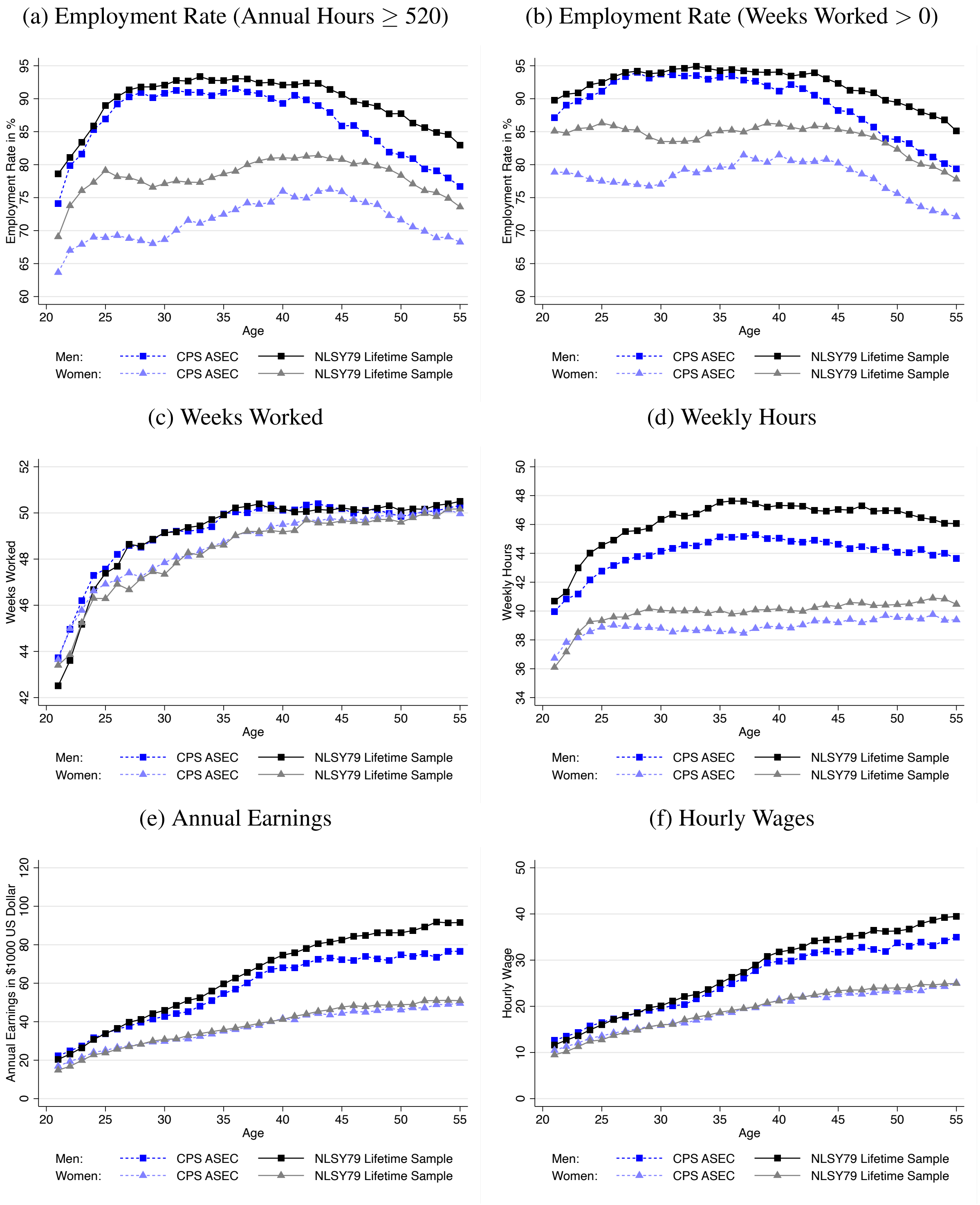
Notes: For the NLSY, we identify the individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower lifetime earnigns and the five individuals with the closest highest lifetime earnings.

D.5 Comparisons by Gender: Details

In the following, we show the same comparison graphs by gender as for the aggregate, mostly excluding those figures that we already show in the main text. For brevity, we focus on the Lifetime Sample with the imputed values for weeks worked, weekly hours, and earnings when applicable. For earnings and wages we only show the results based on total earnings. Again for brevity, we show for the Cross-Sectional Sampleonly the key labor market outcomes and do not provide more detailed earnings comparisons as these are so similar to the Lifetime Sample.

D.5.1 Lifetime Sample

Figure D.7: Means of Labor Market Variables by Gender



(a) Employment Rate (Annual Hours ≥ 520)

(b) Employment Rate (Weeks Worked > 0)

(c) Weeks Worked

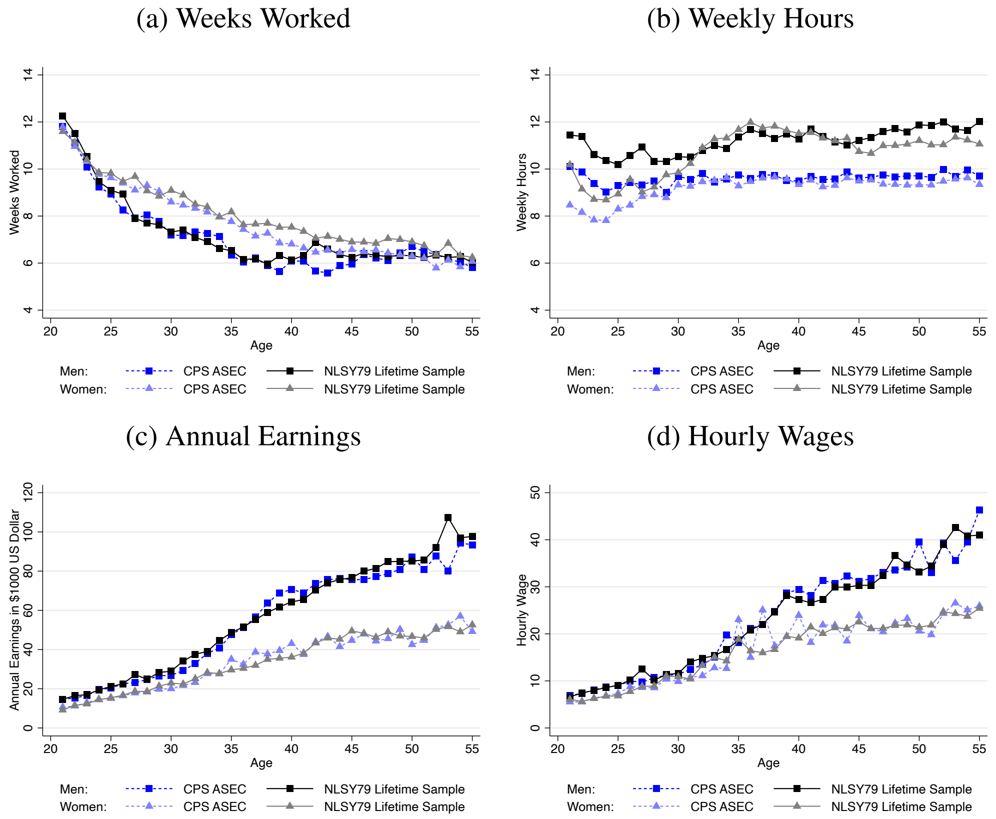
(d) Weekly Hours

(e) Annual Earnings

(f) Hourly Wages

Notes: In Figures D.7c to D.7f we condition on working at least 520 hours per year.

Figure D.8: Standard Deviations of Labor Market Variables by Gender



(a) Weeks Worked

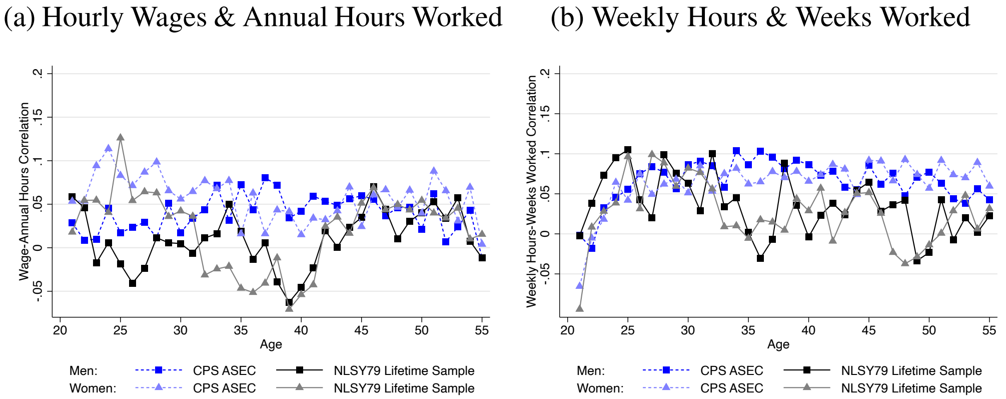
(b) Weekly Hours

(c) Annual Earnings

(d) Hourly Wages

Notes: Only person-year observations for which annual hours are at least 520 are included.

Figure D.9: Correlation of Labor Market Variables

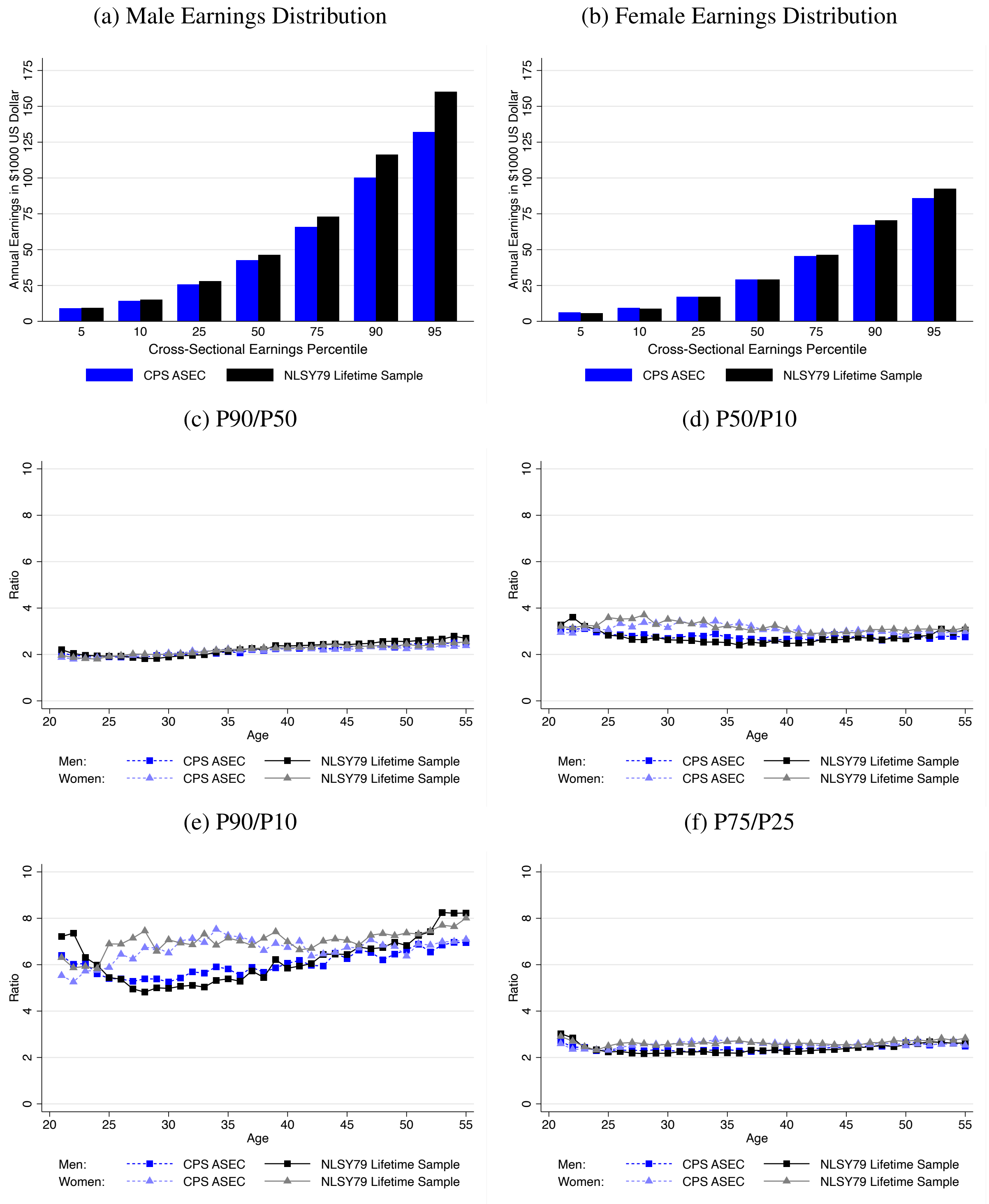


(a) Hourly Wages & Annual Hours Worked

(b) Weekly Hours & Weeks Worked

Notes: Only person-year observations for which annual hours are at least 520 are included.

Figure D.10: Cross-Sectional Earnings



(a) Male Earnings Distribution

(b) Female Earnings Distribution

(c) P90/P50

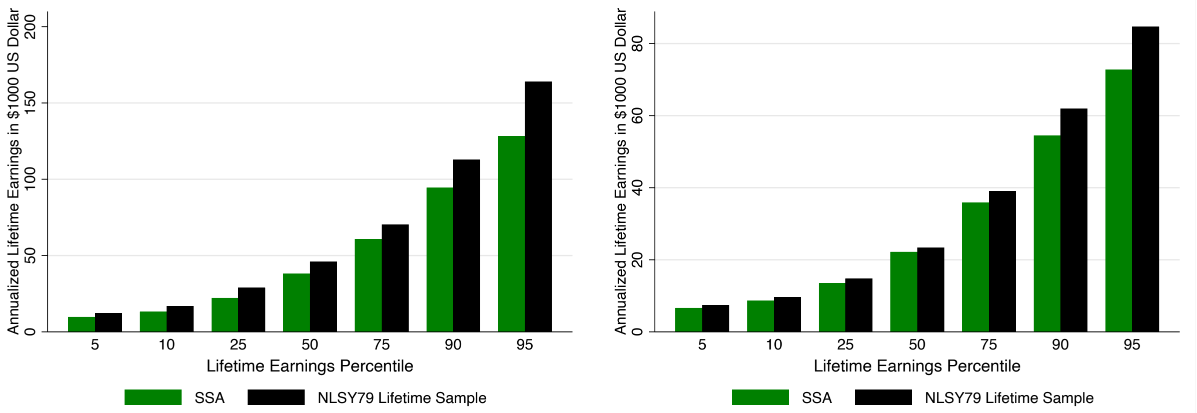
(d) P50/P10

(e) P90/P10

(f) P75/P25

Notes: We first identify the individual at the respective percentile of the cross-sectional earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower cross-sectional earnings and the five individuals with the closest highest cross-sectional earnings. Only person-year observations for which annual hours are at least 520 are included.

Figure D.11: Lifetime Earnings Distribution



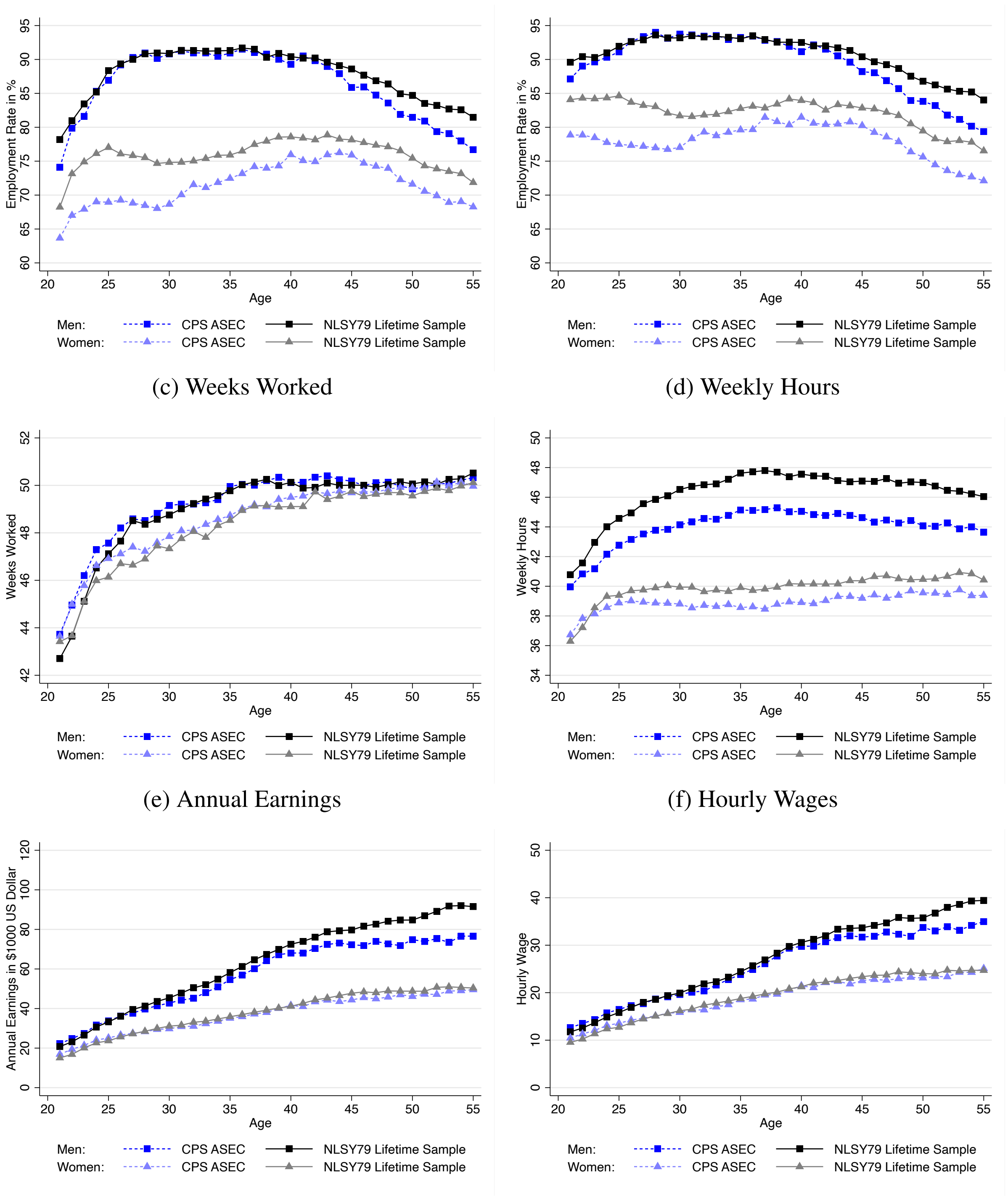
(a) Men

(b) Women

Notes: For the NLSY, we identify the individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution and then construct the value for each percentile by using the unweighted average of earnings of this individual and the five individuals with the closest lower lifetime earnigns and the five individuals with the closest highest lifetime earnings. The Social Security Administration (SSA) data are from Guvenen et al. (2022).

D.5.2 Cross-Sectional Sample

Figure D.12: Means of Labor Market Variables by Gender



(a) Employment Rate (Annual Hours ≥ 520)

(b) Employment Rate (Weeks Worked > 0)

(c) Weeks Worked

(d) Weekly Hours

(e) Annual Earnings

(f) Hourly Wages

Notes: In Figures D.12c to D.12f we condition on working at least 520 hours per year.

D.5.3 Lifetime and Cross-Sectional Sample Differenes for Men

Table D.2: Differences in Labor Market Variables between NLSY79 and CPS ASEC for Men

(a) Average Differences for Figures D.7 and D.12 (Mean of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Employment Rate (Annual Hours ≥ 520)	3.1pp	3.1pp	1.4pp	1.5pp
Employment Rate (Annual Hours > 0)	2.6pp	2.6pp	1.2pp	1.1pp
Weeks Worked	-0.2%	0.0%	-0.4%	-0.2%
Weekly Hours	4.9%	5.0%	5.2%	5.3%
Annual Earnings	11.9%	12.1%	10.1%	10.0%
Hourly wage	6.9%	7.3%	5.4%	5.7%

(b) Average Differences for Figure D.8 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	2.2%	1.1%	5.5%	3.5%
Weekly Hours	16.5%	16.5%	19.9%	19.9%
Annual Earnings	3.1%	3.6%	5.9%	3.5%
Hourly wage	-1.3%	1.5%	1.2%	5.7%

(c) Average Differences for Figure D.9 (Correlations)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Hourly Wages & Annual Hours Worked	-0.03	-0.04	-0.03	-0.04
Weekly Hours & Weeks Worked	-0.03	-0.03	-0.03	-0.04

Table D.3: Differences in Earning Ratios b/t the NLSY79 Lifetime Sample and the ASEC and SSA for Men

	Cross-Section		Lifetime
	Pooled (Figure 11e)	Average 25-55 (Figures D.10c-D.10f)	(Figure 11f)
P90/P50	6.9%	3.8%	0.0%
P50/P10	2.5%	-1.5%	-0.2%
P90/P10	9.6%	2.5%	-0.5%
P75/P25	2.1%	-0.9%	-0.3%

D.5.4 Lifetime and Cross-Sectional Sample Differenes for Women

Table D.4: Differences in Labor Market Variables between NLSY79 and CPS ASEC for Women

(a) Average Differences for Figures D.7 and D.12 (Mean of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Employment Rate (Annual Hours $\geq$ 520)	6.7pp	6.8pp	4.4pp	4.5pp
Employment Rate (Annual Hours $>$ 0)	6.0pp	5.7pp	3.9pp	3.6pp
Weeks Worked	-0.4%	-0.1%	-0.6%	-0.2%
Weekly Hours	2.5%	2.7%	2.5%	2.6%
Annual Earnings	2.5%	1.0%	2.6%	0.6%
Hourly wage	1.0%	0.5%	1.6%	0.8%

(b) Average Differences for Figure D.8 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	4.6%	2.7%	6.6%	4.5%
Weekly Hours	16.5%	16.4%	18.1%	18.0%
Annual Earnings	-0.9%	-7.5%	1.6%	-6.5%
Hourly wage	-1.8%	11.8%	3.6%	17.3%

(c) Average Differences for Figure D.9 (Correlations)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Hourly Wages & Annual Hours Worked	-0.04	-0.05	-0.04	-0.05
Weekly Hours & Weeks Worked	-0.04	-0.04	-0.05	-0.05

Table D.5: Diffs. in Earning Ratios b/t the NLSY79 Lifetime Sample & the ASEC/SSA for Women

	Cross-Section		Lifetime
	Pooled (Figure 11e)	Average 25-55 (Figures D.10c-D.10f)	(Figure 11f)
P90/P50	5.2%	2.8%	0.0%
P50/P10	4.0%	2.7%	0.0%
P90/P10	9.4%	5.3%	0.1%
P75/P25	2.6%	2.7%	0.0%