# The PSID Financial Inequality Project: About the Code

Originally Drafted by Steve Wendel, Morningstar  
Last Updated 25 August 2021

The PSID Financial Inequality Project is a Python application for analyzing the Panel Study of Income Dynamics (PSID) and generating a standardized dataset for research on wealth and income inequality. Morningstar developed the application for its research paper on the racial wealth gap in the United States, [*Unpacking Racial Disparities in Savings Rates*](https://www.morningstar.com/content/dam/marketing/shared/pdfs/Research/Savings-Rates-Differences-By-Race.pdf) (a short [summary article is here](https://www.morningstar.com/articles/1055633)). We have made code, data, and results from the Project freely available on our Github site, <https://github.com/Morningstar/PSID>/PSID.

This document outlines what the PSID is, and how the code works to process it.

## About the Panel Study of Income Dynamics (PSID)

The PSID is the longest running national survey of household finances in the world — started in 1968, the study traces the finances and membership changes of a nationally representative survey of Americans. It is extensive (over 5,000 variables), with a large sample size (5,000-10,000) and includes an oversample of African Americans, which makes it quite valuable especially for research on racial wealth inequality. The PSID has also built a history of trust and expectation between the participant families and the organizers, which has facilitated a high-response rate over decades of data collection.

Unfortunately, the PSID is frankly very difficult to use: it’s richness is also its greatest limitation. Variable names and meanings change over time, and each year of data has its own nuances. The PSID allows the tracing individual household members over time, which is complicated but highly valuable; individual-level tracing is something that most other surveys don’t even attempt because of the complexity.

Three research teams, in particular, have used the PSID to study wealth accumulation and savings rates — Dynan et al. (2004), Gittleman and Wolff (2004), and Juster et al. (2006). All three used the then-new Wealth Supplement to the Panel Study of Income Dynamics: an additional section added to the PSID starting in 1984 that provides detailed information on flows in and out of various assets. When these researchers used the PSID, they did so only for a short time period: two or three waves of the survey. Analyzing the PSID across time appears to have been a major obstacle blocking researchers using this rich dataset thus far.

In this application, we automate the process of standardizing the PSID, and create a unified dataset spanning over thirty years of detailed household wealth information: from 1984 to 2019. Then, we build upon the approach first presented in Juster et al. (2006), Gittleman and Wolff (2004) and Dynan et al. (2004) to separate capital gains from active savings, and analyze the dynamics of saving rates disparities and racial wealth inequality over time.

## **Overview of the Code and Output Files**

The PSID Financial Inequality Project is divided into a set of classes, which are called as needed from a central “Controller.py” script. The Controller takes a parameter file (e.g., params\_AllInequalityAnalyses\_EnrichedPop.py) which specifies both which stages in the process to execute, and parameters specific to those stages.

Conceptually, we can think about the process in 9 stages:

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage** | **Description** | **Folder** | **Main Classes Used** |
| Setup | Determine which parts of the process to run | Controller | Controller, param files |
| 1 | Extract SAS Files and Convert to CSV | PSIDProcessing | RawLoader |
| 2 | Extract Variables We Need, For Each Year | PSIDProcessing | Extractor |
| 3 | Standardize Variable Naming & Coding, Across Years | PSIDProcessing | FamilyDataRecoder, IndividualDataRecoder |
| 4 | Calculate taxes using TAXSIM | TaxSim | TaxsimFormatter |
| 5 | Create cleaned, inflated, two-period longitudinal files. | SavingsRates | InequalityDataPrep |
| 6 | Calculate Savings Rates & Capital Gains | SavingsRates | CalcSavingsRates |
| 7 | QA the processed data | DataQuality | CrossSectionalDescriber; LongitudinalDescriber |
| 8 | Conduct Analyses about savings rates used in MStar Report | MStarReport | SWAnalysisPerPeriod; SWAnalysisLongTerm; AggregatePopulationAnalyzer |
| 9 | Replicate Existing Research on PSID | Replication | DynanAnalyis; GittlemanAnalysis; ZewdeAnalysis |

Stages 1-4 are general purpose tools for using the PSID in a standardized, cleaned way. Stages 5-7 further process the specific data needed for analyzing wealth accumulation: active savings rates, capital gains, inheritances, etc. Step 8 outputs the particular analyses of wealth accumulation used in Morningstar’s report. Step 9 compares those results with prior research in the field.

In addition to the final output for Morningstar’s report (in a Excel file called “FiguresForPaper”), the code is designed to generate extensive data at each step along the way, so that other researchers can either depart from subsequent steps and generate their own analyses, or dig into the intermediate data to check for errors, etc. The output files are located as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage** | **Description** | **Output Directory\*** | **Output Filenames\*** |
| 1 | Extract SAS Files and Convert to CSV | inputData/[PSIDFamilyData\_Prepackaged| PSIDFamilyFiles\_2021Extract]/fam[Year] | FAM[Year] |
| 2 | Extract Variables We Need, For Each Year | outputData/intermediateStages/extractedPSID | extractedPSID\_[Year] |
| 3 | Standardize Variable Naming & Coding, Across Years | outputData/intermediateStages/ mappedAndrecodedPSID | extractedPSID\_Mapped\_Recoded\_[Year] |
| 4 | Calculate taxes using TAXSIM | outputData/intermediateStages/taxsim (Tax Data Only); outputData/processedPSID (Combined with PSID) | taxsim\_[Year]; extractedPSID\_withMRTax\_[Year] |
| 5 | Create cleaned, inflated, two-period longitudinal files. | outputData/ inputForMorningstarAnalysis\_enrichedPop | YearData\_[Year]\_as\_[InflatedToYear]; TwoPeriod\_[StartYear]\_[EndYear]\_as\_ [InflatedToYear] |
| 6 | Calculate Savings Rates & Capital Gains | outputData/morningstarOutput\_enrichedPop | WithSavings\_TwoPeriod\_[StartYear]\_ [EndYear]\_as\_[InflatedToYear] |
| 7 | QA the processed data | outputData/morningstarOutput\_enrichedPop/ descriptives | Tables\_[StartYear]\_[EndYear], ActiveSavingsVars\_\*, QAWealthYears\_\* |
| 8 | Conduct Analyses about savings rates used in MStar Report | outputData/morningstarOutput\_enrichedPop/ analyses | Many, but see FiguresForPaper\_1984\_to\_2019.xlsx especially |
| 9 | Replicate Existing Research on PSID | outputData/ | CombinedDynan\_\*; CombinedGittleman\_\*; Table\* |

## **Data Analysis Process: Detail**

Here is a more detailed description of each of those 9 steps used to analyze the PSID:

1. **Extract SAS Files and Convert to CSV.**  Extracts the annual family data and wealth data, the individual data file, and the variable cross-reference (“crosswalk” file) from the University of Michigan. Throughout this analysis, we use downloads from the PSID data center as of October 15th, 2020. The downloads were formed by selected each year individually (one per download), and then selecting all variables. In the development process, we found irregularities in the pre-packaged data that is posted on the PSID FTP site for 1994. The net wealth information in particular in the pre-packaged data has significant errors. Out of an abundance of caution, we use the current (data center) data for all years.
2. **Extract Variables We Need, For Each Year.** Using the parameter file “Controller/VarsForInequalityAnalysis”, select the relevant variables from each year. The parameter file provides ONE name for each variable from any year; for consistency, the current file uses 2017 variable names when available. The code then automatically determines the name of that variable in other years. If you want to include new variables in your dataset from the PSID, this is where you need to include them first.
3. **Standardize Variable Naming & Coding, Across Years.** Create standardized, annual data files.
   1. Recode each year’s dataset to have consistent, human-readable variable names and meanings (as given in “Controller/VarsForInequalityAnalysis”); remap don’t knows and refuses to missing data.
   2. Where concepts or definitions have changed, uses manual mappings to make the meaning of each variable as consistent as possible across the entire study period.
4. **Calculate taxes using TAXSIM.** The PSID supplies estimated family-level federal taxes from 1970 to 1991, for which they called the NBER’s TAXSIM software. Since tax information is important for a savings-rate analysis, we replicate that process for all other years. The program automatically converts the PSID data into TAXSIM format, calls TAXSIM’s FTP interface, and integrates both state and federal taxes back into the standardized PSID-derived annual data files.
5. **Create cleaned, inflated, two-period longitudinal files.**
   1. **Create:** For each year with wealth data (every five years from 1984 to 1999; biannually afterwards), combine that year with the subsequent year, by household. Households are defined as having the same head of household/reference across the two periods. We use the PSID’s “individual file” to track the head of household across years, as described in the PSID documentation.
   2. **Inflate:** Inflate all values to 2019, using the CPI-U.Stock (point-in-time) variables are inflated to 2019 using the CPI-U for their year; variables providing a flow between years are inflated using the half of the inflation during that period, plus the inflation from the end-year to 2019.
   3. **Clean:**
      1. Remove the non-representative “Latino Sample” from 1991-1995. During this period, the PSID included a subset of Latino immigrants that was not representative of the overall immigrant population, and was subsequently dropped by the PSID itself. All families from the original PSID sample were weighted by the PSID to be nationally representative at the time (and remain representative of their descendants, each wave).
      2. For each two-period longitudinal file, drop records where the household did not answer the survey in one or both years.
      3. For each two-period longitudinal file, drop records where the head of household changes. As described in both Gittleman and Wolff (2004) and Dynan et al. (2004), this is necessary to keep a consistent picture of family finances over time. For the 1984-1989 period, for example, this process drops 1,888 out of 6,381 families; as shown in Gittleman and Wolff (2004), dropping these families create a small upward bias in family wealth.
      4. Fix the ‘moved’ variable used in savings rate calculations. As described in Dynan et al. (2004): **“**If a head indicated they hadn’t moved during the prior year, yet they switched their own/rent status between consecutive years, we forced the move dummy to one [213 observations in 1989-1994]. Or if someone was a renter yet listed a value of a house, we set the house value and mortgage to zero [one observation in 1989-1994]”. In this analysis, we apply the same logic up through 2019.
      5. Apply the house-value adjustment calculation developed by Juster, Smith and Stafford (1999) to separate true home-value appreciation from the purchase of a more (or less) expensive home, using the updated ‘moved’ variable.
      6. Juster, Smith and Stafford (1999) note that the 1984 “Other Assets” variable has 7 apparently miscoded entries, at 9 million. The authors, Gittleman and Wolff (2004), and this analysis exclude those entries from the 1984-1989 dataset.[[1]](#footnote-1)
      7. Drop households where real, pre-tax income averaged across the period is less than $1,000: i.e. households where no one is receiving income. Saving rates are not meaningful for households without income.
      8. Drop households where the head of household is under age 20 or over age 70. Note — a version of the analysis was also conducted with all retirees removed (regardless of age): the substantive results are the same.
   4. Where mean calculations are provided below, two methods are used to remove outliers and bad data: trim off the top and bottom 1% of change in net wealth (following Gittleman and Wolff 2004) and removing cases where the residuals are beyond 2 standard deviations (both provide the same substantive results). When median calculations are provided, we use the untrimmed population.
6. **Calculate Savings Rates & Capital Gains**: See below for description.
7. Conduct additional cleaning and consistency checks.
   1. Look for uncoded values (eg 998s and 999 that mean DK and No Answer, which should be changed before other analyses are run).
   2. Look for major shifts across time in a variable (usually caused by a change in the underlying meaning or coding of the variable), and within a household over time (usually caused by bad data)
   3. Export the exact tables from Gittleman and Wolff (2004), to allow for model docking: creating a baseline scenario in which the results are the same and verified, before changing the scenario to explore other research questions of interest.
8. **Conduct Analyses about savings rates used in MStar Report.** Read Morningstar’s report for more information.
9. **Replicate Existing Research on PSID.** Early in the coding process, the authors replicated analyses from three existing papers using the PSID, to determine exactly how the results here differed and why. While the algorithms and tables generated are correct and match the respective research papers, unfortunately the input data has changed somewhat – and the replication code has not been kept up to date with the new variable names, etc.

## **Savings Rate Calculation**

Building on Dynan et al. (2004) and Gittleman and Wolff (2004), we can track and distinguish between seven, household level savings-related concepts in the PSID:

1. **Change in net wealth**: the difference in the real value of all assets, minus debts, from the start of the period to the end of the period.
2. **Capital gains**: The change in net wealth attributable to the underlying change in the price of the assets.
3. **Gross savings**: The change in net wealth, after capital gains are taken into account.
4. **Large gifts and inheritances**: the inflow of assets due to large ($10,000) gifts or inheritances.
5. **Ongoing smaller dollar support from family members:** the inflow of assets due to smaller (< 10,000) gifts and support. This was not included in Gittleman and Wollf (2004)’s or Dynan et al. (2004) papers, but has been mentioned by subsequent researchers as potential factors driving wealth inequality.
6. **Changes in family composition**: The net value of assets moved into the household during the period, minus assets moved out of the household.
7. **Active savings**: Apparently intentional contributions to increase the value of assets or decrease debts — by any means from contributing cash to a savings account, buying a business, paying off a student loan, to renovating a house.

Active savings is calculated as gross savings, minus large gifts and inheritances, small gifts and support, and changes in family composition. Gross savings and capital gains are calculated per person, per asset class, as described in the next section.

### **Asset Classes and their Treatment**

In this analysis, we consider the ten asset top-level classes that are covered by the PSID where both current value and flow data is provided or is calculatable.

1. Home Equity on Main Home
2. Other Real Estate
3. Vehicles
4. Farms or Businesses
5. Private Annuities and IRAs
6. Employer Sponsored Retirement Plans (Defined Contribution and Defined Benefit)
7. Equity Investments outside of IRAs, Annuities, and Employer-Sponsored Plans.
8. Cash Accounts and Fixed Income Investments, outside of IRAs, Annuities, and Employer-Sponsored Plans.
9. Other Assets (Art, etc.).
10. Other Debts (Credit Cards, Store Debt, Student Loans, Medical Expenses, Legal Bills, Loans from Relatives, Misc. including Personal loans, bank loans, pay-day loans, Overdue Taxes or Rent).

For each asset-type, we calculate three values: savings, capital gains, and open/close transfers. Savings are recorded (or implied) flows into an existing asset during the period. Open/close transfers are a special type of saving (and dissaving) that occurs when the household acquires or disposes of an asset type entirely. Capital gains are calculated as passive increases in value based on the starting value of asset. The aggregate, household level values for capital gains (discussed above) is the sum of capital gains; aggregate, household level values for gross savings is the sum of asset-level savings, and similarly for household open/close transfers. Active savings is defined as (gross savings + open/close transfers) + net movement of funds out of the household — large gifts into the household — small gifts into the household.

The distinction between existing-account saving and open/close transfers allows us to more accurately calculate asset-level savings rates. At an aggregate level the sale of a business (dissaving), and movement of all of the resulting funds into an investment account (saving) would, and should, register as no change in household level saving: the asset mix has changed, but not the total value. However, at an asset-level, the sale of the business would show as an extreme dissaving event — potentially many multiples of the household’s current-year income. These extreme events would warp any calculation of year-by-year savings rates, and thus are separated out. Both are analyzed and accounted for, but separately, to allow for better inferences of intent.

How capital gains is calculated depends on the asset class. Six of the asset classes included in the PSID offer both balance and inflow/outflow data about the asset: primary houses, other real estate, businesses/farms, brokerage stocks, employer retirement plans, and private retirement plans. When full information is available, we can calculate capital gains as the total change in value minus the net flow (dis/saving). For the three investment-related asset types (employer retirement plan, private retirement plans, brokerage accounts), when flow information is *not* available, we use historical Morningstar data on stock and bond performance to infer capital gains rates, calculate capital gains, and then savings. Four asset classes have only balance information, but one can make inferences about flows after making an assumption about capital gains rates: checking/savings (assumed capital gains rate: CPI-U) , vehicles (0% nominal[[2]](#footnote-2)), ‘other assets’ (1% + CPI-U), and ‘other debts’ (CPI-U).

Missing data is a significant issue throughout the PSID’s asset-level data. We handle it by first classifying the data for each household, each asset type into one of 9 categories, as shown in Exhibit A-1. This is a significant departure from Gittleman and Wolff (2004) and Dynan et al. (2004), who effectively treated missing flow or balance data as 0; in our analysis we found that that approach unintentionally generated artifacts — extreme and inaccurate savings rates and capital gains values (positive and negative), depending on the pattern of missingness.

We believe this approach is more appropriate than a cross-household imputation, because of the variables that would be imputed are strongly correlated with each other with each household: for example, if the balance of an investment account is missing in one year (and the account is not closed), the most appropriate value to use is the capital-gains adjusted balance from another year for the same household, rather than a predicted value based on other similar households.

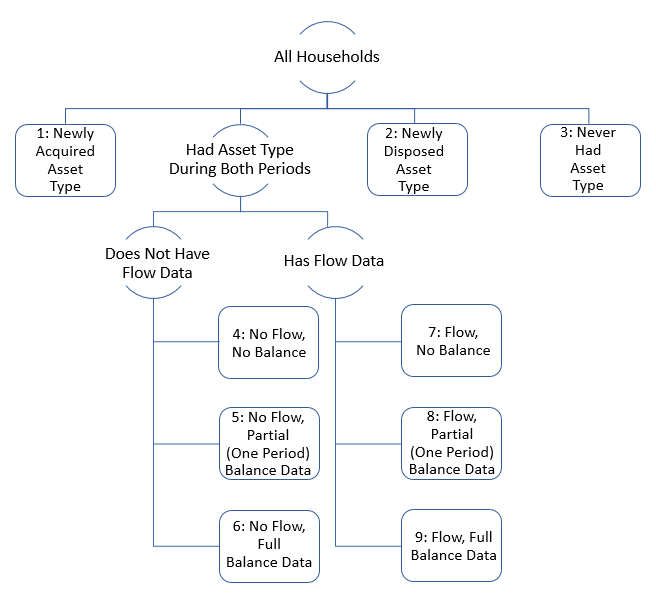
For each category, we fill in data and compute asset-level savings and capital gains as follows:

Figure 4: Process used for cleaning asset data and calculating capital gains and savings

1. Newly Acquired Asset
   1. For example: the household had no vehicle at the start of the period, and had one at the end of the period.
   2. Saving: Zero
   3. Capital Gains: Zero
   4. Transfer: Ending Balance for the Asset Type
   5. Caveats: This process categorizes mid-period contributions/withdrawals to the asset as transfers. The household saving rate is accurate, but the asset-level rate will be over/underestimated.
2. Newly Disposed Asset:
   1. For example: the household had a business at the start of the period, and sold it by the end of the period.
   2. Saving: Zero
   3. Capital Gains: Zero
   4. Transfer: Starting Balance for Asset Type
   5. Caveats: This process categorizes mid-period contributions/withdrawals to the asset as transfers. The household saving rate is accurate, but the asset-level rate will be over/underestimated.
3. Never Had Asset:
   1. For example: a household that did not own real estate, beyond a primary home, at the start or end of the period
   2. Savings: Zero
   3. Capital Gains: Zero
   4. Transfer: Zero
4. No Flow Data Available, no Balance Data Available:
   1. For example: a household that indicated they do have an annuity (at the start and end of the period), but provides no data on its value or flow in or out of the annuity during the period
   2. Savings: Unknown
   3. Capital Gains: Unknown
   4. Transfer: Zero
5. No Flow Data Available, One Period of Balance Data Available:
   1. Savings: Unknown
   2. Capital Gains: Unknown
   3. Transfer: Zero
   4. Caveats & Post-Processing: The missing balance data can cause a large difference in aggregated household net worth — this is an existing problem with the PSID data's net worth data. To more accurately measure net worth, we fill the missing balance information, assuming that the balance hasn't changed. This is a post-processing step, to avoid large swings in household net worth; it does not affect the savings and capital gains calculations.
6. No Flow Data Available, Both Periods of Balance Data Available:
   1. For example: A household that indicates they have an IRA, and provides balance information for it, but does not indicate any flow in or out. In the PSID a lack of flow data can be because there was no flow, or because the household failed to answer the question. For IRAs and annuities, for example, this represented 485 out of 7,304 households in 2017-2019, and for Businesses and Brokerage accounts, this occurred roughly 250 times. It is not common for other accounts.
   2. Capital Gains: Calculate the implied capital gains rate, assuming the missing flow data means there was zero flow. If the absolute value of the implied rate is below a set threshold (here: 20% per annum), then the change in balance is all counted as capital gains. If not, set Capital Gains to Unknown.
   3. Savings: If the implied capital gains rate is below the threshold, savings is zero. If not, Savings is Unknown.
   4. Transfer: Zero
7. Flow Data Available, No Balance Data Available:
   1. For example: A retirement plan where the household provides a contribution rate, but nothing else. This is a relatively common problem for retirement plans (with 568 out of 7,304 households in 2017-2019, for example), but is not common otherwise.
   2. Savings: The net flow indicated by the household
   3. Capital Gains: Unknown
   4. Transfer: Zero
8. Flow Data Available, One Period of Balance Data Available:
   1. For example: A retirement plan where the household provides a contribution rate, but only one balance. This is a relatively common problem for retirement plans (with 720 out of 7,304 households in 2017-2019, for example), but is not common otherwise.
   2. Savings: The net flow indicated by the household
   3. Capital Gains: For investment-based asset types (workplace retirement, private retirement, and brokerage account) use Morningstar’s data on investment returns for that period (workplace & private retirement: assume a 70-30 allocation; brokerage: assume 100% allocation to stocks) to calculate capital gains from the one available balance, adjust adjusting for savings.
   4. Transfer: Zero
   5. Caveats & Post-Processing: As for category #5, we fill in the missing balance information (after accounting for savings and capital gains), to create a more accurate and stable measure of household net worth.
9. Flow Data Available, Both Periods of Balance Data Available:
   1. For example: the household indicates they have a 401(k) (at the start and end) of the period, provides opening and closing balances, and says how much they are contributing as a % of income.
   2. Savings: Reported net flow
   3. Capital Gains: Net change in asset value minus savings.
   4. Transfer: Zero

The analysis of primary homes differs from other asset types. In particular, the housing data is available for every year of data, instead of only on the sample’s “wealth years”. Capital gains and saving are in each year and then summed. If family did not move, the capital gains in each year equals the rise in the value of the home and saving equals the reduction in mortgage principal. In years in which the family moves, the change in the net value of the house is considered saving. In addition, the value of additions or improvements, which is assumed to apply to main home, is added to saving as well.

1. Dynan et al. (2004) included four additional steps of data-removal, which do not appear to be relevant or useful for this analysis. They exclude all households with less than 1,000 in income (1994 dollars), savings of over 750,000 (over the 5-year period of 1984-1989 or 1989-94), with heads of households aged over 59 or under 30, and the 24 people who did not respond to the question on whether or not they had moved in the prior 5 years.  
    [↑](#footnote-ref-1)
2. Cars actually depreciate value over time, at a historical average of approximately 5.5% per annum. However, cars serve as a mix of consumption good and investment. Since other consumption goods, from refrigerators to food, are treated as *expenditures* and not saving in this analysis, we have tried to separate out the depreciation of cars as an non-saving-related expense. Practically speaking, since cars are the largest depreciating asset for most households, when we include car depreciation in capital gains, the median capital gains for all years and races, and for most income levels, is negative. [↑](#footnote-ref-2)