Appendix: Data Sources and Consistency Procedures

This appendix provides detail on the sources of data used in this study, as well as the procedures adopted, to ensure consistency between the income and expenditure data derived from sources.

A. Data

1. Income

We obtain data on before-tax income for the DH model from the 2013 Survey of Consumer Finances (SCF). The SCF data include labor, capital, and transfer income for a representative sample of 6015 households. Labor income is defined as the income from wages and salaries. Capital income is the sum of income from a sole proprietorship or a farm; income from dividends; income from gains or losses from mutual funds or from the sale of stocks, bonds, or real estate; income from other businesses or investments, net rent, trusts, or royalties; income from other interest; income from child support or alimony; and income from other sources, including IRA, IRA/401(k) withdrawal, withdrawal from deferred compensation, and settlement of other employer-provided pension. Transfer income refers to the sum of Social Security benefits, unemployment, or workers' compensation; income from Temporary Assistance for Needy Families (TANF), Supplemental Nutrition Assistance Program (SNAP; food stamps), or other welfare or assistance such as Supplemental Security Income (SSI); and income from nontaxable investments such as municipal bonds.[1] Pension income is excluded from transfer income for consistency with the E3 and DH models, which do not include retirement income.

We derive after-tax income by applying information on tax liabilities from the National Bureau of Economic Research's TAXSIM model (Feenberg and Coutts 1993) to the SCF before-tax data. The SCF microdata do not provide information on place of residence, but this information is needed to determine household state tax liabilities in the DH model. To obtain state tax liability for each household, we randomly assign each household in SCF microdata to a state such that the proportion of households in each state matches the real population share in 2013.

TAXSIM does not calculate tax liability by income source. To obtain this more disaggregated information, we assign tax liabilities to labor, capital, and transfer incomes and calculate the share of before-tax income from each source for each household. Then we multiply these shares by the total tax liability. Specifically, we get total federal tax liability, total state tax liability, and Federal Insurance

[1] Nontaxable investments are defined as transfer income in TAXSIM. For consistency with TAXSIM, we include nontaxable investments in transfer income when calculating before-tax income.

Contributions Act (FICA) tax for each household using TAXSIM. In TAXSIM, FICA is the sum of employee and employer payroll taxes. In the E3 model, we include employee payroll taxes as a tax on labor income. Therefore, we allocate federal and state tax liabilities to labor, capital, and transfer incomes using before-tax income shares. We assume that payroll taxes divide evenly between the portions paid by employees and employers. Households with no income but positive tax liabilities are dropped. After-tax incomes by income source are then calculated as before-tax income minus the tax liability by source.

2. Expenditure

We obtain household expenditures on each consumer good using the 2013 Consumer Expenditure Survey (CEX) microdata collected by the US Department of Labor's Bureau of Labor Statistics (BLS). The CEX provides data on expenditures, income, and demographic characteristics of representative consumers in the United States. These data are collected through two surveys: the Interview Survey and the Diary Survey.[2]

To obtain a complete listing of expenditures for each household, we combine data from the two surveys. The Interview Survey focuses on large consumer goods, such as spending on housing, vehicles, and health care. It collects data on monthly expenditures of households that are selected as samples for five consecutive quarters, after which they are dropped from the sample. The Diary Survey gives greater focus to small, frequently purchased goods, such as expenditures on food, beverages, and personal care products. This survey collects data on weekly expenditures of different households that are followed for only two weeks.

Before integrating data from the two surveys, we calculate a weighted expenditure cost of each consumer good for each household in both surveys.[3] Then we define subgroups using five demographic characteristics: age, education level, marital status, family size, and income decile. Using these subgroups, we combine weighted expenditures for each household in the Interview Survey with each subgroup's average weighted expenditures from the Diary Survey. In this way, we create a large dataset of representative households with combined expenditures from both surveys.

B. Achieving Income-Expenditure Consistency

To produce a complete dataset with both income and expenditure information for each SCF household, we match expenditure data from the CEX to each SCF household.[4] Specifically, we match

[2] There are overlapping consumer good categories across the two surveys. In choosing which data to take from which survey, we follow Bureau of Labor Statistics guidelines.
[3] Each household in the microdata represents a given number of households in the US population.
[4] Our method resembles the approach of Cronin et al. (2017).

expenditure patterns in the CEX to a similar SCF household. First, we define 720 "CEX cells" using demographic characteristics for each household: education (less than high school or high school degree; some college, no degree; college degree), marital status, age (<30, 31–40, 41–50, 51–60, 61–65, 66+), family size (with, without children), and before-tax income deciles.

For each SCF record (there are five records for each household to help maintain anonymity), we assign the SCF record to its corresponding CEX cell. If there is more than one CEX record in the CEX cell, the averages of expenditure data of all CEX records in the cell are used as the expenditure information for the SCF record. If the CEX cell is empty, we do a nearest-neighbor match based on before-tax income, assigning the SCF record to the CEX cell with the nearest before-tax income (and identical nonincome demographics). To find the nearest neighbor, we compare the "distances" between the SCF record's income decile and (1) nearest lower-income neighbor, (2) nearest higher-income neighbor, and (3) "average" neighbor, the average decile of the nearest lower- and nearest higher-income neighbors. We assign the SCF record to the cell with the shortest distance. SCF records with no close-neighbor decile, defined as the nearest neighbor at least 2 deciles away, are dropped from our data set.[5] SCF records that have no corresponding CEX cell with the same demographics are dropped from our data set as well.[6]

The SCF also does not include information on the level of savings for each household. From the CEX, we obtain the level of savings by subtracting total expenditures from after-tax income.[7] The implied savings rate is then matched to the SCF households using the same matching algorithm described above for expenditure shares. Total expenditures for each SCF record are equal to after-tax income less savings. Expenditure levels by good are equal to total expenditures times the matched CEX expenditure shares.

[5] There are 174 SCF records (out of 30,075) that do not have a nearest neighbor.
[6] Only 10 records in the SCF are dropped because CEX cells with identical nonincome demographics do not contain any CEX households.
[7] The CEX tends to overstate savings rates by household, as tax liabilities are generally underreported in self-reported surveys such as the CEX (Metcalf et al. 2012). Savings rates are scaled to match E3 savings rates in the calibration procedure.