Online Appendix for Inequality and Mobility over the Past Half Century
using Income, Consumption and Wealth

Consumption Imputation Methodology and Comparison

This appendix describes the methodology used to impute consumption to the Panel Study of Income Dynamics (PSID). Then, we present several comparisons to help judge the quality of the imputation, comparing the PSID to the Consumer Expenditure (CE) Survey and comparing imputed PSID consumption to reported PSID consumption, when reported consumption is available. We first begin by describing the CE Survey data.

Consumer Expenditure Survey Data

The CE survey has been a continuing quarterly survey since 1980, with an earlier collection in 1972-73. Data are collected from consumer units five times over a 13-month period.[1] The second through fifth interviews are used to collect expenditures for the previous three months; for example, a consumer unit that is visited in March reports expenditures for December, January and February. Also collected in this survey is the inventory of certain durable goods, e.g., homes, real estate, vehicles, and major appliances. To obtain an annual measure of consumption and income, we use consumer units who participate in the survey for all interviews (representing 75-80 percent of all consumer units). The consumer units are then placed in the quarter in which their last interview occurred, and the weights and household demographics are those from the last interview.

Several changes in the CE methodology and sample make creating a consistent measure of consumption more difficult. There was no CE survey from 1968-1971 and 1974-1979. While there was a CE Survey in 1972 and 1973, the survey made significant changes between the 1972-

[1] A consumer unit comprises members of a household who are related or share at least two out of three major expenditures--housing, food, and other living expenses. Since 2015, data are only collected for four quarters.

1973 version and those that followed. In 1972-1973, the CE surveyed every household once and asked about consumption over the previous year. Since 1980 the CE surveyed every household four times and asked about quarterly consumption. The level of inequality differs depending on whether the household is asked about annual consumption or quarterly consumption, resulting in a significant break in the time series when we attempted to impute using 1972-1973 and post-1980 data. Thus, we use the 1980-1981 CE to impute back to 1968 for the PSID. We do not use the 1972-1973 CE.

Another significant change is that the CE only surveyed urban households in 1982 and 1983. We use the 1980-1981 CE, which includes urban and rural households, to impute for 1982 and 1983. The CE also redrew its sample in 1986, 1996, and 2005 to update the sample for the most recent Decennial Census. In practice, this means that the CE would drop many households that were sampled in 1986Q1 that should have been interviewed in 1986Q2 and replaced them with new households. This redrawing of the CE sample results in a significantly smaller number of four-quarter households in those three years. To boost the sample size, we use the year before and the year after the change in the sample. For 1986, we use 1985, 1986, and 1987. The small sample size issue also applies to 1984 when the CE reintroduced rural households. When imputing consumption for the 1984 PSID, we use the 1984 and 1985 CE Surveys.

Our measure of consumption includes the amount that the consumer unit actually spends for current consumption plus the estimated service flows from homeownership and vehicles. It includes expenditures for food, housing, transportation, apparel, medical care, entertainment, and miscellaneous items for the consumer unit. Excluded are expenditures for pensions and social security, savings, life insurance, principal payments on mortgages, and gifts to organizations or persons outside the consumer unit. The service flow from homeownership represents the rental

equivalence of the owned home. For renters, we use the rent. The vehicle service flow is estimated using the make, model (when available), and year of the vehicle to generate the flow of services of the vehicle in a given year.[2]

Imputation Methodology

Several researchers have imputed consumption for the PSID individuals using the CE data. Skinner (1987) first imputed total consumption for the PSID, and most subsequent research has followed this method. Using CE data, Skinner (1987) estimates an equation with total consumption as the dependent variable. In his preferred specification, the independent variables are food at home, food away from home, rent if a renter, utilities, market value of the home if a homeowner, and the number of vehicles owned.

More recently, Blundell, Pistaferri, and Preston (2008) estimate a log-linear demand function for food consumed at home. Blundell et al. (2008) deviate from the Skinner (1987) methodology because they argue that their demand for food equation comes from economic theory rather than a statistical procedure. Their estimated equation is:

$$ln(food\ at\ home) = M'\mu + \beta * ln(C) + e \quad \text{(A1)}$$

The dependent variable equals food at home expenditures divided by the CPI for food at home, resulting in an estimate of the quantity of food purchased. Using the quantity of food purchased is consistent with their estimation of the demand for food. The matrix M contains the CPI for food at home (i.e., the price of food), food away from home expenditures, age of the household reference person, family size, children, race, education, region of residence, and interaction terms with the price of food. C equals the total non-durable expenditures from the CE. Once estimated in the CE, Blundell et al. (2008) invert equation (A1), and the estimated coefficients are used to

[2] See online appendix for Fisher et al. (2015) for details on service flow imputations.

impute consumption in the PSID. This method captures the change in the variance well but overstates the level of the variance and level of consumption.

Ziliak (1998) takes a different approach to estimating total consumption in the PSID. Rather than using another data set, Ziliak (1998) uses the PSID information on household wealth and income. First, he calculates savings by subtracting current asset holdings from next period's asset holdings. Using this measure of savings, he then subtracts savings from current income to generate composite consumption. Ziliak (1998) argues that composite consumption improves upon Skinner's (1987) imputed consumption because Skinner's measure may be unstable if relative prices are changing. We avoid this problem by estimating separate equations for each year of the PSID data. The coefficients are allowed to differ across years, which will reflect price changes.

In this work, our estimated equation expands the Skinner (1987) model, but we also follow the Blundell et al. (2008) method by including demographic characteristics in the estimated equation. To impute total consumption for the PSID, we will estimate the following using the CE:

$$ln(C) = \alpha_0 + X'\alpha_1 + \alpha_2 * food\ home + \alpha_3 * food\ away + v \quad \text{(A2)}$$

The dependent variable, *C*, equals total household consumption as described above. The vector *X* contains:

- a quartic in age of the reference person
- dummies for region of residence
- family size
- dummies for having one child, two child children, and three or more children (omitted is having no children)

- dummies for those with a high school education, those with some college, and those with at least a college degree (omitted is less than a high school degree)
- number of labor income earners
- whether the household owns or rents
- number of automobiles
- labor income decile
- government income decile
- whether the household has business income

Food at home and food away from home were not asked in the 1973, 1988, and 1989 PSID. We impute these values within the PSID before imputing consumption. We use the surrounding values of the variable to impute. For example, we use 1972 food at home and 1974 food at home to help impute 1973 food at home, along with demographic characteristics of the household. We follow a similar methodology to impute number of automobiles in 1973 and 1974, autos from 1987-1997, and rent in 1988 and 1989.

Following our earlier work in imputing income (Fisher, Johnson, and Smeeding, 2015), we use the multiple imputation methodology of Rubin (1987) and produce five estimates of consumption for each wave. Multiple imputation allows researchers to account for the extra uncertainty generated by the imputed values relative to reported values.

Judging Quality of the Imputation – Comparisons to the CE Survey

We first present the simplest comparison – the means of the distributions over time. Figure A1 displays the mean of CE consumption and the five PSID implicates from 1980-2017. The means match well as one would expect when using predictive mean matching for the imputation. The years when the CE re-samples (1986 and 2015 in particular), the CE mean exceeds the imputed

mean, likely because of the smaller sample sizes in the CE those years. Remember that we used three years of CE data to impute to the PSID in those years, which would smooth out the PSID.

Figure A2 presents the kernel density for CE consumption and one PSID implicate for 2017. The two overlap across most of the distribution. Figure A3 shows the Gini inequality from 1980-2017. The patterns tell the same basic story. Inequality is relatively flat from 1980-1995, while it rises from 1995-2007; inequality falls starting with the Great Recession. Figure A6 shows that our Gini is similar to the Gini from Attanasio and Pistaferri (2014) who use an alternative imputation method.

Judging Quality of the Imputation – Comparisons to the Reported Consumption

The PSID introduced a more complete definition of consumption in 1999 and then expanded it again in 2005. From 1999-2003, consumption is the sum total of food, housing, transportation, education, and child care. Beginning, in 2005, consumption also includes spending on travel, clothing, other recreation, home repair, home furnishings, and home phones.

We first compare the Gini coefficient between imputed and reported consumption. As expected, reported consumption inequality is higher than imputed inequality (Figure A4). The trends in inequality within cohorts generally, though there is more year-to-year variation in the reported Gini. The biggest difference is for the second oldest cohort, where reported consumption inequality is increasing while imputed consumption inequality is falling.

Next we compare mobility with reported and imputed consumption using the rank-rank correlation in Figure A5. The rank-rank correlation is higher using reported consumption, meaning mobility is lower with reported consumption. We expect reported mobility to be lower than imputed mobility because of the imperfection of the imputation. The patterns across cohorts generally match between reported and imputed mobility, and the trends over ages within a cohort generally match as well.

Figure A1: Mean of CE Consumption and Five PSID Imputed Consumption, 1980-2017

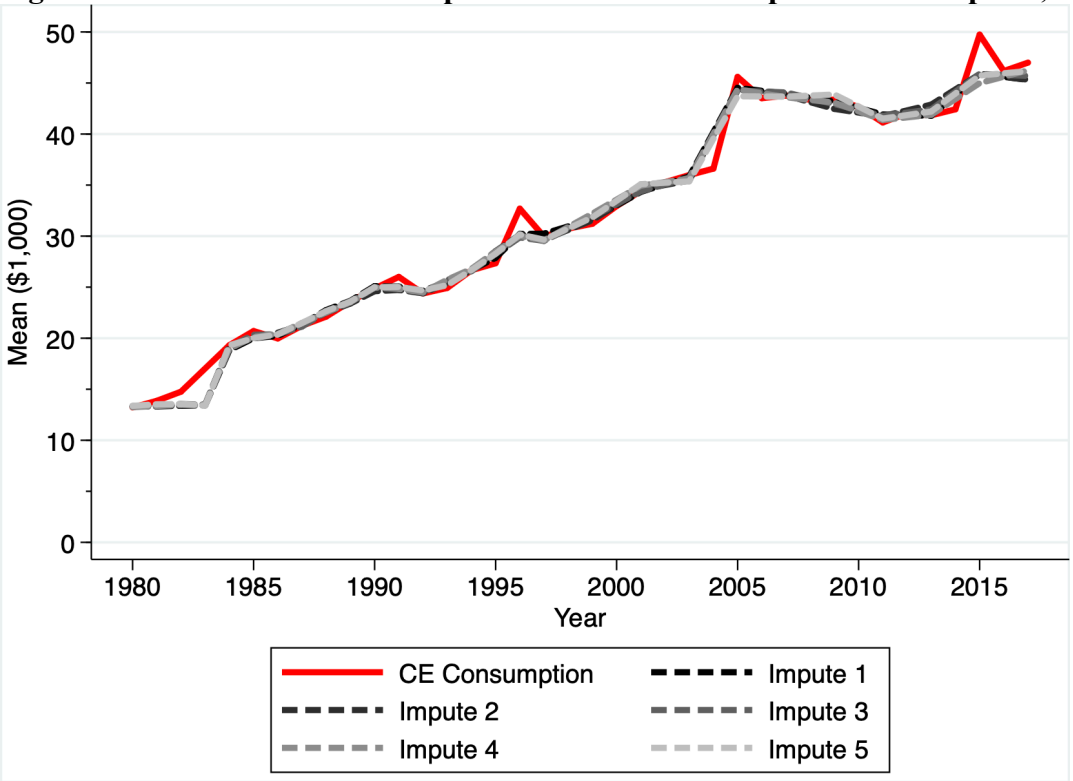


Figure A2: Kernel Density of CE Consumption and PSID Imputed Consumption, 2017

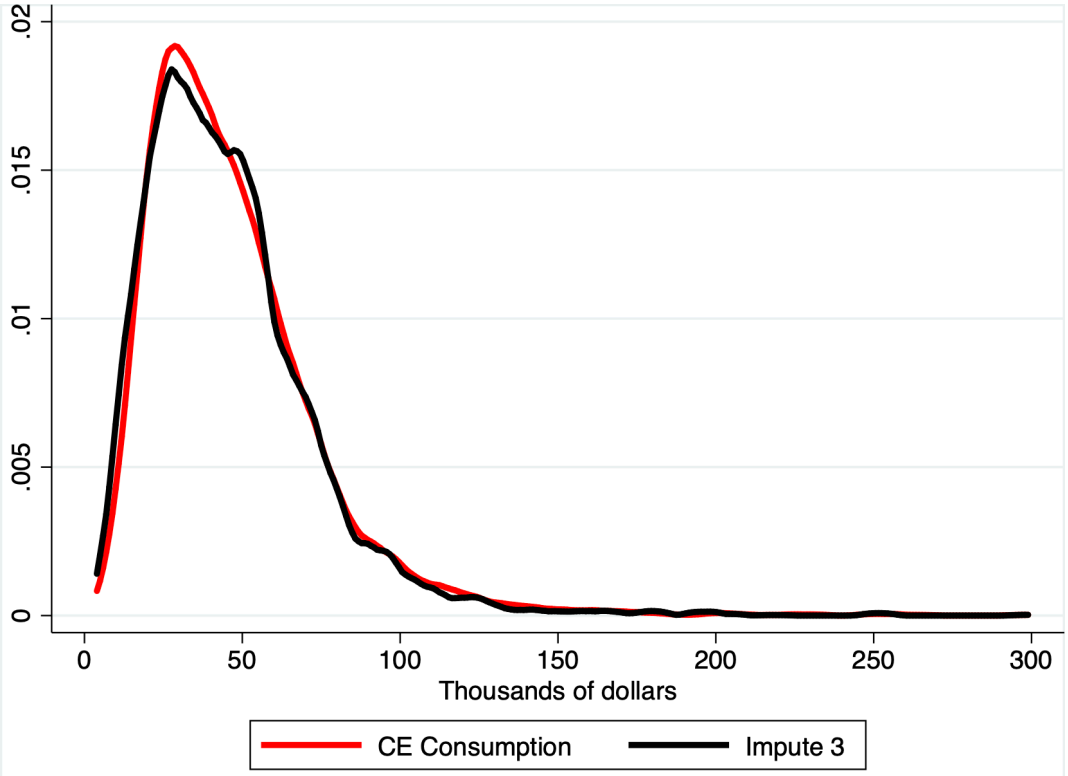


Figure A3: Gini of CE Consumption and Five PSID Imputed Consumption, 1980-2017

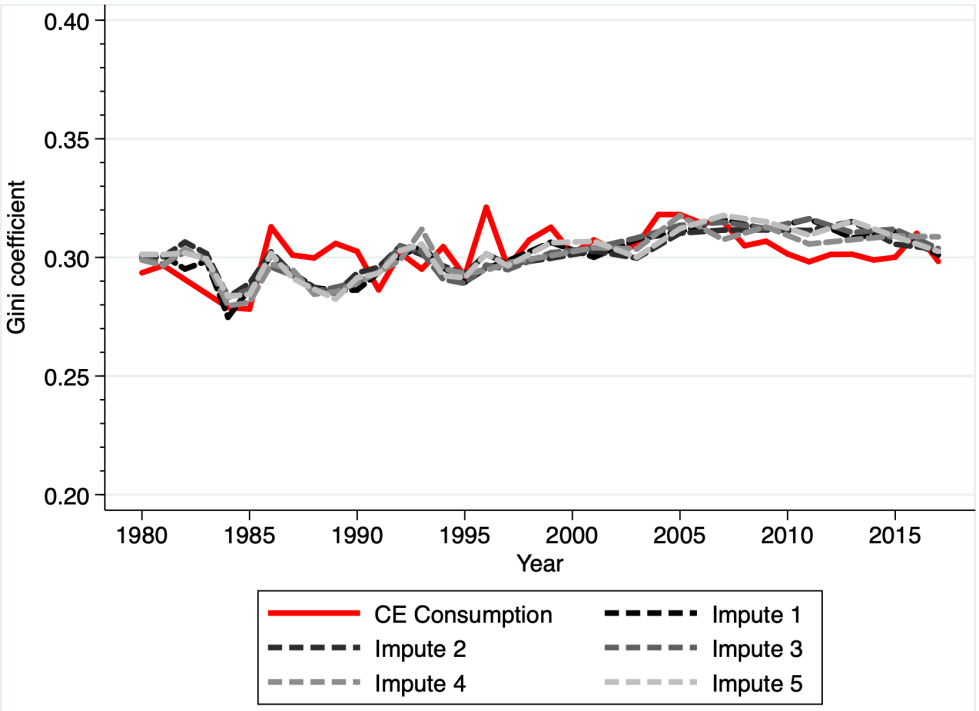
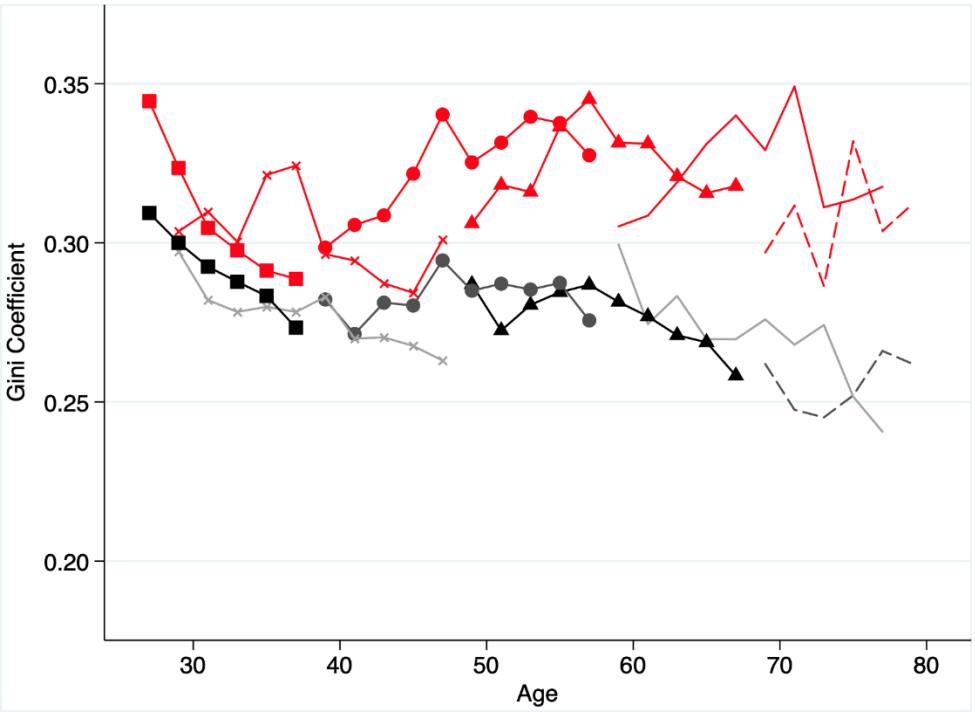
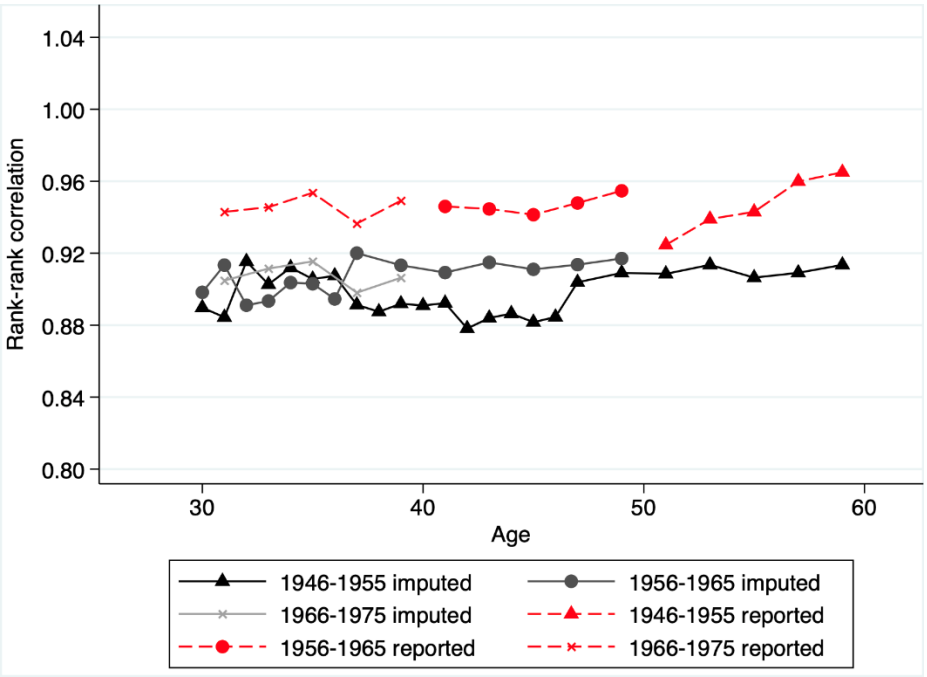


Figure A4: Gini Coefficient for Reported and Imputed Consumption by Cohort (PSID)



Notes: The red lines represent reported consumption inequality using the 1999 consumption definition. The black and gray lines represent imputed consumption. The figure uses 1999-2017.

Figure A5: Rank-Rank Correlation using Reported and Imputed Consumption (PSID)



Notes: The red lines use the 1999 consumption definition.

Figure A6: Comparison of Gini Coefficients for wealth (SCF), income (CPS), and consumption (Attanasio and Pistaferri) to those from PSID