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

Alexander Bick, Adam Blandin, and Richard Rogerson

Abstract

In 1979, the National Longitudinal Study of Youth 1979 (NLSY79) began following a group of U.S. residents born between 1957 and 1964 and has continued to reinterview these same individuals for more than four decades. Despite this long sampling period, attrition remains modest. This article shows that after 40 years of data collection, the remaining NLSY79 sample continues to be broadly representative of their national cohorts regarding 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, the distribution of lifetime earnings over the age range 25 to 55 closely aligns with the distribution found 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

Federal Reserve Bank of St. Louis *Review*, First Quarter 2025, Vol. 107, No. 2, pp. 1-50.
https://doi.org/10.20955/r.2025.02

1. INTRODUCTION

The National Longitudinal Study of Youths 1979 (NLSY79) is a long-running panel dataset for the U.S. It began in 1979 by interviewing a group of U.S. residents aged 14 to 22 (born 1957 to 1964) and has continued to reinterview 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

Alexander Bick is an economic policy advisor at the Federal Reserve Bank of St. Louis. Adam Blandin is an assistant professor of economics at Vanderbilt University. Richard Rogerson is the Charles and Marie Robertson Professor of Public and International Affairs at Princeton University. The authors thank Kevin Bloodworth, Elizabeth Harding, and Siyu Shi for research assistance.

Michael Owyang and Juan Sánchez are editors in chief of the *Review*. They are supported by Research Division economists and research fellows, who provide input and referee reports on the submitted articles.

scientists. For example, between 2010 and 2023, the NLSY79 was used in at least 34 articles published in the "top 5" economics journals.[1]

Work on inequality has long understood the importance of distinguishing between the transitory and persistent components of inequality. Motivated by this understanding, recent work on inequality has used administrative data to document features of lifetime inequality (see, e.g., Guvenen et al. (2022), who use Social Security Administration (SSA) data). A key advantage of administrative data is the large sample size that they offer. However, there are also some disadvantages to relying on administrative datasets: Access to such data is extremely limited, especially in the U.S., and some variables of interest are typically not present. For example, because the SSA data do not include information on hours worked, they cannot distinguish between inequality in earnings and inequality in wage rates.

The NLSY79's long panel provides a publicly available dataset 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, Blandin, and Rogerson (2024) we use the sample constructed in this article to document the relationship between lifetime hours worked, hourly wages, and earnings.

However, the NLSY79 has two major disadvantages relative to administrative data like that from the SSA: a much smaller sample size and voluntary participation. Voluntary participation leads to two related practical issues: attrition and missing observations. While the initial sample was designed to be nationally representative, nonrandom 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 across different surveys. Discarding all individuals with any missing values would severely limit the sample size, which is particularly important when studying second-moment properties of the data, such as inequality. Given our interest in lifetime earnings inequality, we must impute these missing observations. This challenge is further compounded by a particular feature of the NLSY79: the earnings measure we rely on—comparable to the earnings reported in the Current Population Survey's (CPS) Annual Social and Economic Supplement (ASEC) and SSA data—is from 1994 onward only available for odd years.

The goal of this article 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 it is 7155 at age 55. Our second sample, which we refer to as the the lifetime sample, only includes individuals from the cross-sectional sample with at least one direct report at age 55 or older and an observation at each possible age from 21 to 55 (either a direct report or an imputation). This balanced sample comprises 6335 individuals.

For each of our two samples, we can also define subsamples in which we only use only direct reports, i.e., exclude person-year observations that rely on imputation. We show that the 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 nonrandom nature of attrition, we examine whether the NLSY79 is representative at each age over the life cycle. We first establish that attrition is approximately random across several observable dimensions: gender, race, and educational attainment. However, this does not rule out the possibility of 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 CPS's ASEC, also often referred to as the March CPS. We find that life-

1. We based this count on searching for the term "NLSY79" on the web pages of the *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 unaware of any other U.S. survey with comparable information that covers as many individuals for as long. For example, in the Panel Study of Income Dynamics, the sample that can be followed for a similar time period is less than 60 percent of the corresponding NLSY79 sample.

cycle means and standard deviations for these outcomes in all of our samples track those in the ASEC quite closely, though all of our samples 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 the results of 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, we plan to release a replication package including codes construct 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 will be easy to merge with any existing analysis, providing researchers confidence that their data work is based on a representative sample.

Our article is closely related to two existing papers that analyze attrition in the NLSY79. MaCurdy, Mroz, and Gritz (1998) find 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 article 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, Pierret, and Rothstein (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 article is also similar in spirit to Heathcote, Perri, and Violante (2010) and Heathcote et al. (2023), who analyze cross-sectional income, consumption, and wealth inequality across various U.S. datasets, also benchmarking them against National Income and Product Accounts and flow of funds data when possible.

The rest of the article proceeds as follows. Section 2 introduces the NLSY79 and our baseline sample and discusses how our key variables of interest are measured. 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, and Section 5 concludes.

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 U.S. residents aged between 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 oversamples. The initial respondents were reinterviewed 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.

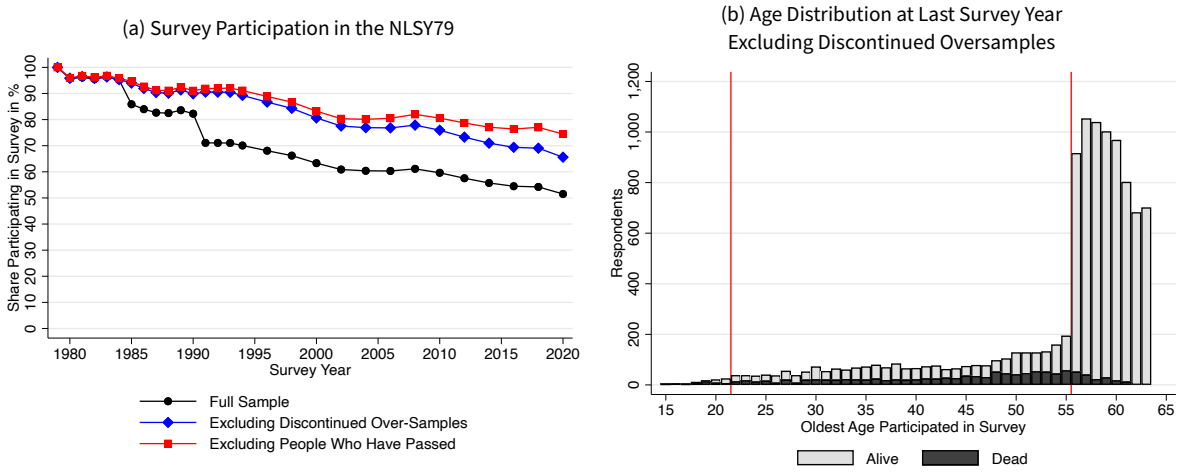
Figure 1a summarizes survey attrition between the 1979 and 2020 interview waves. The black line (circles) shows the share of all initial respondents who participated in the survey in a given year. There are two sharp reductions in survey participation. First, after 1984 the vast majority of a supplemental military oversample were discontinued. Second, after 1990 the economically disadvantaged non-Black/non-Hispanic youths supplemental oversample was discontinued. For the remainder of the article, we exclude these two discontinued supplements, leaving us with 9961 respondents.[5] The blue line (diamonds) in Figure 1a shows the share of these respondents who participated in the

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

4. Aughinbaugh and Gardecki (2007) document evidence of nonrandom attrition in NLSY97 but argue that it is difficult to assess its impact as their analysis only covers the first eight waves, with many sample members who had not yet, or just recently, completed their education.

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

Figure 1
Survey Participation and Age Distribution



NOTE: The discontinued oversamples refer to the military oversample and the economically disadvantaged non-Black/non-Hispanic youths supplemental oversample. The red vertical lines in panel (b) indicate the age range at time of the survey for our final sample.

survey in a given year. Among them, 65.6 percent participated in the 2020 survey. When excluding those who had died by 2020, the participation rate in the 2020 survey increases to 74.5 percent (see the red line (squares)). The high retention rate may be attributable to efforts by the NLSY79 to reinterview all initial respondents, even if they have missed several recent surveys.

Figure 1b shows the distribution of the oldest age a respondent participated in the survey. The youngest NLYS79 cohort is 56 in the most recent survey interview year (2020), which explains the stark difference in the share of respondents whose oldest age they participated the last time in a survey between 55 and 56. To cover the same age range for all respondents, we therefore restrict our analysis to individuals with at least one interview between ages 22 (the youngest age of the oldest NLSY79) and 56. Among the 9961 respondents in the previous paragraph, 9867 respondents have at least one interview between ages 22 and 56.[6] Since the measure of earnings used in our analysis refers to the year before the survey, this implies we have for each respondent in our sample at least one observation between ages 21 and 55.

2.2 Measurement

Since we are primarily interested in life-cycle facts, our results are typically presented by age, defined as the calendar year minus calendar year of birth. 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 NLSY79 creates a weekly employment variable for each individual-year. If an individual had a temporary employment gap during their tenure at a specific employer, it is reflected in the weekly employment arrays. In weeks without employment, an individual's employment status can either be nonemployed or unknown if no information was provided for that week.

For each individual-age, we also construct the average reported weekly hours worked and the number of work-weeks 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 method provides a weekly

6. Note that this sample includes individuals with at least interview at age 22 or later and who pass away or permanently drop out of the NLSY79 before reaching age 56.

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, 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 before an interview.[7] The NLSY79 collects this information for two 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) business/farm income received last year. Going forward, we refer to (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 participated in the survey.

2.3 Data Cleaning

For a small number of observations, we can address missing earnings information by using data on weeks worked or vice versa. Among person-year observations with zero weeks worked in a year and known employment status for the entire year, 1.4 percent have missing wage and salary income and 0.2 percent have missing business/farm income. We assign a value of zero to these missing earnings. Conversely, among person-year observations reporting zero earnings, 3.0 percent indicate zero weeks worked with the employment status missing for some, but not all, weeks of the year. For 5.8 percent, the employment status is missing for the entire year. In either case, we assign a status of nonemployment 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 percent of observations with positive income.[8] Similarly, 5.81 percent of observations with positive weeks worked report zero earnings. In both cases, we have three options: 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 of 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.

3. IMPUTATION OF MISSING VALUES 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 they participated but did not provide some information. We handle missing values in one of two ways: either drop an individual or impute a value using information from nearby interviews from that same individual. Loosely, our guiding principle is to drop individuals with long stretches of missing values that span several years but otherwise impute the missing values when nearby information is available. The rest of this section provides further details on the procedures we adopt.

3.1 How We Impute Missing Values

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). In contrast, earnings 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.

7. The NLSY79 also collects information to construct the weekly earnings for each job held since the last interview. 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.

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

3.1.1 Imputing Weeks Worked and Weekly Hours When Some but Not All Weeks Are 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 equals one if v equals x and zero otherwise.

Imputing Weeks Worked. When an observation has a missing employment status for at least one week but not all weeks of a year, 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 nonmissing weeks for each year t, $s_{i,t}$:[9]

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

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 nonmissing 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 are more recent and have fewer weeks with missing employment status.[10]

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 the hours worked for those weeks to missing. Note that the second summand is zero for observations where the employment status is known for the entire 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 the same procedure for weekly hours as we do for weeks worked, adjusting only the variable applied. The detailed formulas are available in Appendix 1.1.

3.1.2 Imputing Values for Full-Year Missing Data

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 we

9. A few years have 53 weeks, in which case we adjust the summations in equation (1) and all other equations in this section.

10. 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. Note that for the younger NLSY79 cohorts, we may also have observations before age 21 and for the older cohorts after age 55. However, to treat all cohorts symmetrically, we do not use this information.

cannot use data from year t itself, as it is entirely missing. Instead, we linearly interpolate these full-year missing values using the most recent prior year $\underline{t}$ 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 equation (4). We then impute missing earnings by multiplying the imputed wage by annual hours worked:

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

Before imputing wages, we make two adjustments to reported wages. First, we set a floor for hourly wages at half the federal minimum hourly wage, adjusting any wages below this threshold to match it. This affects 4.3 percent of person-year observations with nonmissing earnings. Among respondents who work at least one year, 36.2 percent have at least one year affected by this adjustment. Appendix 1.2 provides more details and a discussion.

Second, we flag and adjust unreasonably high-wage observations. Specifically, for all observations with hourly wages in the top 0.1 percent of the wage distribution, we set weeks worked and weekly hours to missing and impute them using the linear interpolation procedure in equation (4). We then impute wages using the original earnings and imputed annual hours via equation (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 worked.[11] We provide additional details on this topic in Appendix 1.3.

3.2 Sample Selection

The previous section described how we imputed missing values by linearly interpolating between the nearest prior and subsequent years with nonmissing 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 $[t - \overline{T}, t + \overline{T}]$.[12]

Lifetime Sample. Our initial sample (see Section 2.1) of 9867 respondents reduces to 9075 after removing individuals who died before turning 55, and it reduces 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 $\bar{T}$.

11. For example, among these top wage outliers, the average annual earnings are \$153424 compared with \$238285 in the top 99–99.9 percent of wage earners, and \$36509 in the bottom 99 percent of wage earners. Meanwhile, among these top wage outliers, the average annual hours worked are 222 compared with 1566 in the top 99–99.9 percent of wage earners, and 1966 in the bottom the top 99 percent of wage earners.

12. Appendix 2 provides a formal description of the criteria.

Table 1

Full-Year Missing Data: 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	6825	7045	7088	7117	7126	7133	7141	7149	7151
Annual Earnings	3070	4933	5707	6069	6335	6483	6625	6707	6779
% of Initial Sample	31.1%	50.0%	57.8%	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	82.5%	82.2%	81.5%	81.2%	80.9%	80.7%	80.6%	80.4%	80.3%
Avg. Annual Hours	2144	2114	2103	2099	2097	2096	2096	2095	2094
Avg. Annual Earnings	51299	48130	47231	46812	46610	46640	46587	46524	46519
Observations	107450	172655	199745	212415	221725	226905	231875	234745	237265

(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	82.4%	82.1%	81.3%	81.0%	80.7%	80.5%	80.4%	80.2%	80.1%
Avg. Annual Hours	2147	2118	2106	2102	2101	2099	2100	2099	2098
Observations	307279	307846	308055	308181	308256	308314	308363	308399	308424
Avg. Annual Earnings	51366	48198	47302	46883	46686	46718	46665	46602	46598
Observations	221350	230285	234031	236185	237627	238617	239372	239917	240338

NOTE: In panels 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.

How $\bar{T}$ Affects the Lifetime Sample. The lower panel of Table 1a shows how many 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: Increasing $\bar{T}$ from one to three years increases the sample size from 31.1 percent to 57.8 percent of the initial sample. This increase 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) decreases 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 individuals when $\bar{T}$ is low.[13] A second possible explanation is that individuals with lower attachment to the labor market 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 $\bar{T}$.

13. For example, consider hypothetical individuals A and B, who both participate in all surveys, for year $t < 1994$. Suppose that A is employed every year and B is employed every year except $t-1$ and $t+1$. Also suppose that 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 will remain in the sample for $\bar{T} \geq 2$. However, B will be dropped with $\bar{T}=1$ because earnings in year t could not be imputed as they did not work in $t-1$ or $t+1$.

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 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. Consequently, 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. As in the lifetime sample, mean employment, annual hours, and annual earnings in the cross-sectional sample decrease in $\bar{T}$.

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 finally, earnings.

3.2.1 Missing Weekly Employment Status

Of the total 250985 person-year observations in our initial sample, 4.2 percent have at least one week with missing employment status.[14] Figure 2a shows the distribution of weeks with missing employment status among person-year observations with at least one week of missing employment status. Most cases involve missing employment status for the entire year; a smaller yet still notable number miss only 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. We find that 39.6 percent respondents do not have a single week with missing employment status, while 24.2 percent have just one year with at least one missing week. Figure 2c shows the distribution of full-year missing employment, in which the weekly employment status is missing for the entire year. For 52.3 percent, this never occurs, while 26.2 percent have just one full year is 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.

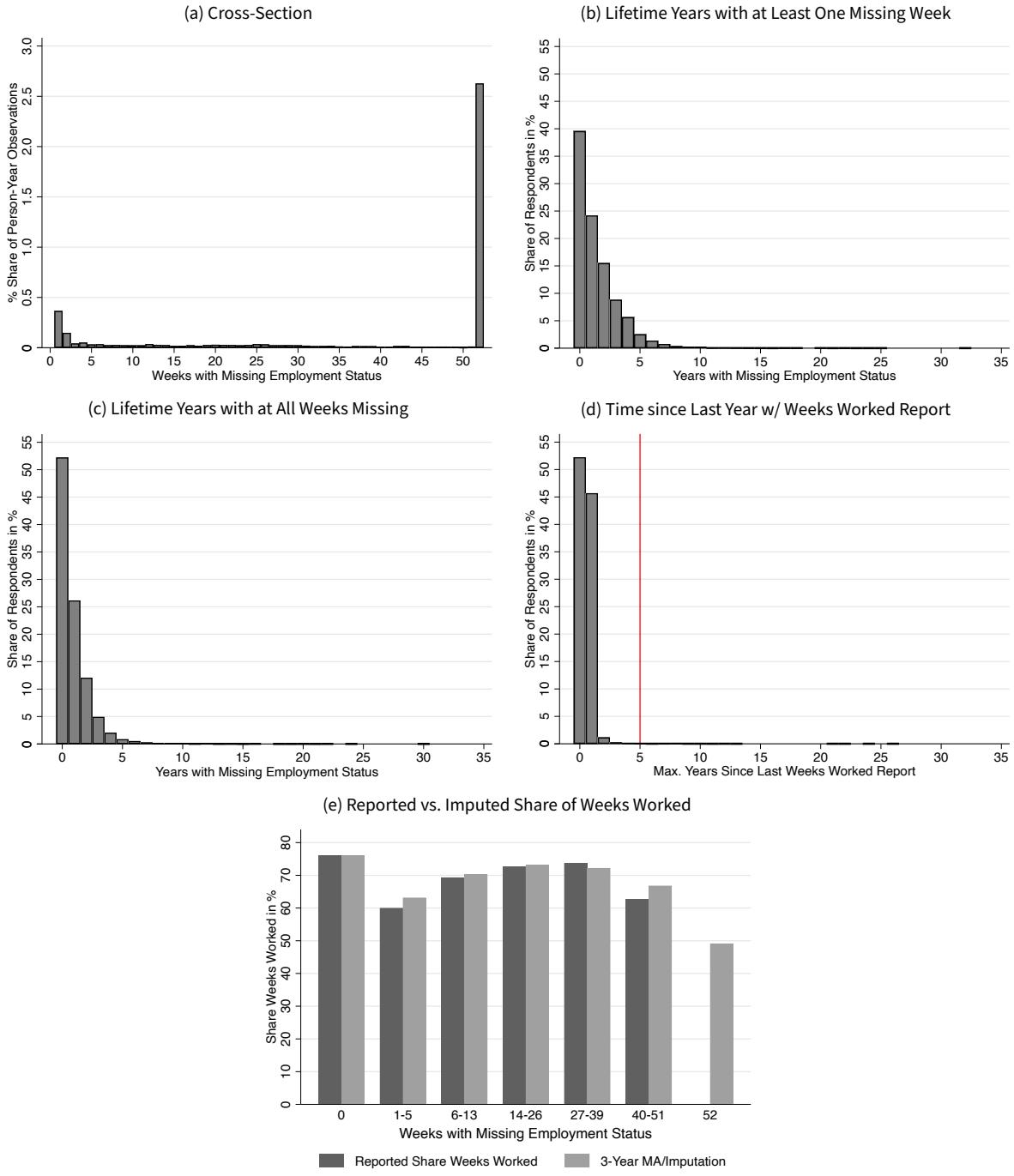
Figure 2e compares (i) the mean share of weeks worked among weeks with nonmissing employment (dark gray) with (ii) the mean imputed share of weeks worked among weeks with missing employment (light gray), shown separately for different bins of missing weeks worked. The main takeaway is that the reported and imputed share of weeks worked are similar. When the employment status is missing for all weeks (represented by the 52-hour bin), there is no reported value of weeks worked for comparison. In this case, the low mean imputed employment rate implies that individuals missing a full year of weeks worked 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 worked may be missing. Among the 7147 remaining individuals, there are 199296 person-year observations with positive weeks worked. Of these, 5.4 percent 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 unreported are in light gray, and those where the employment status is missing and weeks worked have been imputed (and thus weekly hours are missing by construction) are in dark gray. Cases with missing hours are heavily concentrated on missing either one or all workweeks in a year.

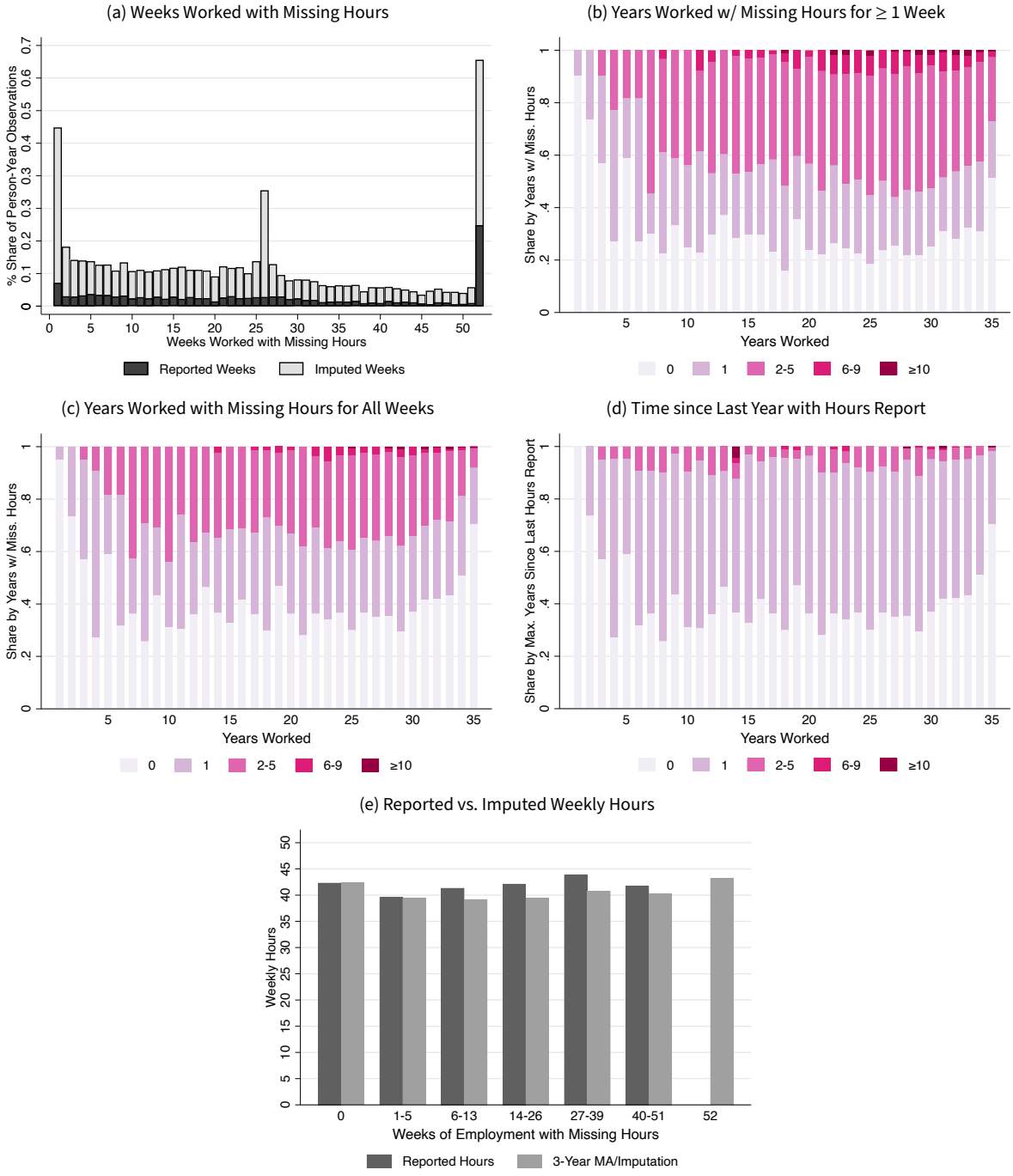
14. Figure Appendix 3.1 shows that years with the employment status missing for at least some weeks are much more prevalent if an interview was missed that year.

Figure 2
Missing Employment Status



NOTE: In Figures 2a to 2d, we exclude the respective group with no missing employment status. For illustrative purposes, in 2a and 2e, years in which 53 weeks are worked, but employment status for these 53 weeks are missing, are included in the 52-week bin.

Figure 3
Missing Weekly Hours



NOTE: In this figure, a year of work is defined as a year with positive weeks worked. In Figures 3a to 3d, we exclude the respective group with no missing employment status. For illustrative purposes, in 3a and 3e, years in which 53 weeks are worked, but hours for these 53 weeks is missing, are included in the 52-week bin.

Figure 3b takes a lifetime perspective. We group workers by their years worked (the horizontal axis) and 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 percent do not have a single year with some missing hours, and 21.6 percent have only one year with some missing hours. Figure 3c is similarly constructed except the vertical axis now represents years worked with no hours information at all. The main takeaway from both 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. We drop anyone with more than five years between two such observations, leaving us with a sample of 7126 individuals.

Figure 3e compares mean hours worked among nonmissing weeks (dark gray) with mean imputed hours worked among missing weeks (light gray), shown 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.

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 from farm/business income. 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 nonmissing source of earnings.

Unlike weeks worked and hours worked, which are in principle collected retrospectively for all jobs since the last interview, our earnings measure is only available for the reference year with an interview. Among the 7126 remaining individuals, we have 134195 person-year observations in a reference year in which an individual worked and took the interview. Among those, 7.1 percent have missing earnings.

From a lifetime perspective, Figure 4a shows that 44.4 percent 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 noninterview years. This distribution includes (i) years where an individual was interviewed but did not report earnings, (ii) interview years where they did not participate in the interview, and (iii) noninterview years.[15] Focusing on the last case, recall that after 1994, the NLSY79 switched to a biennial schedule. Depending on their birth cohort, individuals who worked every year automatically have missing earnings in at least 10–13 years simply due to the switch. The mass of years worked with missing earnings is concentrated in precisely this range: 54.2 percent of our remaining respondents have between 10 and 13 years of missing earnings. Note that some individuals have fewer years of missing earnings because they did not work in all noninterview years.

Figure 4c shows how the distribution of years with missing earnings varies according to years worked. For example, as previously described, every worker who works for all 35 years has at least 10 years of missing earnings due to the biennial switch. However, few of these workers have more than 15 years of missing earnings.

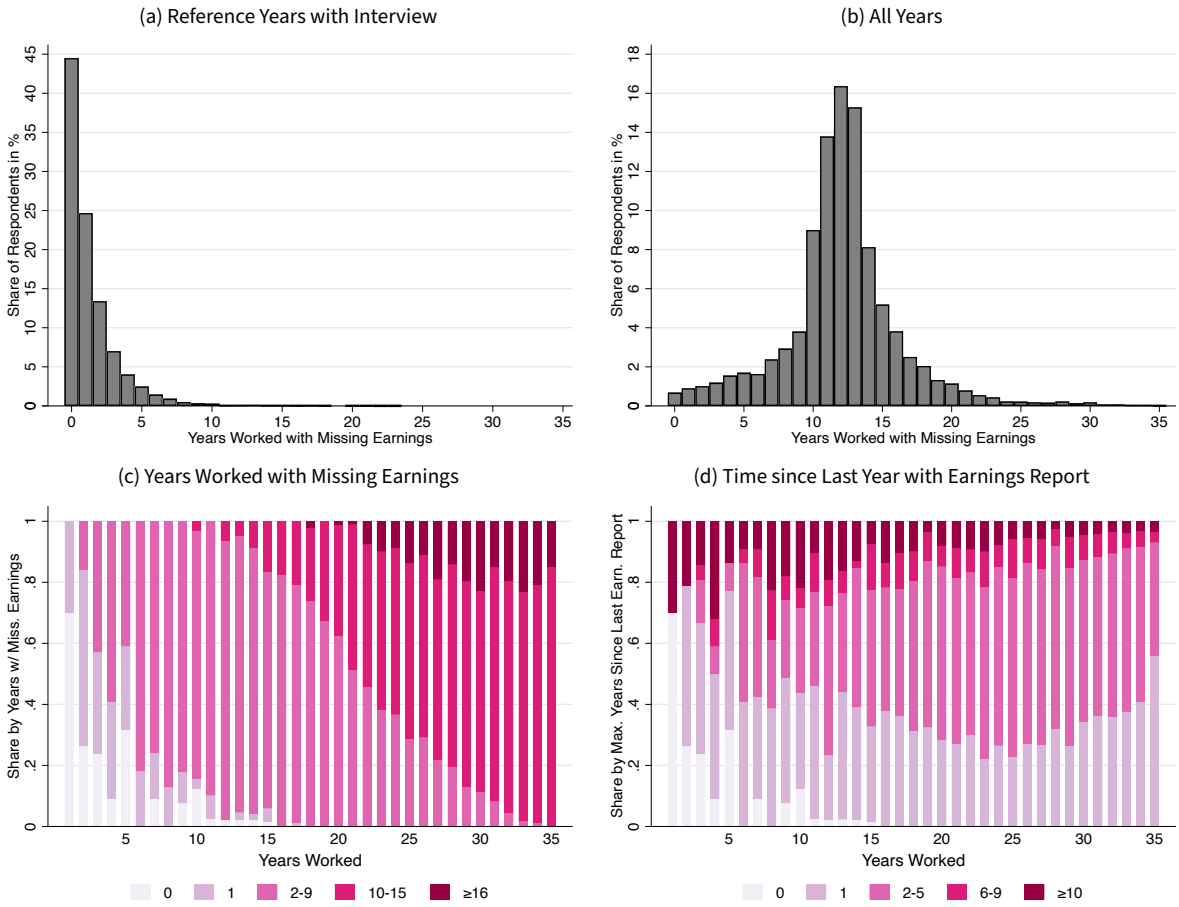
Figure 4d reports the largest gap between a year worked with missing earnings and the nearest year worked (either prior or subsequent) with an earnings report, conditional on the number of years worked. We drop anyone with more than five years between two such observations, leaving us with a lifetime sample of 6335 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, and 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 percent. This discrepancy is attributable to a wage-hours correlation that is somewhat higher in the imputed data compared with direct reports.

The last two columns of Table 2 repeat the exercise for income solely from wages and salary. These results are highlighted because they provide a relevant measure for comparison of lifetime earnings with Social Security earnings histories from Section 4.3. The conclusions are essentially the same as those for total earnings, partly because income from wages and salary is by far the dominant source of total earnings (see Appendix 1.4 for details).

15. Figure Appendix 3.2 shows the prevalence of missing earnings in reference years, distinguishing (i) and (ii).

Figure 4
Years Worked with Missing Earnings



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

Table 2
In-Sample Fit of Imputed Wages and Earnings

	Total Income		Wages and Salary	
	Report	Imputation	Report	Imputation
Hourly Wage	20.20	20.18	19.71	19.63
Annual Earnings	40193.6	42343.6	39041.2	40839.0
Wage-Annual Hours Correlation	0.01	0.15	0.01	0.16

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 CPS's ASEC, which we obtain from the IPUMS CPS database by Flood et al. (2023). Crucially, the ASEC collects annual-level information on labor market outcomes that correspond to measures in the NLSY79. In particular, weeks worked include weeks with only a few hours or paid time off, and annual earnings are reported separately for income from wages and salary and 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 and 1964.[16] We further condition on having lived in the U.S. in 1979 for those who were not born in the country. (Because this criterion is only fully observable in the ASEC beginning in 1994, we do not impose it before then; Appendix 4.1 shows that after 1994, this criterion does not substantially change our results.)

Going forward, all statistics reported for the NLSY79 use the initial 1979 cross-sectional sample weights. In Appendix 4.2 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 the ASEC (solid lines). Figure 5a shows a very similar share of men and women by age in the ASEC and the NLSY79 cross-sectional sample. In both datasets the share of women increases with age due to differential mortality. Figure 5b displays the same data from the ASEC, now presented alongside the NLSY79 lifetime sample (the constant sex distribution is due to the balanced panel). We can see that women are slightly overrepresented 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 overrepresentation of Black respondents and slight underrepresentation of non-Hispanic/non-Black respondents in the NLSY79. Figure 5d compares education shares. During the early 20s, the share of having completed a bachelor's 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 high school/some college individuals, and the ASEC has a slightly higher share of individuals who did not graduate from high school.

To summarize, 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.[17] In almost all cases, imputations have a minimal impact on our results. Therefore, our discussion will primarily 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.

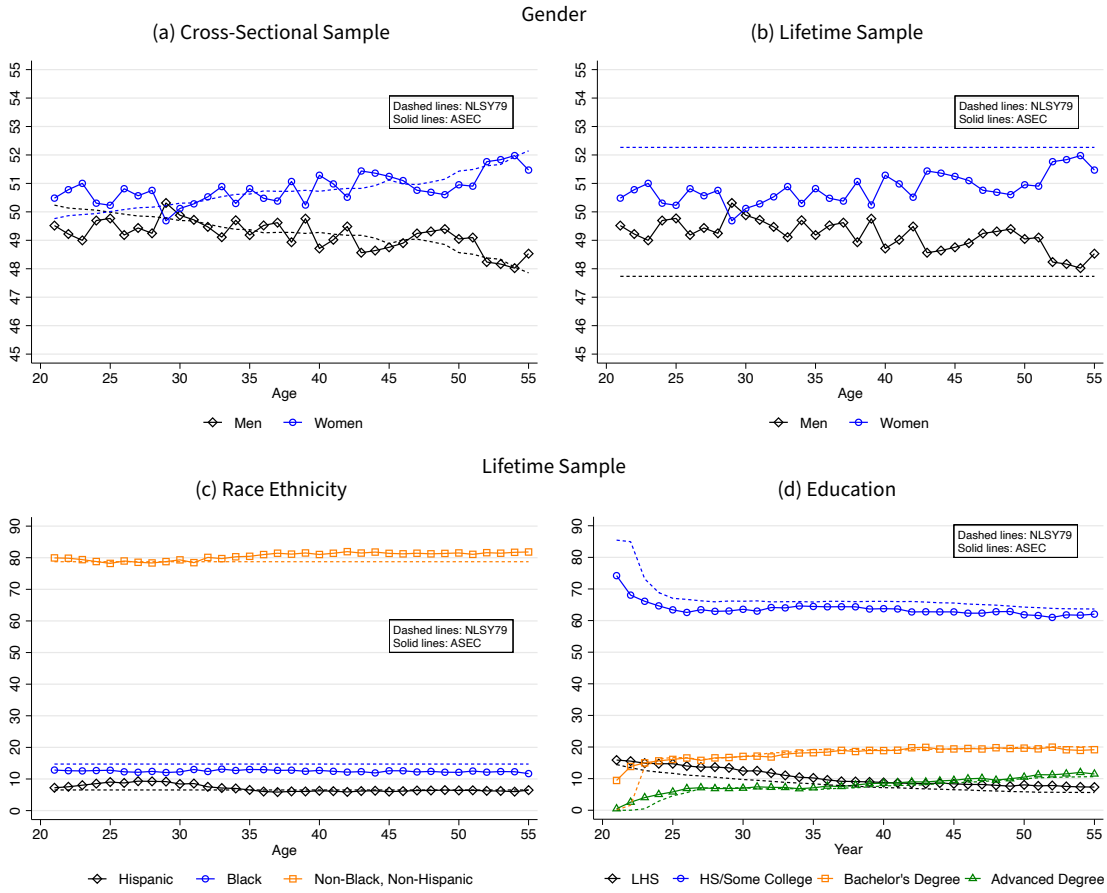
Figure 6a displays employment rates, defined as having worked at least 520 hours in a year. First, the employment rate in the NLSY79 cross-sectional sample is higher than in the ASEC. At age 21, the employment rate in the cross-sectional sample is 4.4 percentage points higher than in the ASEC. The gap narrows until the mid-40s and then widens again; on average, it is 3.0 percentage points, and by age 55, it is 4.1 percentage points. We emphasize that this disparity is not attributable to nonrandom sample attrition in the later years of the survey since it is apparent even at very early ages. Second, the lifetime sample is further positively selected based on 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

16. As in the NLSY79, we calculate the birth year as the survey year minus age at the time of the survey.

17. For the directly reported data, we make the following adjustments. We set weeks with missing employment status to nonemployment. In situations where hours are available for only a subset of weeks worked, we use 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.

Figure 5

Demographic Composition in the ASEC and NLSY79



NOTE: In 1992 the CPS changed how education was recorded. Until 1991, as shown in Figure 5d, we classify individuals by their highest grade completed (LHS = completed at most 11 grades, HS/some college = completed 12–15 grades, bachelor's degree = completed 16 grades, advanced degree = completed 17 or more grades) and from 1992 onward by their highest degree completed.

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 ≥ 520)	4.7pp	4.8pp	2.9pp	3.0pp
Employment Rate (Annual Hours > 0)	4.1pp	4.0pp	2.5pp	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	3.8%	3.8%	3.5%	3.4%

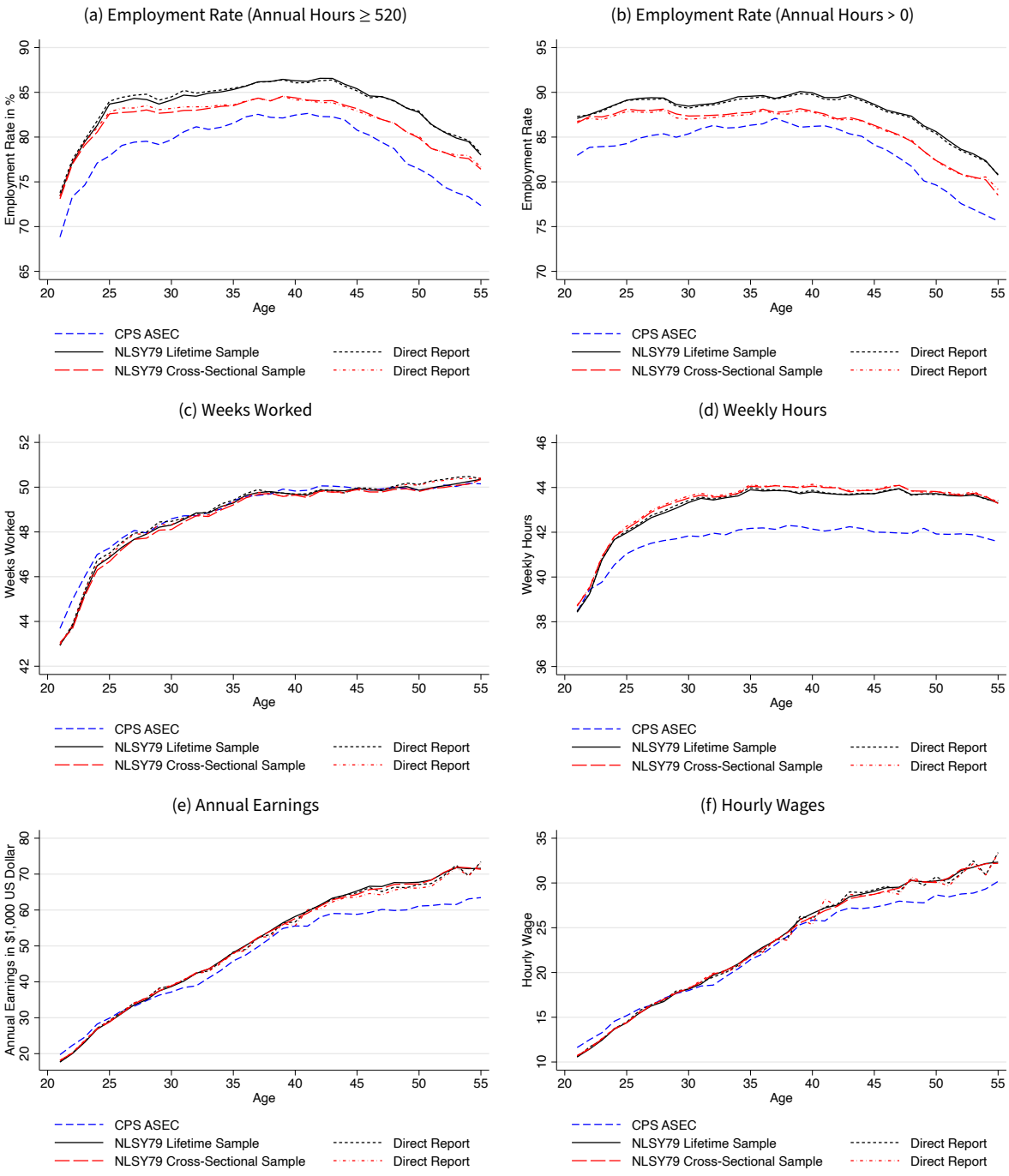
(b) Average Differences for Figure 7 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	4.1%	2.3%	6.3%	4.2%
Weekly Hours	16.8%	16.7%	19.5%	19.4%
Annual Earnings	2.5%	1.6%	5.4%	2.1%
Hourly wage	-1.4%	6.9%	1.8%	11.4%

(c) Average Differences for Figure 8 (Correlations)

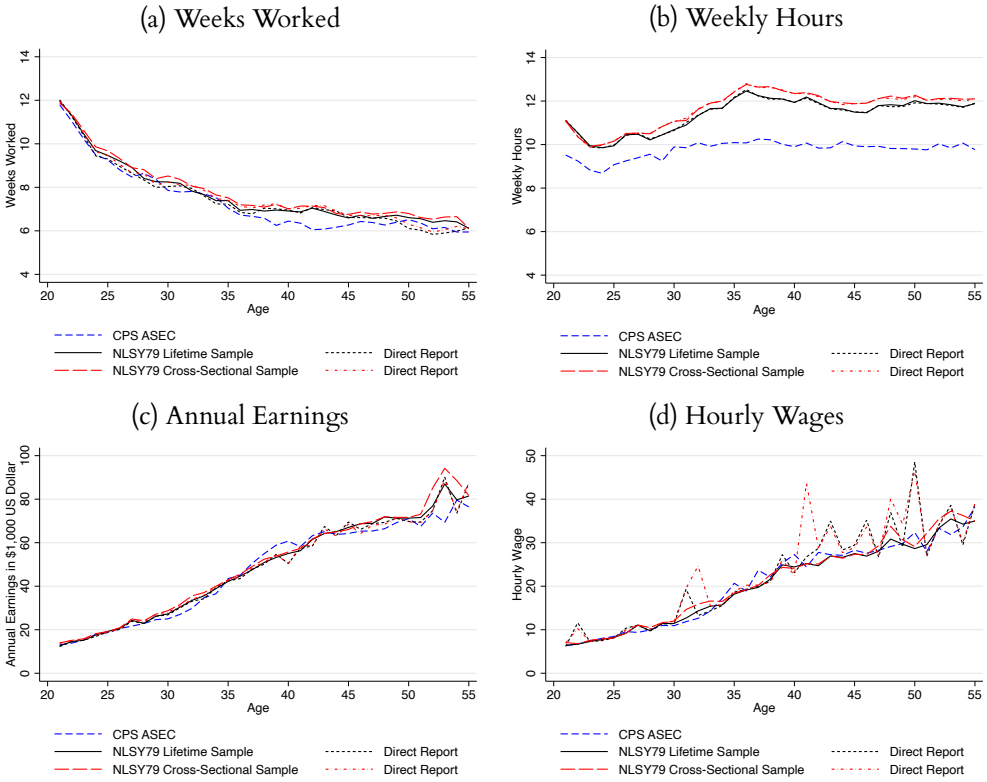
	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Hourly Wages & Annual Hours Worked	-0.02	-0.03	-0.02	-0.03
Weekly Hours & Weeks Worked	-0.03	-0.04	-0.04	-0.04

Figure 6
Means of Labor Market Variables



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

Figure 7
Standard Deviations of Labor Market Variable



NOTE: We include only those person-year observations where annual hours worked are 520 or more.

observation that cannot be imputed. Third, 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 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, 1.5 (3.5 percent) higher in the lifetime sample than in the ASEC (see Figure 6d). The largest gap is 2.0 hours (4.7 percent). Figures 6e and 6f displays life-cycle profiles for mean annual earnings and hourly wages.[18] (Appendix 4.4a and Appendix 4.4b report the corresponding statistics for income from wages and salary only.) Before age 40, mean earnings in the two datasets track each other quite closely, with a slightly higher growth rate in the NLSY79. This finding aligns with those of MaCurdy, Mroz, and Gritz (1998), who analyze 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 12.9 percent higher than in the ASEC, partly driven by the higher hours per week in Figure 6d. However, Figure 6f also shows a similar pattern in mean hourly wages: By age 55, they are 7.1 percent 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

18. 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.

Figure 8

Correlation of Labor Market Variables

NOTE: We include only those person-year observations where annual hours worked are 520 or more.

Table 4

Differences in Earning Ratios between 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	6.3%	3.7%	0.2%
P50/P10	4.7%	2.9%	-0.2%
P90/P10	11.3%	6.8%	0.1%
P75/P25	4.8%	2.3%	0.0%

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 datasets. Figure 7b shows that the standard deviation of weekly hours worked is higher in the NLSY79 than in the ASEC. Both of 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 the NLSY79 closely align with those in the ASEC. In particular, while the means in the NLSY79 are somewhat higher than in the ASEC at later ages, this is not reflected in different standard deviations. These figures represent one instance where 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 nonemployment for the directly reported data, the implied hourly wages are very high and erratic.[19]

Finally, Figure 8 shows that the NLSY79 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 8 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 versus 0.08 in the ASEC), and in both there is little systematic variation over the life cycle.[20] Figure 8 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 versus 0.07 in the ASEC). Additionally, in both datasets the correlation increases from slightly negative to slightly positive throughout the early 20s and then gradually declines for the remainder of the life cycle.

In summary, the major discrepancies in labor market outcomes between the NLSY79 and the ASEC include an

19. In fact, for expositional purposes, we cap the maximum directly reported wage in the NLSY79 at twice the maximum of the imputed wages. Without this cap, the last spike at age 50 would be about twice as large for both direct reports.

20. In Bick, Blandin, and Rogerson (2022), we show that the nonlinear relationship between hourly wages and weekly hours worked in the NLSY79 and the ASEC (and other standard datasets) are very similar.

employment rate that is 4.7 percentage points higher on average, weekly hours worked that are 3.5 percent higher on average, mean hourly wages that are 2.6 percent on average, and annual earnings that are 5.5 percent higher on average (all these gaps refer to the lifetime sample, for which the gaps relative to the ASEC are the largest). These discrepancies, however, do not affect the shape of the life-cycle profiles. In 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 1957–64 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 presents various percentiles of the earnings distribution in the NLSY79 and the ASEC. The gap between the two datasets increases as we move up the earnings percentile range. Thus, the upper part of the earnings distribution explains why mean earnings in the NLSY79 exceed those in the ASEC, as documented in Figure 6e.

Figure 9b displays four commonly used earnings percentile ratios in the earnings inequality literature: 90/50, 50/10, 75/25, and 90/10. The deviations range from 4.7 percent to 6.3 percent for the first three ratios and are naturally largest for the P90/P10 ratio with 11.3 percent, see the first column of Table 4.

Figures 9c to 9f show that the four inequality measures in the NLSY79 align closely with their ASEC counterparts over the life cycle. The NLSY79 shows higher inequality at the very beginning of the life cycle and from the late 30s onward, which is also around the time when mean earnings in the NLSY79 continue to grow much faster than in the ASEC (Figure 6e). Notably, the difference between the NLSY79 and the ASEC averages over the life cycle for each ratio is substantially smaller than when the data are pooled (compare columns one and two of Table 4).

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), who use SSA data on W2 wage and salary earnings to document lifetime earnings for several U.S. 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 ages 25 to 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 the results from our lifetime sample; to maximize our sample size, we include all 1957–64 birth cohorts.

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 they (i) survived until at least age 55; (ii) had earnings larger than a year-specific threshold level, denoted by Y_t, for 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 they were 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.[21,22]

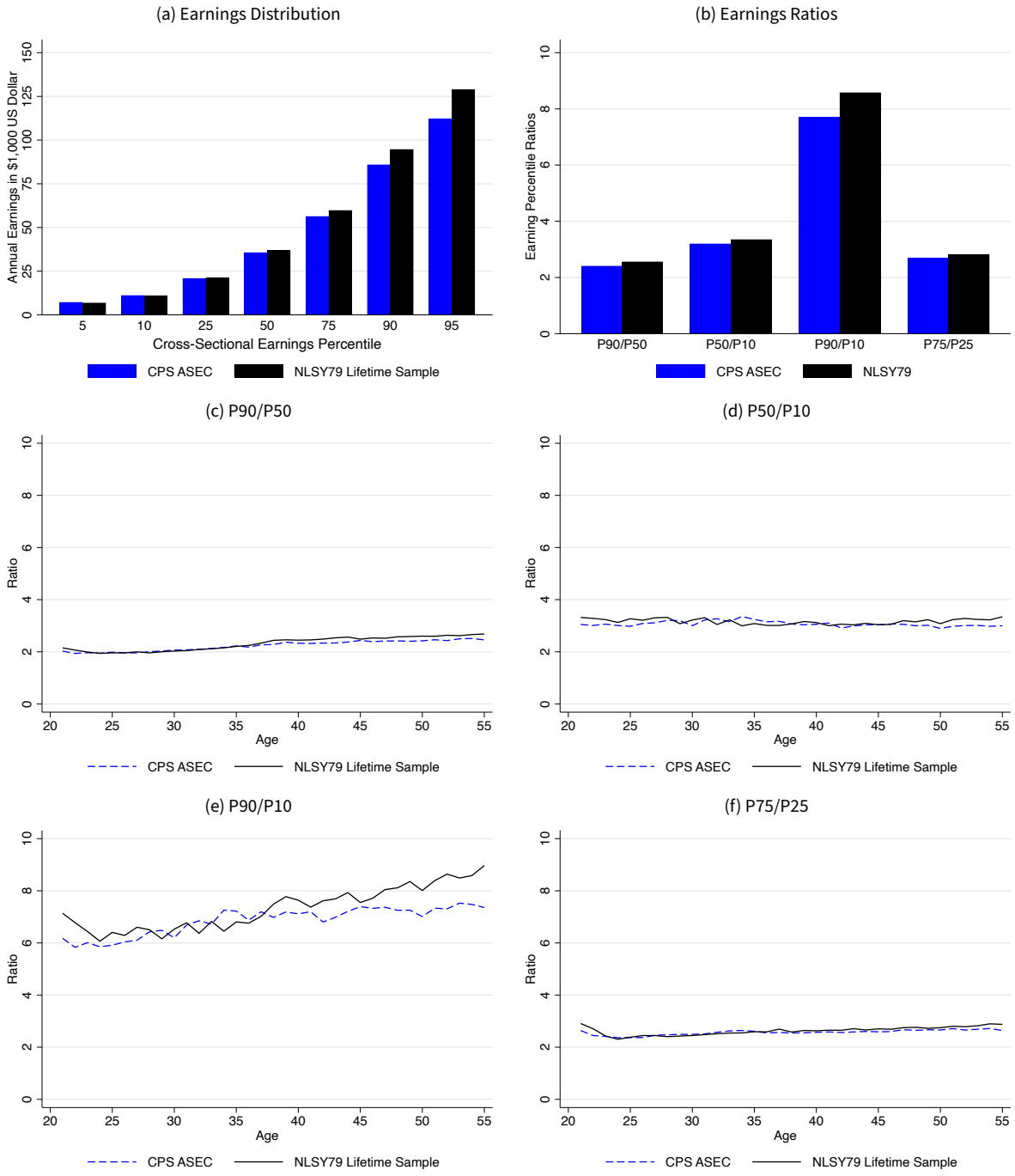
Also following Guvenen et al. (2022), we construct lifetime earnings as the average over those earnings without discounting. The NLSY79 provides information on the industry of the most recent job until 1993, and from 1994 onward, we use information from the first job reported for that year. For years of employment with missing industry information, we use the last reported industry. Out of our final lifetime sample of 6335 individuals, 3963 satisfy these criteria.

Figure 10a displays lifetime earnings percentiles in the NLSY79 and in Guvenen et al. (2022). Lifetime earnings

21. In Appendix 1.4 we explain how we impute missing earnings from salary and wages, which we do without conditioning on an indiviual's industry.

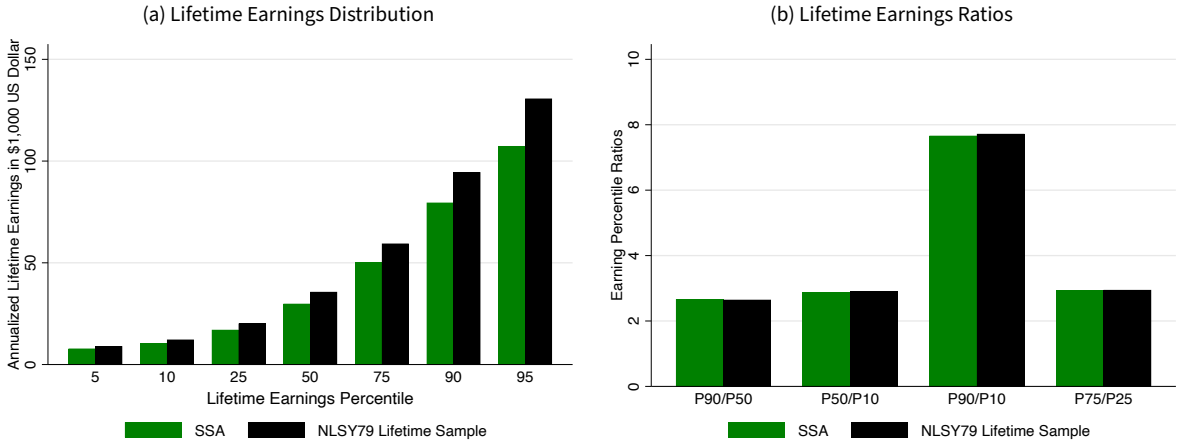
22. Guvenen et al. (2022) define "commerce and industry" workers to include those in all industries except agriculture, forestry and fishing, hospitals, educational services, social service, religious organizations and nonclassified membership organizations, private households, and public administration.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

Figure 9
Cross-Sectional Earnings



NOTE: 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 10
Lifetime Earnings



NOTE: For the NLSY79, we first identify individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution. To calculate the value for each percentile, we then use the unweighted average of earnings of the identified individual, 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).

are, on average, 14.8 percent higher in the NLSY79, with the magnitude differing across the distribution.[23] Figure 10b compares the earnings ratios in both datasets, and they are remarkably similar (also see the last column of Table 4).[24] We conclude that the mean of lifetime earnings in the NLSY79 is modestly higher than in the SSA, but the inequality in lifetime earnings is very similar.

4.4 Comparisons 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 4.5. Overall, our comparisons by sex yield similar results, though discrepancies are occasionally larger for one sex than the other.

Figure 11a shows larger average employment rate gaps for women (on average, by 6.6 percentage points) than men (on average, by 3.0 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 that in the ASEC by 5.3 percentage points for women and 6.2 percentage points for men.

Figure 11b shows that, conditional on being employed, men work 103.7 (4.8 percent) more hours per year in the NLSY79, on average. For women, the gap is with 38.6 (2.0 percent), somewhat smaller than for men. For both men and women, the gaps are again entirely driven by higher weekly hours worked, while weeks worked in both datasets are almost identical (see the figures in Appendix 4.7c and Appendix 4.7d).

Figures 11c and 11d show that the aggregate gap in mean earnings and wages is almost entirely driven by higher mean earnings and men's wages. By age 55, men's earnings in the NLSY79 exceed those in the ASEC by 11.9 percent, while this difference is only 2.6 percent for women.[25]

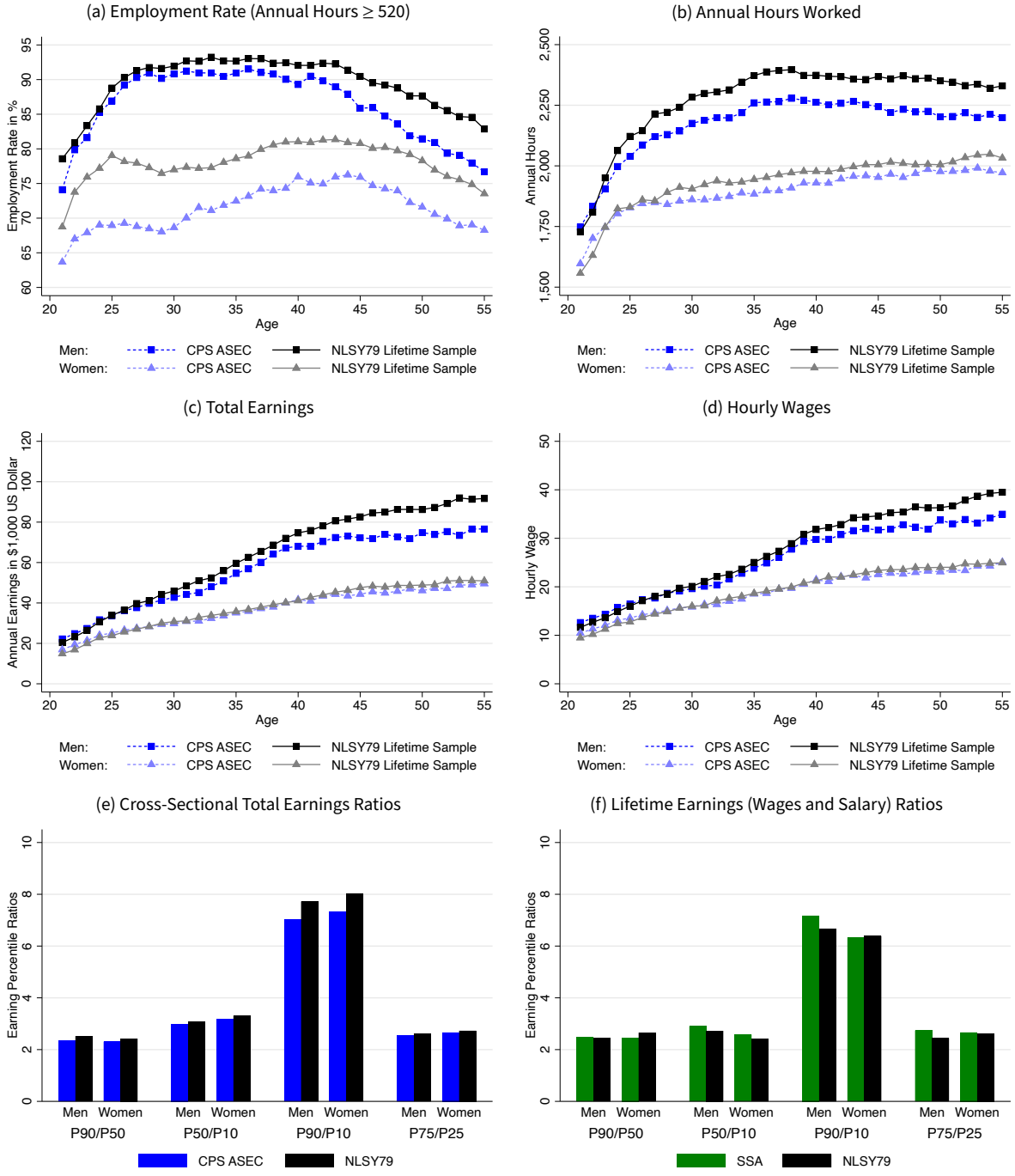
Finally, Figures 11e and 11f show that earnings percentile ratios also align fairly closely with the ASEC and the SSA data for both men and women. Cross-sectional earnings inequality is slightly higher in the NLSY79 than in the ASEC, both for men and women. Lifetime earnings inequality is slightly lower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

23. 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 11.1 percent larger than in the SSA data but are 16.1 percent higher for the six youngest cohorts.

24. Appendix 4.6 expands the analysis to include 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.

25. Goldin, Pekkala Kerr, and Olivetti (2023) show that the gender earnings gap for college graduates in the NLSY79 closely aligns with that of the Longitudinal Employer-Household Dynamics, a U.S.-linked administrative employer-employee database.

Figure 11
Key Results by Gender



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

5. CONCLUSION

For more than 40 years, the NLSY79 has reinterviewed a sample of individuals born between 1957 and 1964. When excluding two supplemental samples that were (almost entirely) discontinued early in the survey, 74.5 percent of the surviving initial respondents participated in the most recent available survey, conducted in 2020–21. In this article, we document that demographics and labor market outcomes in the NLSY79 remain broadly representative of their birth cohorts, from both cross-sectional and lifetime perspectives.

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 U.S. To facilitate these analyses, we plan to release a replication package providings codes 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 will be easy tomerge with any existing analysis, providing researchers confidence that their data work is based on a representative sample.

REFERENCES

Aughinbaugh, Alison, and Rosella 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, Alison, Charles R. Pierret, and Donna S. Rothstein. 2017. *Attrition and Its Implications in the National Longitudinal Survey of Youth 1979.* Statistical Survey Paper. Bureau of Labor Statistics.

Bick, Alexander, Adam Blandin, and Richard Rogerson. 2022. Hours and Wages. *Quarterly Journal of Economics* 137, no. 3 (August): 1901–62.

———. 2024. *Hours Worked and Lifetime Earnings Inequality.* Working Paper.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, et al. 2023. *Integrated public use microdata series, current population survey: version 11.0.* [dataset]. Minneapolis, MN: IPUMS. https://doi.org/https://doi.org/10.18128/D030.V11.0.

Goldin, Claudia, Sari Pekkala Kerr, and Claudia 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, Fatih, Greg Kaplan, Jae Song, and Justin Weidner. 2022. Lifetime earnings in the united states over six decades. *American Economic Journal: Applied Economics* 14, no. 4 (October): 446–79. https://doi.org/10.1257/app.20190489. https://www.aeaweb.org/articles?id=10.1257/app.20190489.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. Unequal we stand: an empirical analysis of economic inequality in the united states, 1967–2006. Special issue: Cross-Sectional Facts for Macroeconomists, *Review of Economic Dynamics* 13 (1): 15–51. ISSN: 1094-2025. https://doi.org/https://doi.org/10.1016/j.red.2009.10.010. https://www.sciencedirect.com/science/article/pii/S1094202509000659.

Heathcote, Jonathan, Fabrizio Perri, Giovanni L. Violante, and Lichen Zhang. 2023. Unequal we stand: nequality dynamics in the united states 1967-2021. *Review of Economic Dynamics* forthcoming.

MaCurdy, Thomas, Thomas Mroz, and R. Mark Gritz. 1998. An evaluation of the national longitudinal survey on youth. *The Journal of Human Resources* 33 (2): 345–436.

MaCurdy, Thomas, and Christopher Timmins. 2001. *Bounding the Influence of Attrition on Intertemporal Wage Variation in the NLSY.* Working Paper. Duke University and Stanford University.

APPENDIX 1. IMPUTATIONS

Appendix 1.1 Weekly Hours

When an observation has missing hours for at least one week worked but not all weeks worked of a year, 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 $q^h_{i,j} \forall\, j = \underline{t}, t, \bar{t}$:

$$\bar{h}_{i,t} = \frac{\sum_{j=\underline{t},t,f} q^h_{i,j} \times h_{i,j}}{\sum_{j=\underline{t},t,\bar{t}} q^h_{i,j}}. \tag{Appendix 1.1}$$

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[26]

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

We construct weights such that observations have more weight if they are more recent and have more weeks worked with an hours report.[27] We use this moving average $\bar{h}_{i,t}$ to impute average weekly hours for year t:

$$\widehat{h}_{i,t} = \frac{\sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}>0} \times h_{i,t,k} + \left(\widehat{wks}_{i,t} - \sum_{k=1}^{52} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}>0}\right) \times \bar{h}_{i,t}}{\widehat{wks}_{i,t}}. \tag{Appendix 1.2}$$

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 (Appendix 1.2) is the average reported hours worked for this year.

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

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

Appendix 1.2 Wages below Half the Minimum Wage

Figure Appendix 1.1 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.

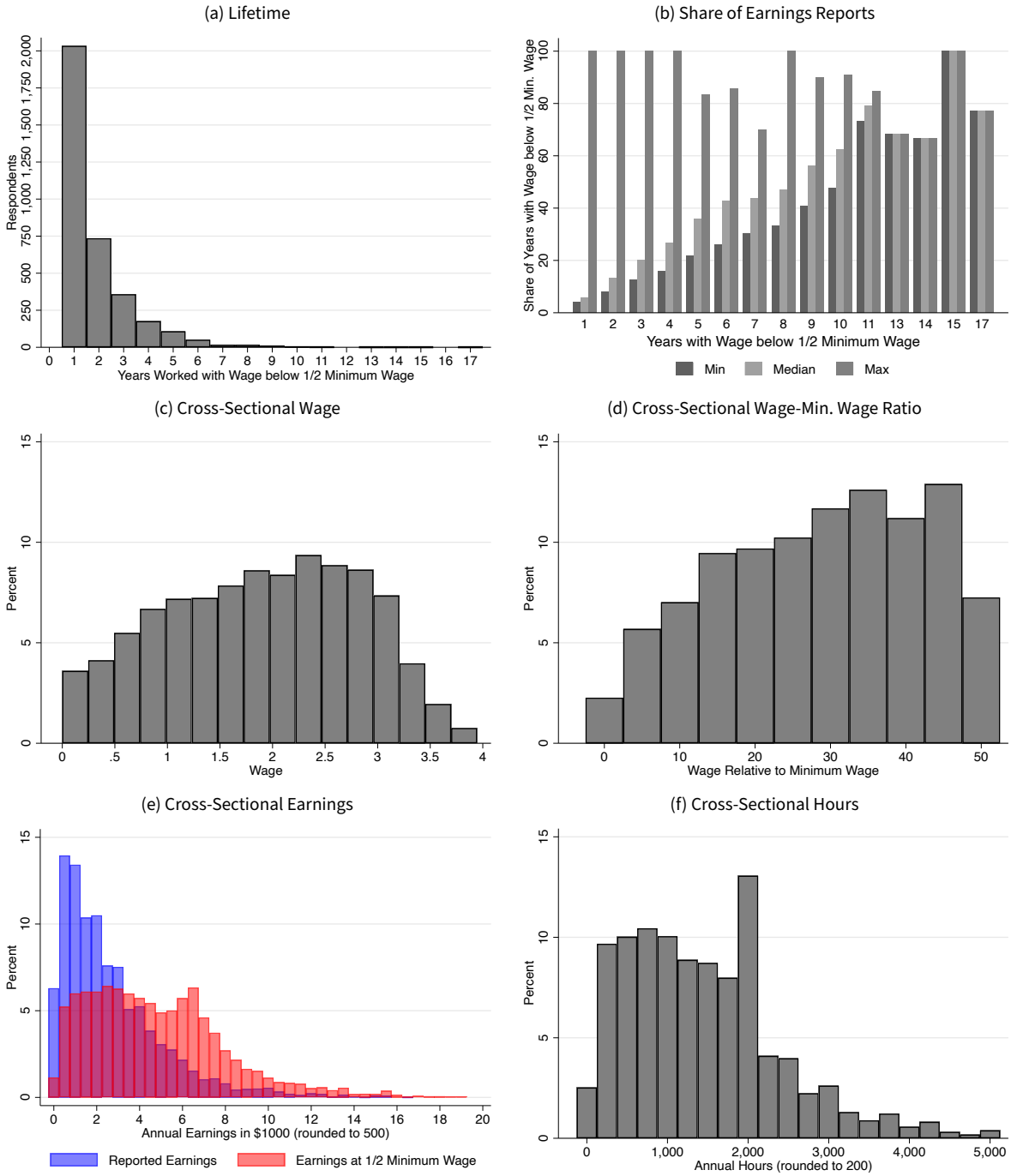
Figure Appendix 1.1a plots the lifetime distribution of years worked with a wage below half the minimum wage in reference years with a positive earnings report, conditional on having at least one such year. This is the case for 36.2 percent of individuals in our final sample who have worked at least one year. Figure Appendix 1.1b 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 Appendix 1.1c shows the distribution of wages below half the respective minimum wage, and Figure Appendix 1.1d shows the same wages relative to the respective minimum wage.

Figure Appendix 1.1e 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 Appendix 1.1f) and half the minimum wage in the respective year. Average reported earnings are \$2689 compared to \$4790 after the adjustment.

Figure Appendix 1.1

Distributional Facts on Wages below Half the Minimum Wage



Appendix 1.3 Top Earnings and Wages

Figure Appendix 1.2 provides more information on the top 1 percent of the cross-sectional wage distribution implied by equation (5). To put those numbers into perspective, we also provide information on the top 1 percent of the cross-sectional distribution of directly reported earnings.

Each panel of Figure Appendix 1.2 shows an outcome variable for individuals who are in the 99th–99.9th percentile of the earnings distribution (first three bars), in the top 0.1 percent of the earnings distribution (next three bars), in the 99th–99.9th percentile of the wage distribution (next three bars), and in the top 0.1 percent of the wage distribution (last three bars). The "current" (dark gray) 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 of which earnings group they were in 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 panels.

Figure Appendix 1.2a shows that individuals in the top 0.01 percent 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 Appendix 1.2b repeats this exercise but reports average earnings rather than average wages. While the patterns are more similar across all four groups, top wage earners have lower earnings than top earners, and the top 0.1 percent of wage earners have even lower earnings than the top 99–99.9 percent of wage earners.

Figure Appendix 1.2c explains these patterns. Annual hours worked by top wage earners are lower than those by top earners. Moreover, the hours worked for top earners change fairly little, unlike those at the top of the wage distribution. This is particularly true for the top 0.01 percent of wage earners, whose hours worked essentially collapse—leading to a temporary increase in their wage rates—before immediately recovering again. Figure Appendix 1.2d and Figure Appendix 1.2e show that this drop in annual hours comes from a reduction in both weeks worked and weekly hours worked, with the former effect being larger.

Finally, Figure Appendix 1.2f shows the probability of also being in the respective earnings group in the year before and after (the "current" bar is not well defined in the sense that it is 100% by construction, and we therefore do not display it at all). Top 0.01 percent earners are much more likely to have been, and remain, in that group compared to those with top 0.01 percent wages.

Table Appendix 1.1 provides additional results. The first two rows in each panel present the corresponding numbers for the top 1 percent of wage earners. The third row in each panel shows the updated outcome for those previously in the top 0.1 percent 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 now aligns more with the top 99–99.9 percent of wage earners, and consequently, wages are similarly affected. The two bottom panels further show that almost all person-year observations in the top 1 percent of the wage distribution are direct reports.

Figure Appendix 1.2
Top Earnings and Wages

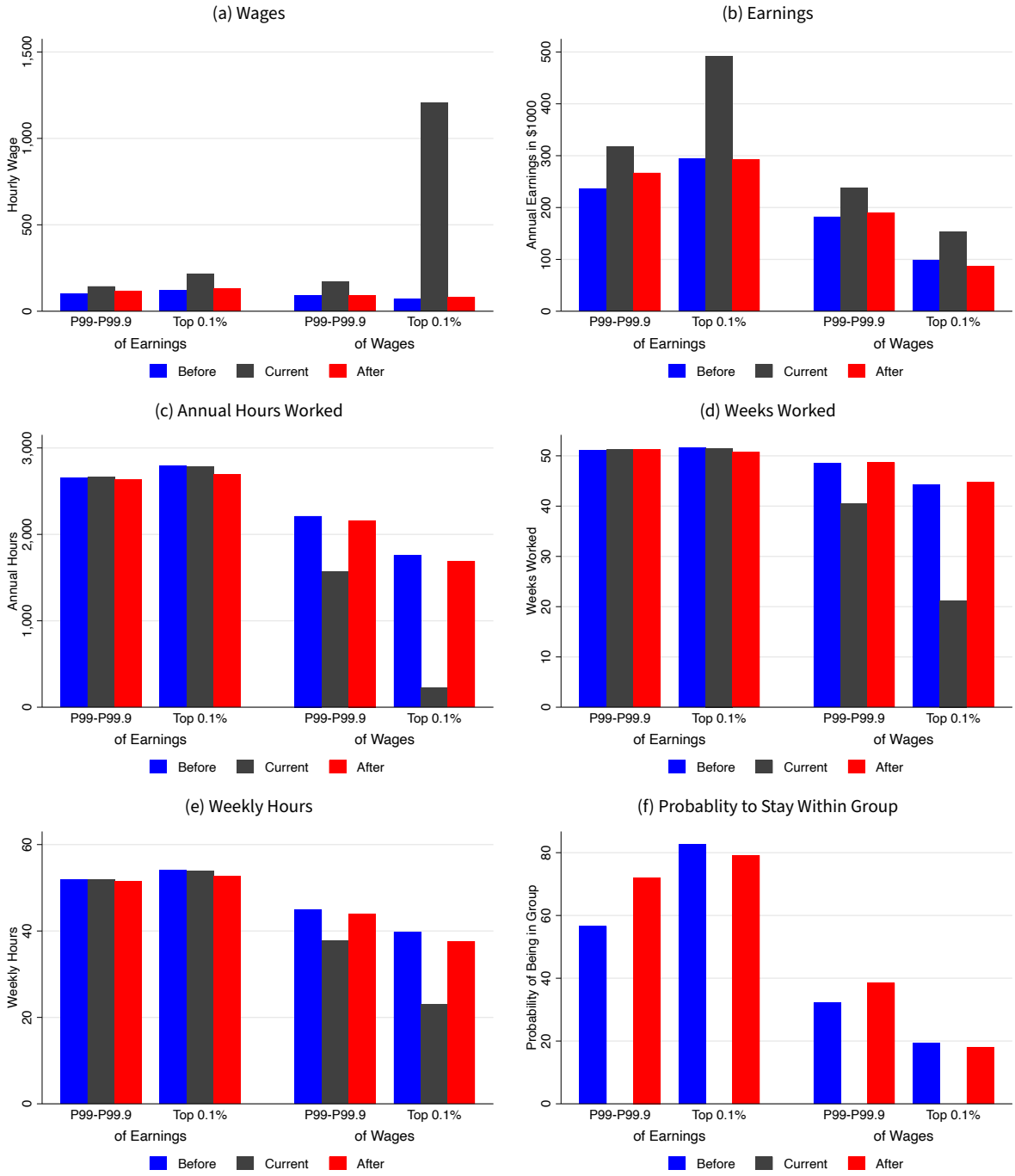


Table Appendix 1.1
Initial Top 1 Percent of Wages

	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	1151	1695
Weeks Worked			
Top 99%-99.9%	49	40	49
Top 0.1%	44	21	45
Top 0.1% - Adjusted	44	38	45
Weekly Hours			
Top 99%-99.9%	45	38	44
Top 0.1%	40	23	38
Top 0.1% - Adjusted	40	30	38
Average Hourly Wage			
Top 99%-99.9%	91	170	93
Top 0.1%	71	1207	82
Top 0.1% - Adjusted	70	196	96
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

Appendix 1.4 Income from Wages and Salary

In Section 4 we compare lifetime earnings in the NLSY79 with Guvenen et al. (2022), who analyze earnings on W2 forms to the SSA, thus exclusively covering this source of income. In our sample, in the cross-section of employed in reference years, 95.3 percent of person-year observations feature earnings from wages and salary, 1.2 percent are from farm/business income, and 3.5 percent are from both sources. From a lifetime perspective, nearly everyone who has worked at least one year has received earnings from salaries or wages during at least one year of their working life; 30.3 percent have at least one year with earnings from farm/business income; and 26.7 percent have at least one year with both types of earnings.

For the lifetime earnings comparison we therefore impute earnings from wages and salary separately. In particular, we only use earnings observations for imputation from years when individuals earned wages and salary, ensuring these figures are not mixed with other income types. Moreover, if earnings from wages and salary are missing, we require at least one report of such earnings from the previous or following five years. For the years we impute earnings, we assume that all reported hours are from employment despite the possibility of some being from self-employment. However, we believe 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. Additionally, for those earning both types of income, income from wages and salary is, on average, the dominant source. The share of income from wages and salary, from total earnings for those having both sources of earnings, is 48.6 percent at the 10th percentile, 51.9 percent at the 25th percentile, 76.9 percent at the 50th percentile, and 92.6 percent at the 75th percentile.

APPENDIX 2. SAMPLE SELECTION

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

(Appendix 2.1)
$$m^e_{i,t} = \begin{cases} 0 & \text{if } \sum_{k=1}^{K_t} \mathbb{I}_{e_{i,t,k}=-1} = K_t \\ 1 & \text{otherwise;} \end{cases}$$

i.e., $m^e_{i,t}$ equals zero when the employment status is missing for all weeks in year t. K_t equals to 52 or 53, depending on the number of weeks in year t. We keep an individual if for any $m^e_{i,t} = 0$, there is 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

(Appendix 2.2)
$$m^h_{i,t} = \begin{cases} 0 & \text{if } \sum_{k=1}^{K_t} \mathbb{I}_{e_{i,t,k}=1, h_{i,t,k}=-1} = \widehat{wks}_{i,t} \;\&\; \widehat{wks}_{i,t} > 0 \\ 0 & \text{if } \widehat{wks}_{i,t} = 0, \\ 1 & \text{otherwise;} \end{cases}$$

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 nonemployment equal to years of all missing weekly hours because neither 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 $\widehat{wks}_{i,t} > 0$, there is at least one observation within the past $\bar{T}$ or next $\bar{T}$ years in 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

(Appendix 2.3)
$$m^y_{i,t} = \begin{cases} 0 & \text{if } y_{i,t} = -1 \;\&\; \widehat{wks}_{i,t} > 0 \\ 0 & \text{if } \widehat{wks}_{i,t} = 0, \\ 1 & \text{otherwise;} \end{cases}$$

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 nonemployment equal to years of missing employment because neither 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 $\widehat{wks}_{i,t} > 0$, there is at least one observation within the past $\bar{T}$ or next $\bar{T}$ years in 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 $\widehat{wks}_{i,t} > 0$, $\sum_{j=t-\bar{T}}^{J=t+\bar{T}} m^y_{i,j} \geq 1$.

APPENDIX 3. MORE ON MISSING VALUES

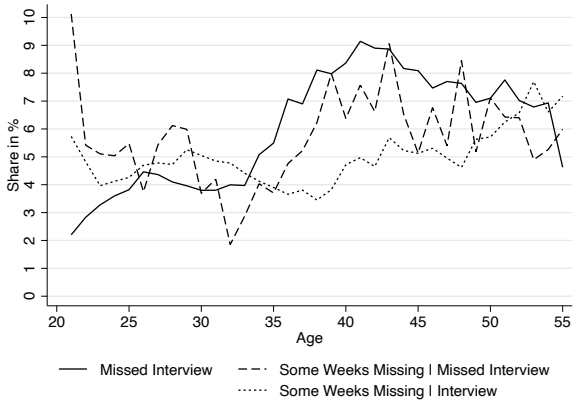
Appendix 3.1 Weeks with Missing Employment Status

The solid line in Figure Appendix 3.1 shows the share of individuals, from our sample of 7171, who missed an interview in reference years—years preceding an interview round. This share increases from around 2 percent to a peak of around 9 percent in the early 40s, and the sharp drop at age 55 is due to the selection criteria requiring at least one interview after this age. Additionally, the two youngest cohorts must 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 other two lines in the figure report the probability of having some missing weeks, conditional on having missed the interview (long dashes) and conditional on having participated in it (short dashes). Not surprisingly, the probability of having weeks with missing employment status is much larger for those who also missed an interview.

Figure Appendix 3.1

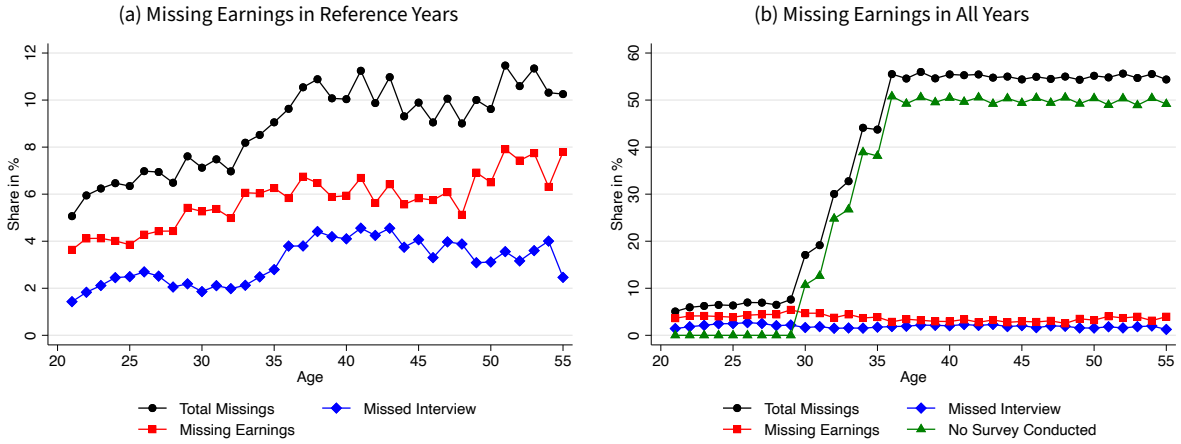
Interview Status and Prevalence of Weeks with Missing Employment Status



Appendix 3.2 Missing Earnings

Figure Appendix 3.2a documents the prevalence of missing earnings during references years. The top line with circles shows the total share of missing earnings. The two remaining lines break down total missing values between a missed interview (blue line with diamonds) and actual missing earnings, i.e., an individual participating in the survey but not reporting their earnings (red line with squares), with the latter being somewhat more important. Figure Appendix 3.2b includes missing earnings in nonreference years. The increase in missing data between ages 29 and 36 reflects the different ages at which participants switched to the NLYS79 becoming biennial. As a result of this change, about 50 percent of earnings are missing by construction in the second half of the life cycle.

Figure Appendix 3.2
Missing Earnings



APPENDIX 4. REPRESENTATIVENESS OF THE NLSY79

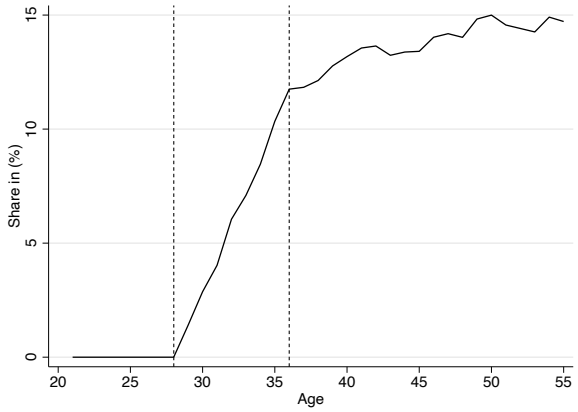
Appendix 4.1 The Effect of Immigration in the ASEC

The NLSY79 cohorts only include individuals who lived in the U.S. in 1979, while the ASEC includes anyone living in the U.S. independent of whether they were born in the U.S. or immigrated to the country. The information required to identify when someone moved to the U.S. (independent of whether they were already an U.S. citizen by then or not) is only available from 1994 onward in the ASEC. The oldest NLSY79 cohort was 37 in 1994 and the youngest was 30. Hence, only once the youngest cohort reaches age 37, the corresponding ASEC sample is fully comparable to the NLSY79 sample. Figure Appendix 4.1 shows the share of individuals identified as not living in the U.S. in 1979. Here, we shift all data by one year since all our labor market variables refer to the previous year.

Given this difference in sample composition, Figure Appendix 4.2 compares our main variables of interest for everyone born between 1957 and 1964 to individuals already living in the U.S. in 1979. In each panel, the immigration status year is unavailable for anyone at ages to the left of the first vertical line, and available for ages to the right of the second vertical line. We define someone as employed if they worked at least 520 hours per year (Figure Appendix 4.2a), and we report weeks worked (Figure Appendix 4.2b), hours worked per week (Figure Appendix 4.2c), and total annual earnings (Figure Appendix 4.2d) conditional on being employed according to this definition. We conclude that the gaps are relatively small between the two samples. For maximum consistency with the NLSY79, we choose a sample that excludes, as much as possible, individuals who were not living in the U.S. before 1979.

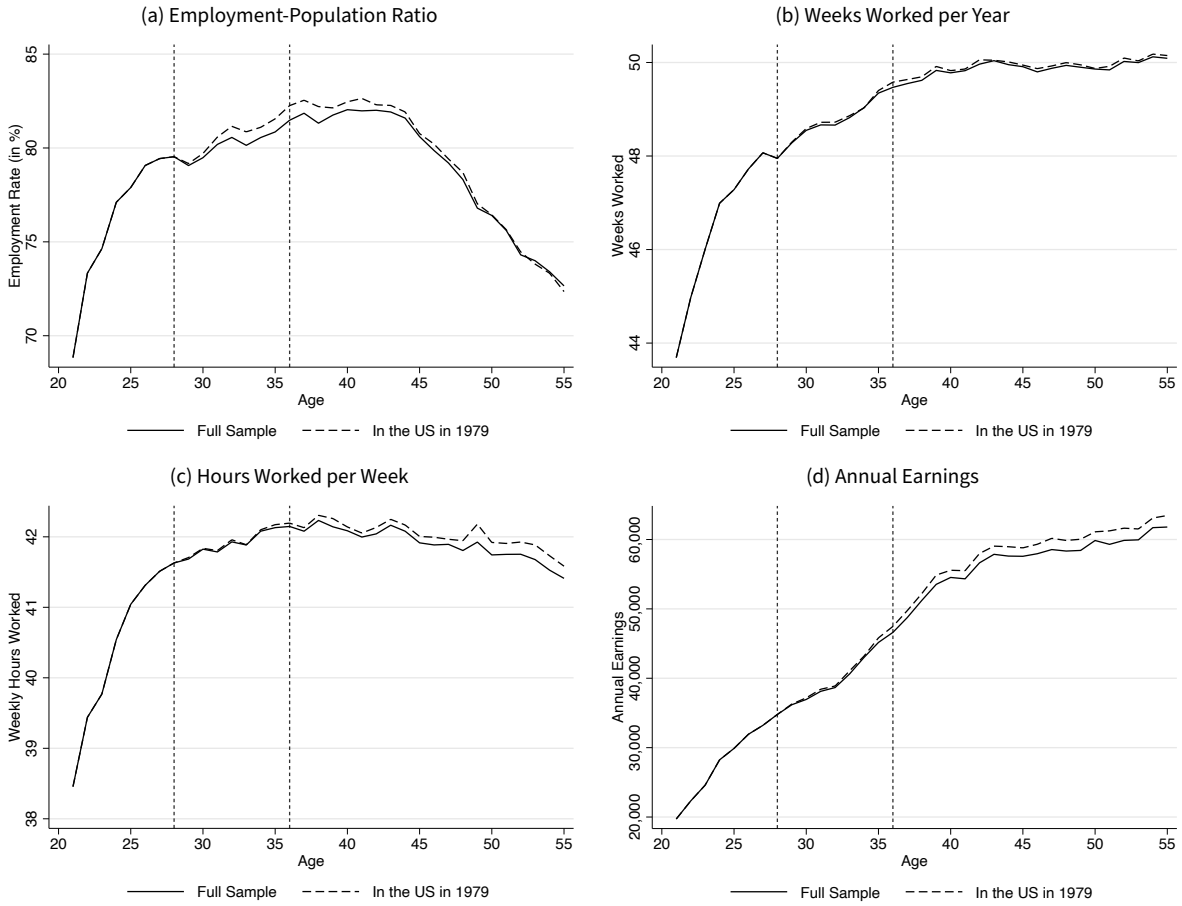
Figure Appendix 4.1

Share of the 1957–64 Cohorts Not Living in the U.S. before 1979 in the ASEC



Notes: The first dashed line indicates the oldest age at which the immigration status is unknown for any cohort, while the second dashed line indicates the youngest age at which the immigration status is known for all cohorts.

Figure Appendix 4.2
The Role of Immigration



Notes: The first vertical dashed line indicates the oldest age at which the immigration status is unknown for any cohort, while the second vertical dashed line indicates the youngest age at which the immigration status is known for all cohorts.

Appendix 4.2 Custom Weights

In our main analysis, we use the initial weights for weighting that include subsamples of military and economically disadvantaged non-Black/non-Hispanic youths, as well as individuals who had no interviews after age 21. We drop all of those individuals such that the initial weights are no longer necessarily representative of the 1957–64 cohorts residing in the U.S. in 1979. In addition, for the lifetime sample, we drop even more respondents as they do not have an interview after age 55, either because they have passed away or stopped participating in the survey. The "Weight IDs" option on https://nlsinfo.org/weights/nlsy79 allows us to separately construct custom weights for the set of individuals in either of the two samples. Table Appendix 4.1 shows that for our main variables of interest, the difference between using the initial or the custom weights is minimal.

Table Appendix 4.1
The Effect on Custom Weights

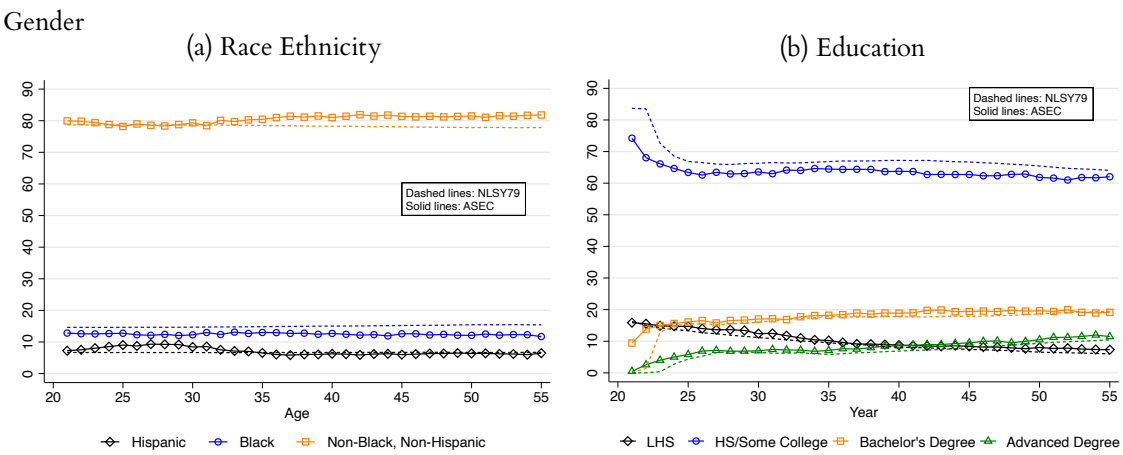
	Cross-Sectional Sample			Lifetime Sample		
	Initial Wgts	Custom Wgts	Diff	Initial Wgts	Custom Wgts	Diff
Employment Rate	81.0	80.6	-0.4pp	83.2	83.2	0.0pp
Annual Hours	2,113.9	2,113.8	0.0%	2,118.0	2,124.7	0.3%
Annual Earnings	49,808.7	49,511.4	-0.6%	51,529.9	51,717.6	0.4%
Hourly Wage	22.9	22.8	-0.6%	23.6	23.6	0.0%

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

Appendix 4.3 Comparisons with the CPS

Figure Appendix 4.3

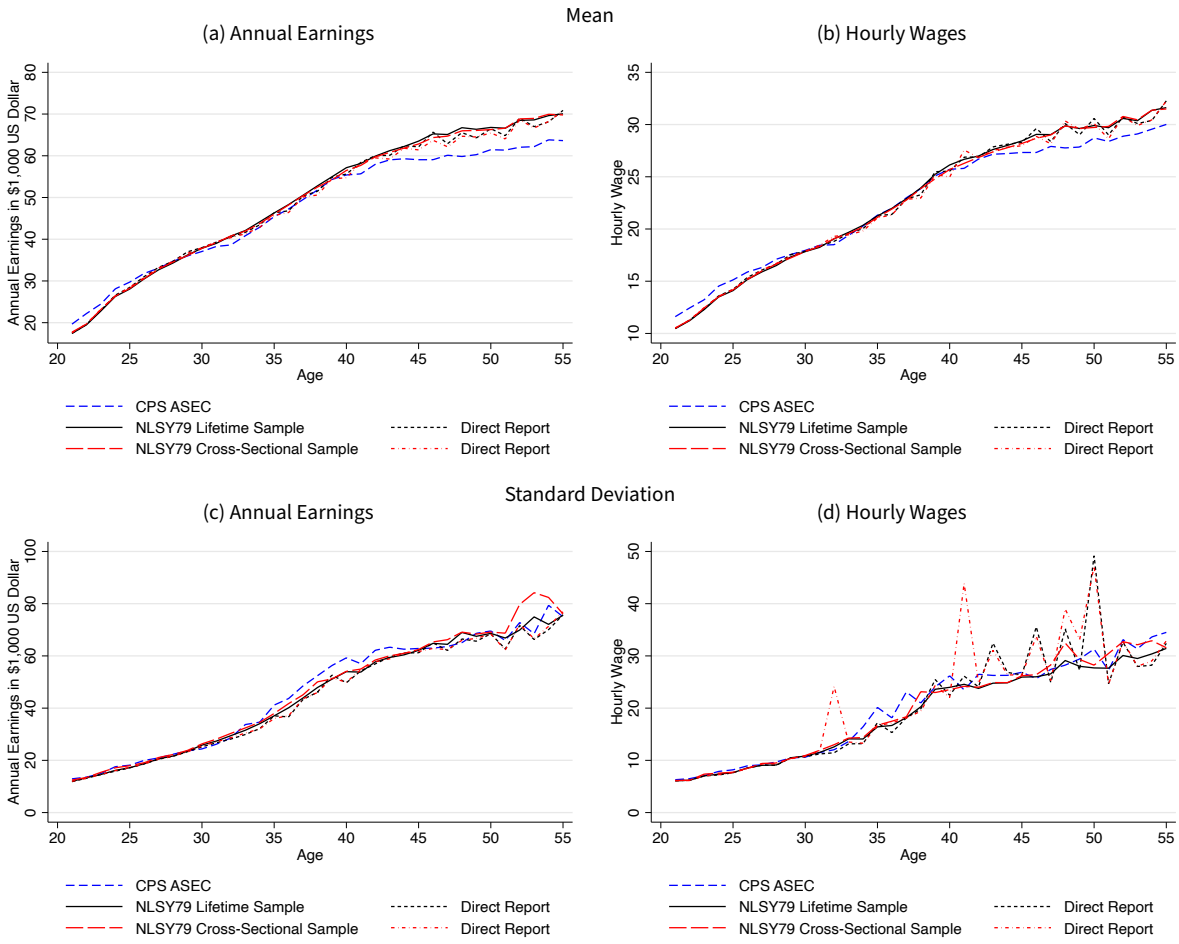
Demographic Composition in the CPS ASEC and NLSY79 Cross-Sectional Sample



Notes: In 1992 the CPS changed how education was recorded. Until 1991, as shown in Figure 5, we classify individuals by their highest grade completed (LHS = completed at most 11 grades, HS/some college = completed 12–15 grades, bachelor's degree = completed 16 grades, advanced degree = completed 17 or more grades) and from 1992 onward by their highest degree completed.

Figure Appendix 4.4

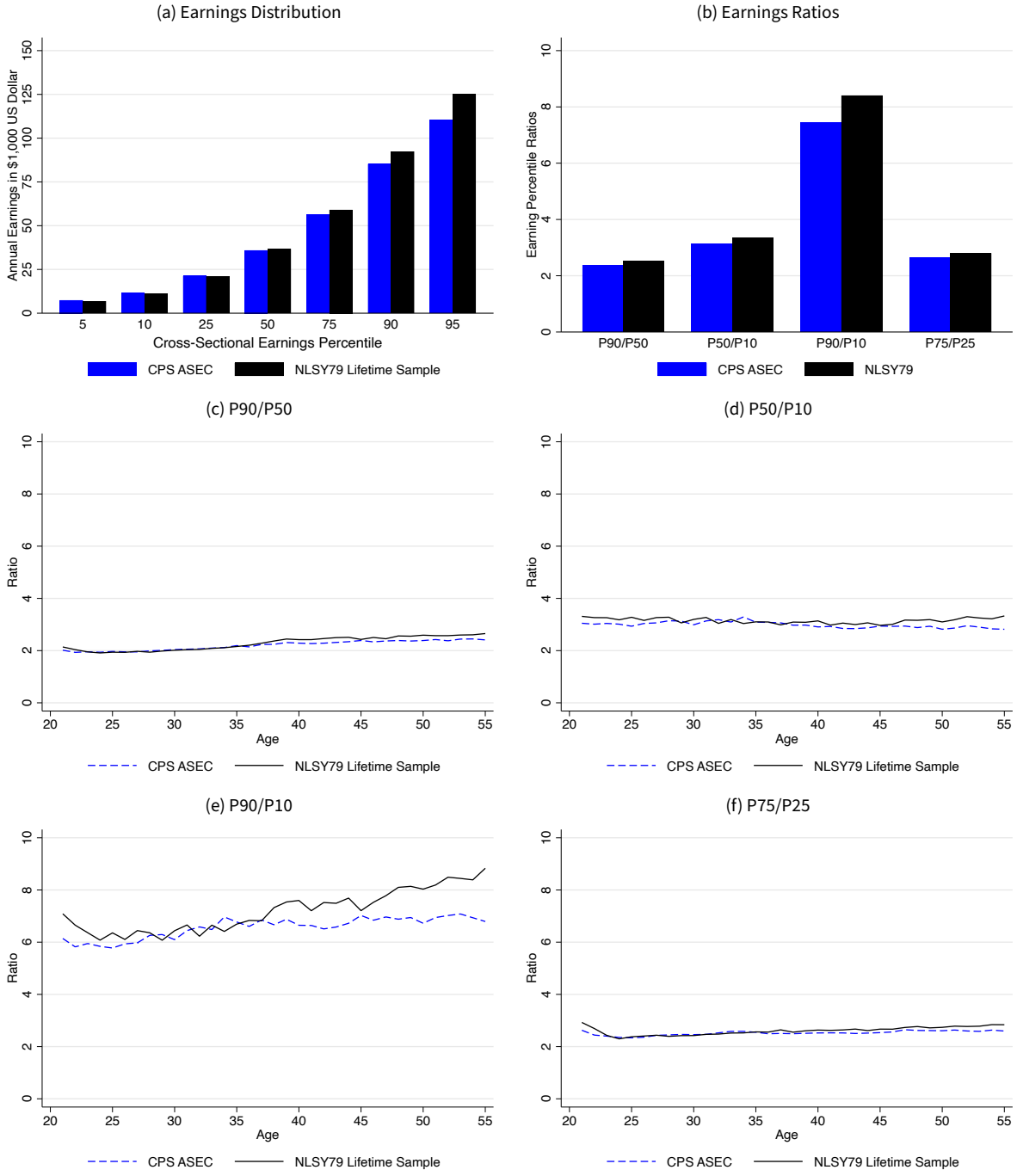
Income from Wages and Salary



Notes: We include only those person-year observations where annual hours worked are 520 or more.

Figure Appendix 4.5

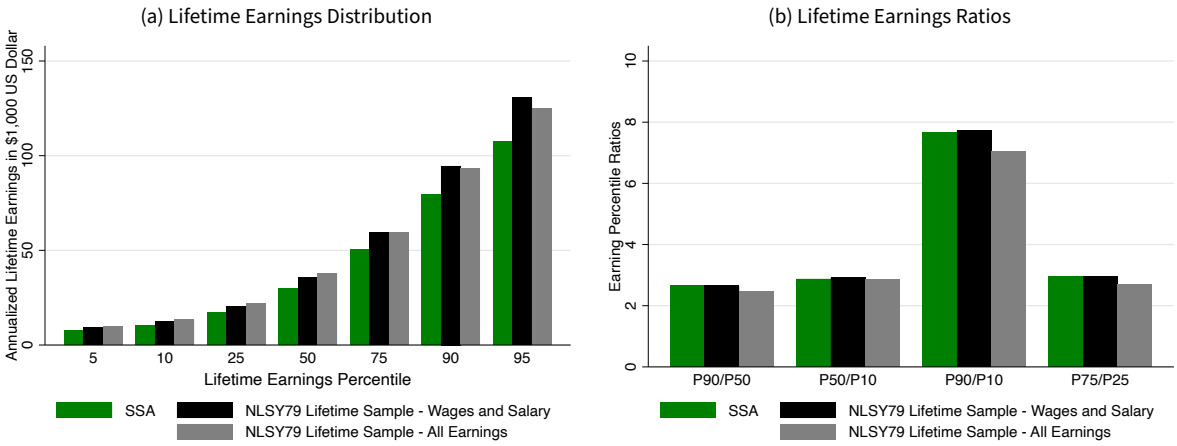
Cross-Sectional Earnings from Wages and Salary



Notes: We first identify individuals at the respective percentile of the cross-sectional earnings distribution. To calculate the value for each percentile, we then use the unweighted average of earnings of the identified individual, the five individuals with the closest lower cross-sectional earnings, and the five individuals with the closest highest cross-sectional earnings. We include only those person-year observations where annual hours worked are 520 or more.

Appendix 4.4 Comparisons with the SSA Data

Figure Appendix 4.6
Lifetime Earnings



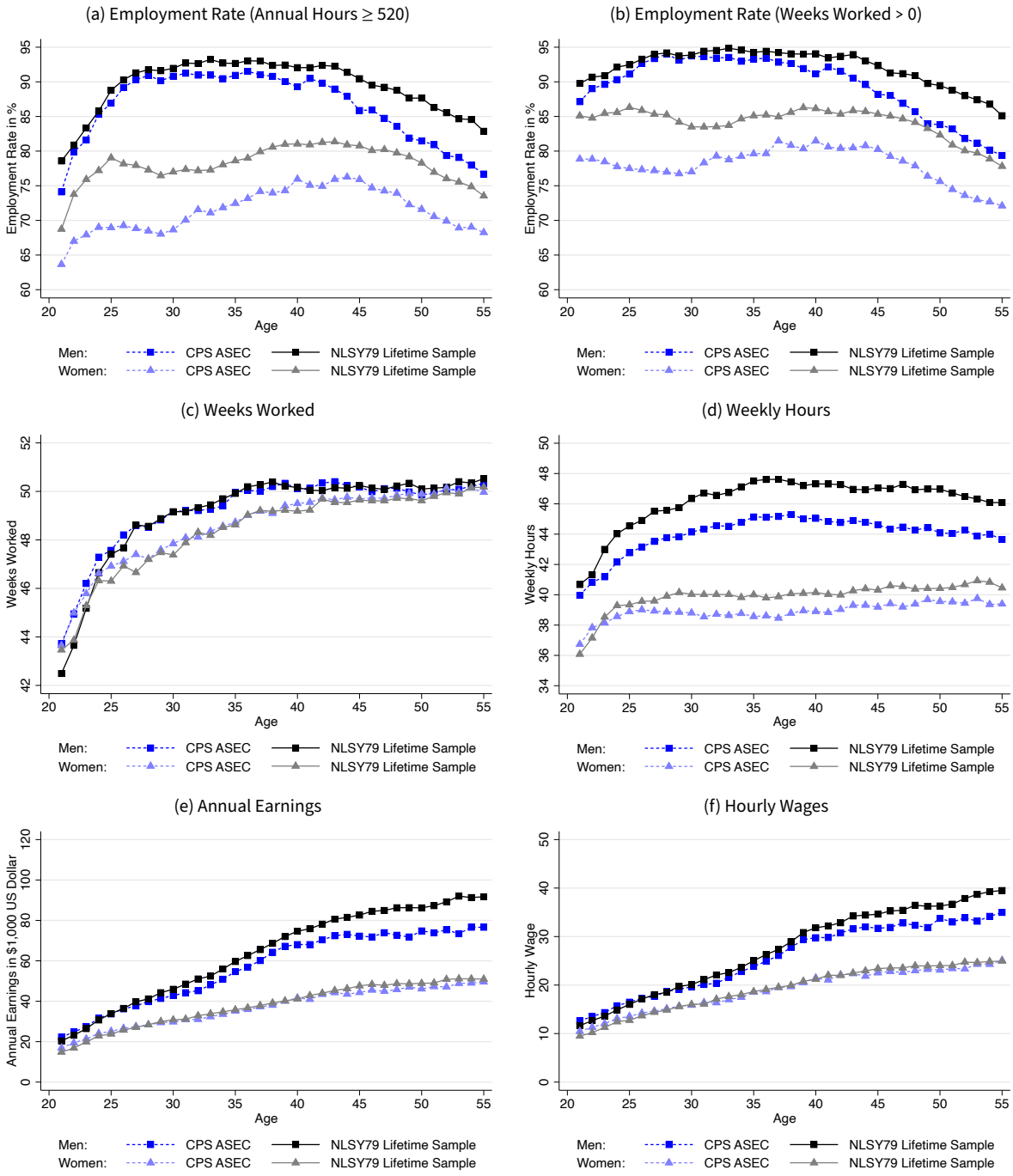
Notes: For the NLSY79, we first identify individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution. To calculate the value for each percentile, we then use the unweighted average of earnings of the identified individual, the five individuals with the closest lower lifetime earnings, and the five individuals with the closest highest lifetime earnings.

Appendix 4.5 Comparisons by Gender: Details

In this section, we present gender comparison graphs similar to those for the aggregate, mostly excluding figures from 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 only show key labor market outcomes for the cross-sectional sample and omit more detailed earnings comparisons, as they closely mirror those of the lifetime sample.

Appendix 4.5.1 Lifetime Sample

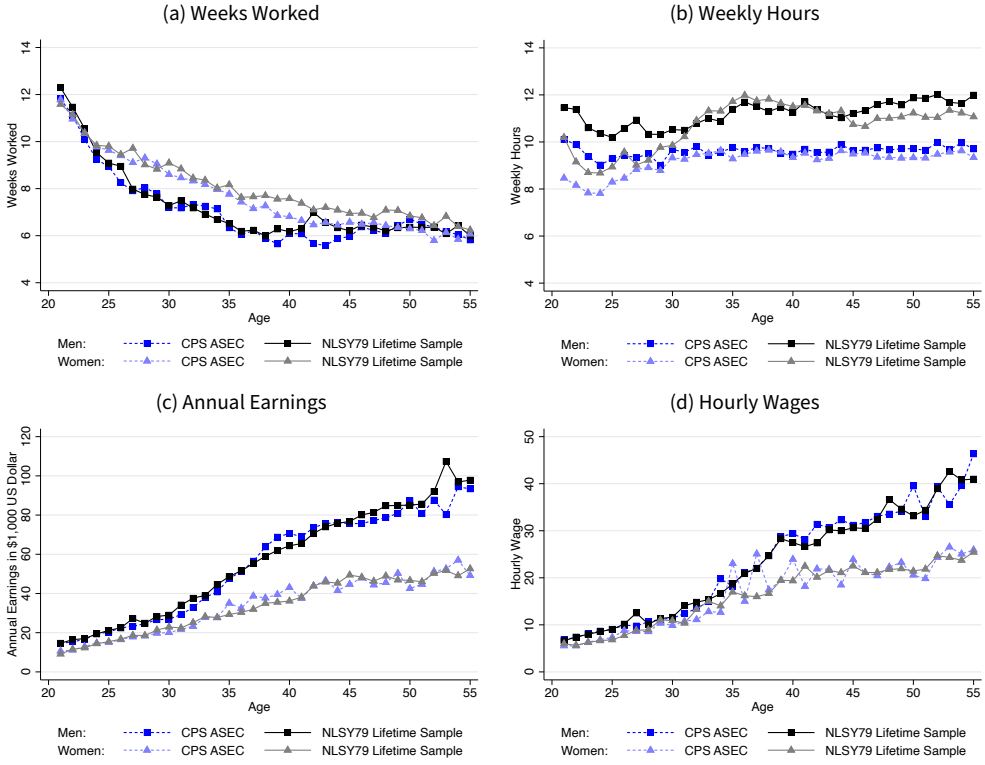
Figure Appendix 4.7
Means of Labor Market Variables by Gender



Notes: In Figures Appendix 4.7 to Appendix 4.7, we condition on working at least 520 hours per year.

Figure Appendix 4.8

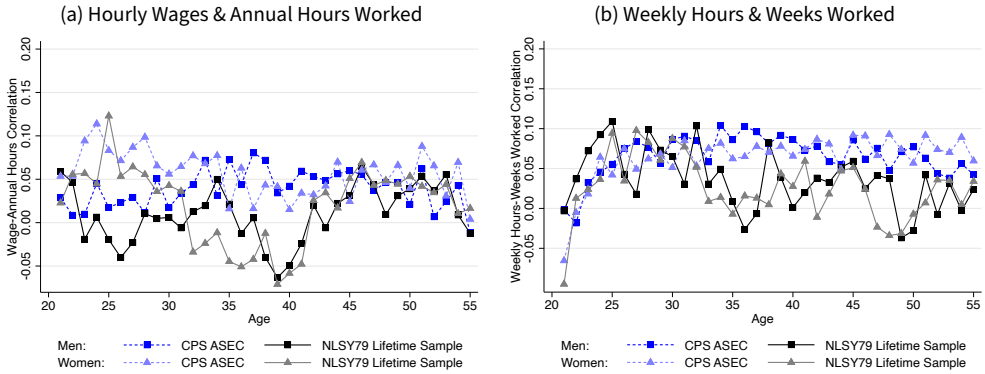
Standard Deviations of Labor Market Variables by Gender



Notes: We include only those person-year observations where annual hours worked are 520 or more.

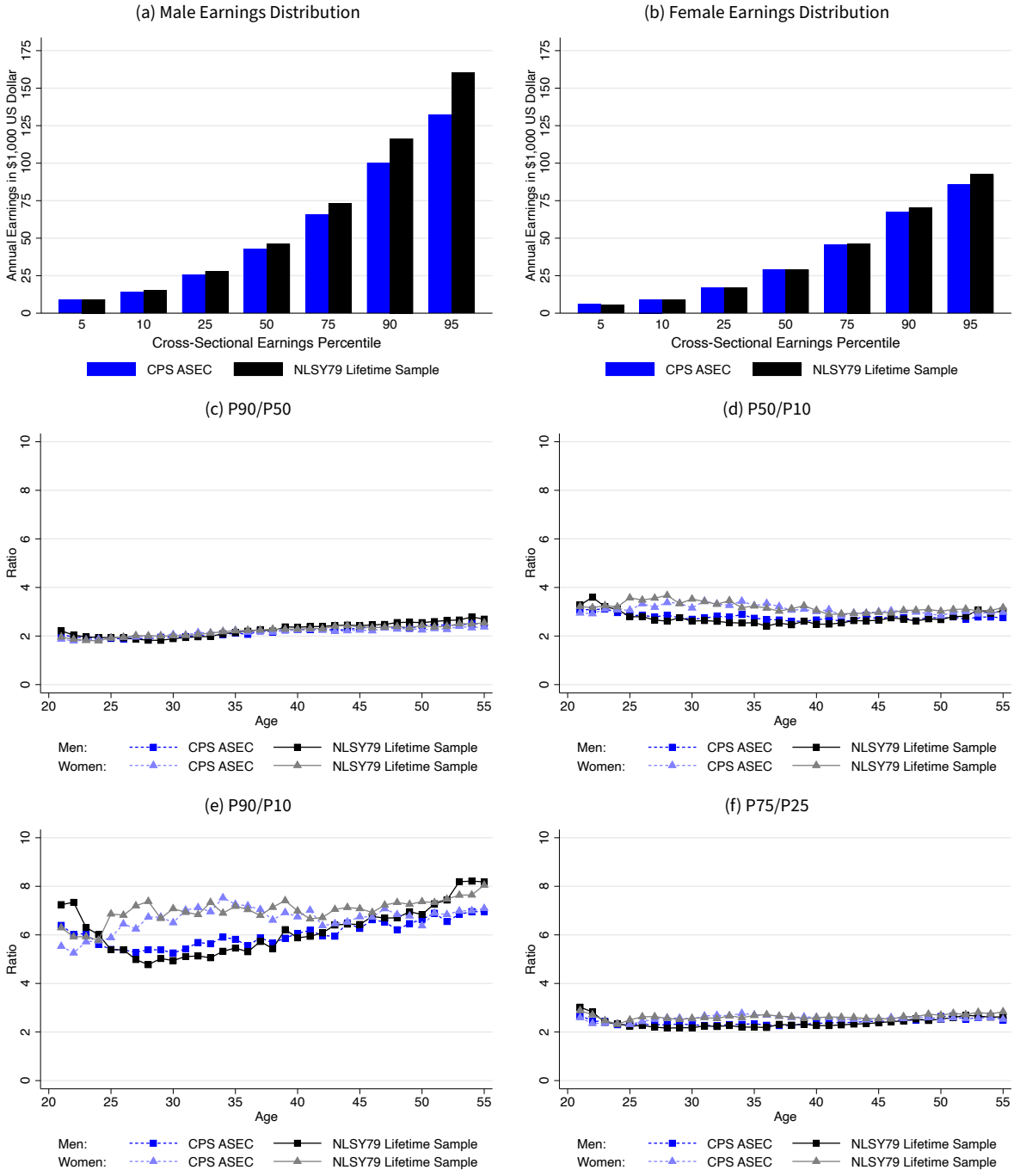
Figure Appendix 4.9

Correlation of Labor Market Variables



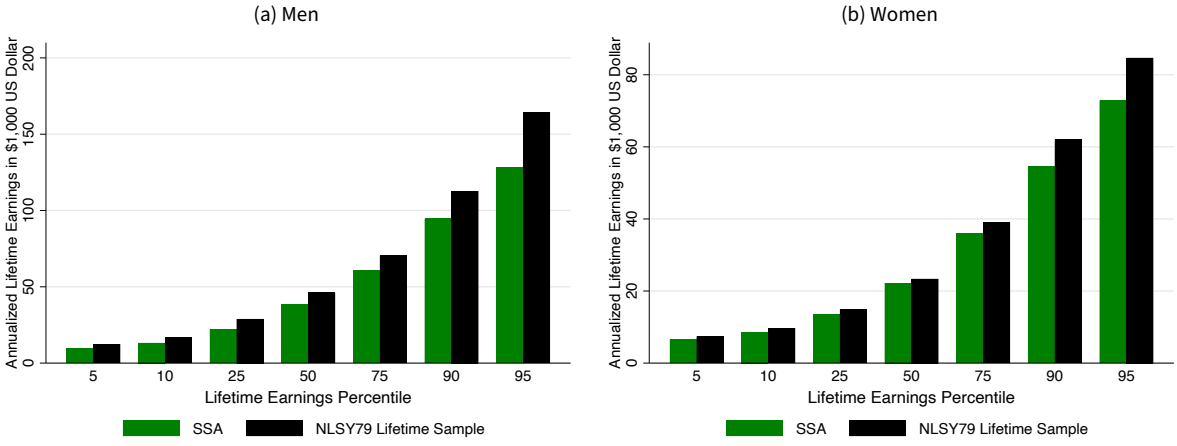
Notes: We include only those person-year observations where annual hours worked are 520 or more.

Figure Appendix 4.10
Cross-Sectional Earnings



Notes: We first identify individuals at the respective percentile of the cross-sectional earnings distribution. To calculate the value for each percentile, we then use the unweighted average of earnings of the identified individual, the five individuals with the closest lower cross-sectional earnings, and the five individuals with the closest highest cross-sectional earnings. We include only those person-year observations where annual hours worked are 520 or more.

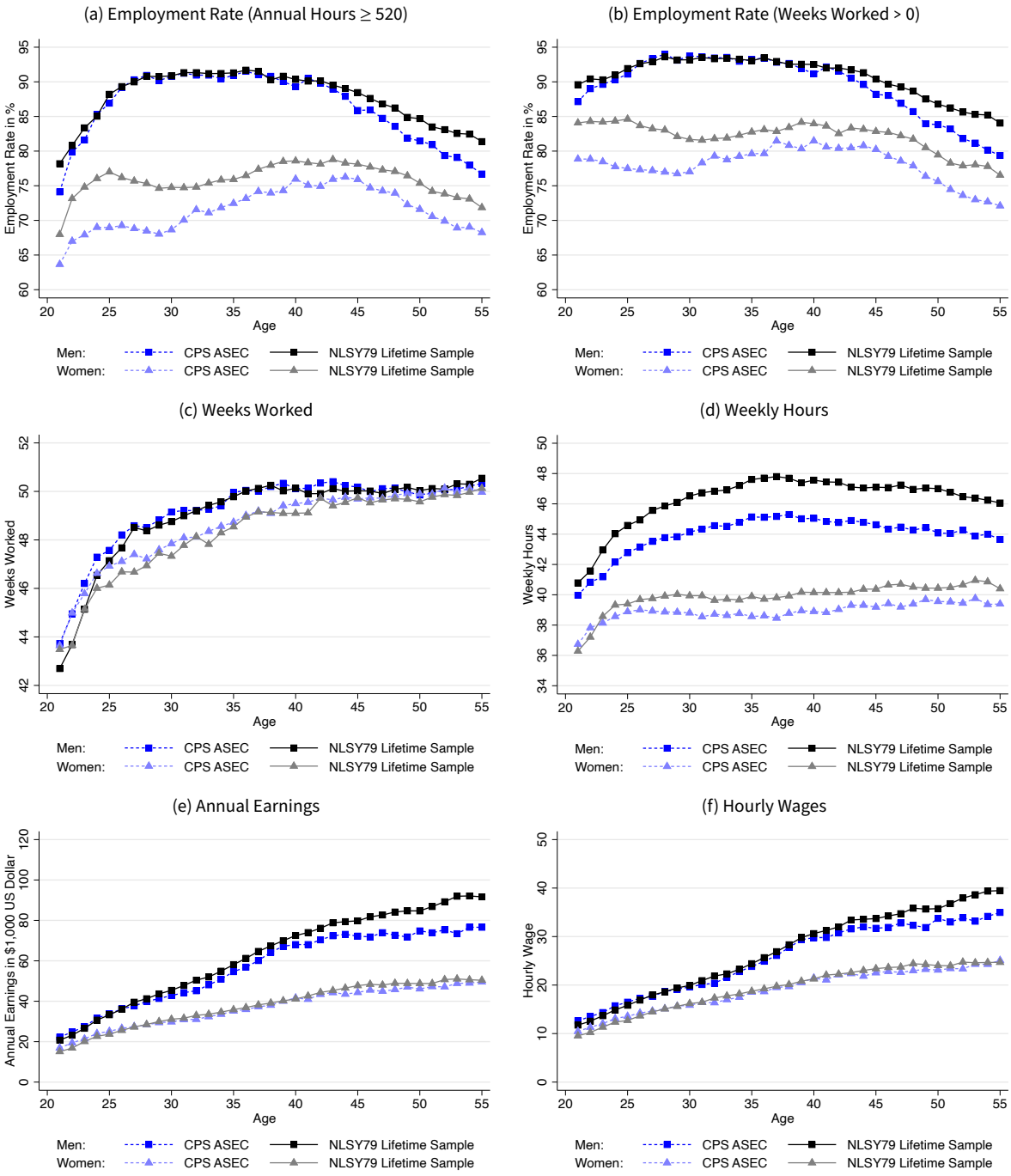
Figure Appendix 4.11
Lifetime Earnings Distribution



Notes: For the NLSY79, we first identify individuals at the 10th, 25th, 50th, 75th, and 90th percentile of the lifetime earnings distribution. To calculate the value for each percentile, we then use the unweighted average of earnings of the identified individual, 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).

Appendix 4.5.2 Cross-Sectional Sample

Figure Appendix 4.12
Means of Labor Market Variables by Gender



Notes: In Figures Appendix 4.12 to Appendix 4.12, we condition on working at least 520 hours per year.

Appendix 4.5.3 Lifetime and Cross-Sectional Sample Differences for Men

Table Appendix 4.2

Differences in Labor Market Variables between NLSY79 and CPS ASEC for Men

(a) Average Differences for Figures Appendix 4.7 and Appendix 4.12 (Mean of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Employment Rate (Annual Hours ≥ 520)	3.0pp	3.1pp	1.4pp	1.5pp
Employment Rate (Annual Hours > 0)	2.6pp	2.5pp	1.1pp	1.1pp
Weeks Worked	-0.2%	0.0%	-0.4%	-0.2%
Weekly Hours	5.0%	5.0%	5.2%	5.3%
Annual Earnings	11.9%	12.1%	10.2%	10.0%
Hourly wage	7.0%	7.3%	5.4%	5.7%

(b) Average Differences for Figure Appendix 4.8 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	2.5%	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	-1.1%	1.7%	1.4%	5.5%

(c) Average Differences for Figure Appendix 4.9 (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 Appendix 4.3

Differences in Earning Ratios between the NLSY79 Lifetime Sample and the ASEC and SSA for Men

	Cross-Section		Lifetime
	Pooled (Figure 11e)	Average 25-55 (Figures Appendix 4.10c-Appendix 4.10f)	(Figure 11f)
P90/P50	6.9%	3.8%	0.0%
P50/P10	2.7%	-1.4%	-0.2%
P90/P10	9.8%	2.6%	-0.5%
P75/P25	2.1%	-0.9%	-0.3%

Appendix 4.5.4 Lifetime and Cross-Sectional Sample Differences for Women

Table Appendix 4.4

Differences in Labor Market Variables between NLSY79 and CPS ASEC for Women

(a) Average Differences for Figures Appendix 4.7 and Appendix 4.12 (Mean of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Employment Rate (Annual Hours ≥ 520)	6.6pp	6.8pp	4.4pp	4.5pp
Employment Rate (Annual Hours > 0)	5.9pp	5.7pp	3.8pp	3.6pp
Weeks Worked	-0.4%	-0.1%	-0.6%	-0.2%
Weekly Hours	2.5%	2.7%	2.5%	2.6%
Annual Earnings	2.6%	1.0%	2.6%	0.6%
Hourly wage	1.1%	0.4%	1.5%	0.8%

(b) Average Differences for Figure Appendix 4.8 (Standard Deviations of Variables)

	Lifetime Sample		Cross-Sectional Sample	
	All Obs.	Direct Reports	All Obs.	Direct Reports
Weeks Worked	4.8%	2.7%	6.7%	4.5%
Weekly Hours	16.5%	16.4%	18.2%	18.0%
Annual Earnings	-0.9%	-7.5%	1.4%	-6.5%
Hourly wage	-1.9%	10.7%	2.6%	16.2%

(c) Average Differences for Figure Appendix 4.9 (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 Appendix 4.5

Differences in Earning Ratios between the NLSY79 Lifetime Sample and the ASEC/SSA for Women

	Cross-Section		Lifetime
	Pooled (Figure 11e)	Average 25-55 (Figures Appendix 4.10c-Appendix 4.10f)	(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	2.5%	2.7%	0.0%