**DATA CLEANING CODE REVIEW: OVERVIEW**

The code file PSID\_RAWTOCLEAN\_DECADES takes several inputs from the raw data folder, merges, and transforms them. The output file is a csv file titled psid\_clean\_2019.csv, which is then used in the imputation procedure.

The main inputs are psid\_wrk.dta, psid\_fertility.dta, psid\_marital.dta, and an IPUMS extract. The psid\_wrk.dta file contains all raw data from the individual and family files, which are downloaded from a single, multi-year extract from the PSID website. The “cart” used to create the data extract will be publicly available to download directly from the PSID’s data portal.

The structure of the code file first selects and cleans the “main” data from the psid\_wrk.dta data file: i.e., it selects and recodes the variables we draw on from the PSID that are collected in our target years (and the latest survey wave prior to the target year). It then draws on the same data file to create data tables that use information drawn from other years to create variables that match our target years, which are then merged to our main data table containing only our target years. For example, we use information on reported race across all survey years to construct a measure of race ever reported, which we then merge to our main data table by individual id. Similarly, we use information on hours and weeks worked across all survey years to construct a measure of work experience up to our target year, and merge to our main data by individual id and survey year.

We then use the fertility and marital history files to construct measures of fertility timing, parity, and marital status by our target years. Because detailed marriage and fertility history reports are only collected starting in 1985, some respondents in 1981 (one of our target years) have incomplete records, as they have attrited by the sample by 1985. For these respondents, we draw on information drawn from the individual and family files contained in the “main” data to construct best-approximation measures of fertility and marital status. We then merge this information with our main data by respondent id and survey year.

Finally, we draw on data from the decennial census and ACS to construct measures that capture features of the respondents’ occupational profiles by year. We then merge these measures by year and occupation codes, using detailed crosswalks provided by IPUMS and, when there are discrepancies in the coding schemes between IPUMS and the PSID, recoding the detailed occupation to match a similar occupation in the broader meso-level class of occupations. The details of these procedures are documented on a case-by-case basis when recoding the PSID occupation variables in the “main” data at the very beginning of the coding file.

**LINES 12-43**

The code file begins by loading in the packages that contain the required functions, as well as some custom functions.

**LINES 53-69**

These lines load IPUMS-provided crosswalks that allow us to harmonize industry and occupation codes across survey years

**1. MAIN DATA**

**LINES 73-621**

I then begin by selecting the variables we use in our target years and, when appropriate, simultaneously recoding them (for example, by converting numeric codes representing missingness like “999” to NA). For some variables, the recoding procedure differs by year; when this is the case, details about year-by-year recoding procedures are detailed in the code comments. For example, housework is topcoded at 98 hours in early years, but extends beyond 98 hours in later years starting at 1999. At this stage, we topcode all years at 98 hours for consistency.

While most of these recodings are straightforward and explained in the code comments, the industry and occupation codes are worth expanding upon. The PSID uses different occupation and industry codes across several years. To harmonize these codes, we use crosswalks provided by IPUMS and the census bureau, standardizing each year’s codes to 2010 occupation codes and 1990 industry codes. For the 2017 and 2019 waves of the PSID, occupations are set to 2010 codes, so no recoding is necessary. In the 19810-81 waves, we use an IPUMS-provided crosswalk to recode 1970 occupations to 2010 occupations. All but one of the occupations has a match in the crosswalk: I use the PSID codebook to recode this occupation into a similar occupation (600- Armed Forces in PSID to 580-Military rank not specified in the 2010 occupations). Since all military occupations are ultimately dropped from the analysis, this recoding does not affect the results. For the 2010-11 wave of the PSID, occupations are set to 2000 codes. We again use the IPUMS crosswalk to recode these occupations to their matching 2010 codes. One 2000 occupation code (787) has no match in 2010: while the PSID categorizes this as a production occupation, it does not provide a detailed code. We recode this occupation to 785, another production occupation that would be classified in the same higher-level occupational categories. For the industry codes, we again use crosswalks provided by the census bureau to recode all 2012 industry codes in 2017-19 to 1990 codes and crosswalks provided by IPUMS to recode 1970 industry codes used in years 1980-81, 1990-91, and 2000-01 to 1990 codes. All 1970 and 2012 industry codes have corresponding 1990 codes. In 2009-2011, the PSID uses 2000 industry codes. Two industry values in the PSID have no match in the crosswalk nor any matches we could find for existing 2000 industry codes. One value is 120 (assigned to two male respondents): as it falls within the range of non-durable manufacturing industries, we recode this value to 127, also in the non-durable manufacturing range. The other value is 775 (assigned to 5 female respondents): because it falls within the range of Management, Administrative and Support, and Waste Management Services rage, we recode this value to 779, which is coded as sanitary services.

We then transform the data into long format and using PSID sequence numbers and information on relationship to head, create some indicator variables for whether individuals should be included in the sample (for example, whether they are part of the immigrant or Latino samples, whether they are interviews as a head or a wife).

**LINES 626-640**

These lines create a plot that details the proportion of PSID respondents that are identified as household heads/wives by year and age- I am keeping it in for referral for now, but we can edit this out in the future if we decide we no longer need this either as a number to reference or as a figure.

**LINES 644-778**

These lines further clean the coding file by selecting only individuals who are heads or wives in our target years and coding their values individually- for example, if an individual is coded as a “head”, then these lines code the individual’s union status according to union\_hd, as opposed to union\_wf. These lines also conduct some additional recoding for individual-level variables, such as using occupation and industry code to code whether an individual works in agriculture or the military. ***Lines 732-778*** use the detailed industry codes to create 2-digit industry codes- I’m not sure we use these anymore beyond excluding individuals who are in military or agriculture, so we could cut these lines.

**2. EVER REPORTED RACE**

**LINES 785-1031**

These lines draw on the psid\_wrk.dta file, but use all reported questions on racial identification and whether the person is of Spanish/Hispanic descent to create a measure that codes respondents’ race according to all of the available respondents’ reports. From 1985 to 1993, interviewees can report up to two races for the head/spouse. In 1994 to 1996, individuals can report up to three races. Starting in 1997 and up to 2003, individuals can report up to four races, but "Hispanic" is not asked separately and is an option in the race questions. Starting in 2005 up to 2017, the PSID asks about Hispanic ancestry separately and individuals can report up to four races for the head/spouse. Following Blau & Kahn, we code individuals as “Hispanic” if they report having Hispanic descent, “Black” if they report being Black in any of the available race questions, “Other” if they report another race in any of the available race questions (American Indian, Aleut, Eskimp; Asian, Pacific Islander; Other), or “White” if they report being White. This file then converts this table to a wide format, where each row is an individual and each column takes on the coded race for that year. Using this “wide” format, ***lines 1015-1023*** create a person-specific indicator, race.ever, that codes the respondent as exclusively “Hispanic”, “Black”, “Other”, or “White” if the respondent is coded as Hispanic, “Black”, “Other”, or “White” across any year (in that order of priority). ***Lines 1026-1031*** merge this data on ever reported race to the “main” data by individual id: it then creates the variable “race” using this “race.ever” measure. If there is no available “race.ever” measure, we use the reported race for the head of household in the main family files.

After this procedure, we use the census bureau’s two-digit occupation categories to classify these occupations into broader groups. Following Blau & Kahn’s occupational categories, we depart from the census bureau categories by grouping physicians/surgeons with dentists, lawyers, and judges, classifying other legal staff (such as paralegals) as education, training, library, and legal support, and keeping health practitioners (excluding physicians and dentists) together.

**3. WORK EXPERIENCE**

Work experience is a cumulative measure constructed with reported information across many survey years. ***Lines 1034-1198*** gather information on individuals’ reported age, relationship to head, annual work hours, years worked full-time, and years worked for each respondent and each year, creating a person-year record for each individual, recoding missing data as missing as per the PSID codebook, and then selecting only individuals who are included in our main data.

***Lines 1207-1304*** use this data to compute our measure of experience. In each year that a respondent joins the PSID as a new head or spouse, they are asked to report total years of experience and total years of full-time experience since age 18. All respondents are also asked this question in 1976 and 1985. For all years in which this baseline experience is reported, if the measure of experience since age 18 is greater than the current age minus age 18, I censor experience to equal current age minus 18 in ***lines 1217-1220***.

One complication that arises is that starting in 1999, the PSID stops collecting annual data and only collects data on alternate years. While at each wave, the PSID reports annual hours worked in the prior year, the PSID does not report annual hours worked in the gap year. The PSID reports different kinds of information on hours, weeks, and months worked for different gap years. In 1999 and 2001, the PSID reports about how many weeks the respondent worked on all jobs in the gap year (1997 and 1999), how many months the respondent worked for pay, and how many hours the respondent usually worked per week. For these waves, in ***lines 1229-1232*** I compute the annual number of hours worked in the intial gap years by multiplying the total number of weeks worked by the hours usually worked per week if the respondent was employed. If data is missing on the total number of weeks worked in the gap year, I compute the annual number of hours worked in the gap year by multiplying the number of months the respondent worked in the gap year by 4.33 and then multiply this product by the hours usually worked per week. If the respondent reports not being employed during the gap year, their hours are coded as zero.

For the subsequent gap years, the PSID reports whether the respondent was employed during the gap year and the average number of hours the respondent worked a week on all jobs in the gap year. The PSID also provides a constructed variable on the number of weeks the respondent was employed in the gap year, based on the employment history calendar. For these waves, I compute hours worked during the gap year by multiplying the number of weeks the respondent was employed during the gap year times the average hours worked per week across all jobs, if the respondent was employed. If the respondent reports not being employed during the gap year, their hours are coded as zero.

In ***lines 1256-1287***, for each year, I take the last year the individual was asked their years of experience since age 18 as a baseline and add a year of total experience for each subsequent year up to the target year if the individual reports working positive hours and a year of full-time experience for each subsequent year if the individual reports working at least 1500 hours a year. Then, in ***lines 1292-1302***, I restrict to our target years and merge with the main data.

**4. FERTILITY**

Fertility is computed using the fertility history files. ***Lines 1308-1333*** download the fertility history files, which contain individual ids and child-specific records: a record of whether the child is biological or not, the year the child was born, and the total number of children born to that respondent. Each individual has a record, regardless of whether they’ve had a child or not. ***Lines 1321-1333*** use these records to create a year-by-year record of number of children born to each respondent by the end of that year. ***Lines 1335-1358*** join these fertility records to the main data, and do some data cleaning (commented out in detail in the code file).

The fertility history files only started being collected in 1985. As a result, there is a subset of respondents observed in the early years who have attrited by 1985 (473, ~9 % of respondents) and are not in the fertility history file. For these respondents, we use data collected on the total number of kids and the age of the oldest child born to the head to create synthetic measures of fertility in ***lines 1364-1393***, and merge this data to the main data in ***lines 1396-1417***. These lines also conduct some data cleaning, commented out in detail in the code: for example, if the number of children after this procedure is still missing in 1980-81, we assign the number of children born to the head. Some respondents have a value that indicates they have no children in the measure of age at first birth for individuals that report having children in the fertility file. For individuals who report children in the fertility files but no children in the age of first birth variable, we code age at first birth as missing and then use the fertility history files to create alternative measures of first birth by observing the first year in which a respondent transitions from having no children to having a child and taking the respondent's age in that year as the age of first birth (***lines 1430-1493***), merge this data with the main data (***lines 1496-1519***), and use these measures to create categorical measures of number of children to date and age at first birth.

**5. MARITAL STATUS**

Marital status is computed using the marital history files. These files contain a record for each marriage by individual id (all individuals have at least one record, even if never married): the order of the marriage, its status as of the latest interview date, and dates for the marriage. Our procedure takes these records to create measures of marital status as of the last day of the year in which wages are measured (so, the last day of the calendar year prior to the survey year). ***Lines 1540-1547*** code the date of the marriage record’s dissolution (separation or divorce) using the recorded month/dates. ***Lines 1554-1690*** aim to capture the status of a given marriage record at the end of the wage year. In ***lines 1554-1613***, we code, for each marriage record, the last date at which the respondent was in that marriage within the target year. That is, if respondents were never married, the variable is coded as missing: if the marriage is intact or the marriage end date is greater than the last day of the wage year, then the respondent is coded as married: if the marriage has dissolved, it codes the date at which that marriage was dissolved. In ***lines 1662-1665***, we use the marriage start dates to code whether the respondent was ***not yet*** married in that marriage by the end of the survey year. These two procedures allow us to distinguish between marriage records for respondents who are no longer married, but remain married for the duration of the wage year, and respondents who will be married, but are not yet married as of the end of the wage year.

For some respondents, we have missing values for the end date of the marriage, but nonetheless can ascertain whether respondents were still in that marriage as of the wage year. For example, if the last year in which marital history information is collected precedes the wage year and the status of that marriage is coded as a dissolution, then we know this dissolution must have occurred prior to the end of the wage year. We use this information in ***lines 1673-1691*** to code marital status of that marriage as of the wage year. Then, in ***lines 1696-1722***, we get the status of each marriage as of the target year, and tag respondents’ current marital status according to these statuses. In the following lines, we merge the marital status data with the main data by id and year and recode some additional variables (for example, creating categorical variables for education and region, coding marital status as the marital status of the head if marital status is missing).

**6. IPUMS OCCUPATIONS**

***Lines 1757-1820*** recode the 2010 occ codes in the PSID to to match the IPUMS occupational data. There are 28 OCC2010 codes in the PSID that do not have a match in the IPUMS data, depending on the survey year. These lines recode these occupations to related occupations in the same broader classification: details on this recoding are commented out line by line in the code.

***Lines 1838-1845*** read in the IPUMS data, and lines 1848-1860 restrict the data to individuals who are employed, working for wages, and between the ages of 35 and 45. We compute occupational characteristics for each detailed occupation among individuals who meet these work hours/age range criteria in ***lines 1868-1881***. In ***lines 1888-1987***, we merge this occupational data to the main data by year and detailed 2010 occupation, and we create indicator variables for sample restrictions in ***lines 1901-1932***. Finally, line 1935 writes this data file to a csv, which is then used in the imputation procedure.