

## ADDITION TO NLSY79 USER GUIDE

### Overview of Revisions to Asset and Debts:

In the spring of 2008 a revised set of asset and debt variables were released to the public. These revised asset and debt variables fixed a number of problems with the NLSY79 data by eliminating some implausible outliers, generating uniform topcodes for all rounds, and constructing a total net worth variable. This section of the guide provides details on revision process.

### *What Users See:*

Prior to the spring 2008 release users saw a single asset or debt question for each item in the wealth section of the questionnaire. For example in 1987 the questionnaire asked each respondent who owned a home or apartment the market value of their residential property. The questionnaire asked respondents “About how much do you think this property would sell for on today's market?” Until the spring of 2008 the respondent answers were found in a single 1987 variable that had the following R and Q numbers:

R23627.00 [Q1947] (TRUNC)

After the revision was done, two more asset variables were added to the data set based on the same underlying property responses. The two new variables are

R23627.01 [\*Created] (TRUNC) (REVISED)

And

R23627.02 [\*Created] (TRUNC) (IMPUTED)

What is the difference between R23627.00, R23627.01 and R23627.02? The variable that ends in (.00) R23627.00 is the original variable in the dataset and is left so that researchers can reproduce previous results. The variable that ends in (.01) R23627.01 is a new variable which uses a revised topcoding algorithm. By revising the variable researchers are now provided with some extra information that was not available before. The variable that ends in (.02) R23627.02 is a new variable which imputes missing and unknown responses if possible as well as using the revised topcoding algorithm.

There are two new variables because some users will not want to use imputed data. The (.01) variables are cleaned and re-topcoded but do not have any imputed values. The (.02) variables have as many missing or unknown values imputed as possible. In general the survey staff recommends that users without a strong preference should use the (.02) asset or debt value that ends in the label “(TRUNC) (IMPUTED).”

Table 1 gives an example using the 1987 property value question of how seven different types of cases were handled by the revision and imputation process. Please note the “\$” are not in the NLSY79 data but are added to make it easier to read the table.

Table 1: Examples of how NLSY79 Asset/Debt Data Were Modified.

Public ID	Original R23627.00	Revised R23627.01	Imputed R23627.02	Explanation
200	\$150,001	\$276,984	\$276,984	Originally above the topcode and the value is still above the topcode but the topcode is now higher, revealing more information.
40	\$150,001	\$153,000	\$153,000	Originally above topcode and now below topcode. Value is no longer topcoded.
9083	-1	-1	\$93,333	Originally refused. Now contains an imputed value
205	-2	-2	\$276,984	Originally a don't know. Now contains an imputed but since the imputed value is above the topcode the topcode is used as the value.
526	-3	-3	\$100,000	Originally an invalid skip. Now contains an imputed value
2	-4	-4	\$0	Originally a valid skip. Since valid skip means does not have the asset the item is changed to zero.
1336	-2	-2	-2	Originally a don't know. Since it was not possible to impute value, the value was left as a don't know.

Not every asset or debt variable has a new revised or imputed offspring. Instead to keep the project manageable only 15 asset/debt categories were created in each year. These 15 categories match up exactly with the categories found in the NLSY79 wealth module that was used in the 1990s. The categories are: Home Value, Mortgage Value, Property Debt Value, Cash Saving, Stocks/Bonds, Trusts, Business Assets, Business Debts, Car Value, Car Debt, Possession Value, Other Debt Value, IRA, 401K, Certificate of Deposit Value. However starting in 2000 and then in 2004 the wealth module became more complex. For these later rounds each asset/debt category corresponds to multiple individual asset/debt variables.

For example, in 2004 respondents were asked to report the values of two homes. Their values are combined to form the “home value” category. Similarly, in 2004 the “stocks/bonds” category represents the individually-reported values of government bonds, mutual funds, life insurance surrender values, stocks, corporate bonds, and money owed to the respondent.

***Details of the Revision and Imputation:***

In addition to creating these combined variables the NLS asset and debt revision project did six other steps. This six step process started off with cleaning the raw data and culminated in a new

net worth variable and new top coding for most respondents. The details of the six steps are as follows.

*Step 1 – Cleaning Raw Data:* The original raw data has a number of out of range codes. These out of range codes were originally given the top code value when released to the public. Examination of the out of range cases suggests most of these out of range flags were mistakes and not actually out of range. Most of these out of range codes occurred in the 1988 and 1989 surveys, but this issue arises in other PAPI years. All out of range codes were changed to an “invalid skip” (-3) in the revised (.01) variables. If possible these variables were then imputed in the (.02) variable). Researchers are able to determine which items were incorrectly marked as out of range by looking for items that were top coded originally and then changed to a -3 value in the revised (.01) variables.

*Step 2 – Unfolding brackets:* Unfolding brackets were used for four asset/debt categories in 2000 and for all categories in 2004. These unfolding brackets were not used prior to 2000. Unfolding brackets are used if a respondent fails to report a particular asset’s or liability’s value. For example, suppose a respondent refuses or does not know the value of his certificate of deposit (CD). The respondent is first asked if his CD is worth more than an entry amount, which is \$10,000 for some respondents and \$20,000 for others. If the value is not above the entry amount, the respondent is asked if the value of his CD is \$5,000 or more. If the value is above the entry amount, he is asked if the value would amount to \$30,000 or more. These three questions result in four potential reported ranges: below \$5,000; between \$5,000 and the entry amount; between the entry amount and \$30,000; and above \$30,000.

Whenever an unfolding bracket is used, we replaced the reported range with the median value among respondents whose reported value falls in the given range. For example, respondents who revealed via unfolding brackets that their CDs are valued below \$5,000 were assigned the median CD value among all responses who report directly (not via unfolding brackets) a value between \$0 and \$4,999. The 2004 median values used for each bracket for each asset/debt category are shown in tables 2 and 3.

Table 2: 2004 Median Values Used to Impute Unfolding Brackets.

<i>Asset/debt item</i>	<i>Low</i>	<i>Middle 1</i>	<i>Middle 2</i>	<i>High</i>
Collections	\$1,500	\$7,000	\$15,000	\$40,000
Items worth more than \$1k	\$2,000	\$7,000	\$15,000	\$50,000
Credit card debt	\$1,500	\$7,000	\$15,000	\$40,000
Student loans for R/SP	\$1,500	\$7,000	\$15,000	\$40,000
Student loans for children	\$2,300	\$8,000	\$15,000	\$35,000
Debt to business	\$700	\$7,000	\$16,000	\$50,000
Other debts	\$2,000	\$7,000	\$16,500	\$50,000
Business value	\$1,500	\$6,500	\$20,000	\$200,000
Business debt	\$3,000	\$8,000	\$17,000	\$140,000

Cash	\$1,000	\$7,000	\$20,000	\$50,000
CDs	\$2,000	\$8,250	\$20,000	\$55,000
Government bonds	\$1,000	\$6,000	\$15,000	\$50,000
Mutual funds	\$2,000	\$9,000	\$20,000	\$60,000
Life Insurance	\$2,000	\$7,425	\$20,000	\$100,000
Stocks	\$2,000	\$8,000	\$20,000	\$80,000
Corporate bonds	\$500	\$6,500	\$14,000	\$100,000
Money owed to R	\$1,000	\$6,000	\$16,000	\$50,000
Home value	\$2,000	\$10,000	\$20,000	\$160,000
Mortgage	\$2,500	\$10,000	\$20,000	\$100,000
Property debt	\$2,000	\$6,500	\$20,000	\$44,500
2 <sup>nd</sup> Home value	\$3,000	\$10,000	\$20,000	\$140,000
2 <sup>nd</sup> Mortgage	\$2,500	\$7,750	\$20,000	\$100,000
2 <sup>nd</sup> Property debt	\$1,500	\$10,000	\$20,000	\$90,000
Value of cars, trucks	\$2,200	\$9,000	\$20,000	\$40,000
Debt on cars, trucks	\$2,500	\$8,000	\$17,000	\$32,000
Value of other vehicles	\$3,000	\$8,000	\$16,000	\$45,000
Debt of other vehicles	\$3,000	\$8,000	\$16,500	\$50,350
R retirement plan	\$2,000	\$8,000	\$20,000	\$61,500
Spouse retirement plan	\$2,000	\$8,000	\$20,000	\$70,000

Table 3: 2004 Median Values Used to Impute Unfolding Brackets for Retirement Items.

<i>Asset item</i>	<i>Low</i>	<i>High</i>
IRA	\$5,000	\$40,000
Roth IRA	\$5,000	\$26,500
Coverdell IRA	\$4,400	\$20,000
Keogh plan	\$5,000	\$25,000
Variable annuities	\$5,000	\$32,000
529 plans	\$3,500	\$34,500
Other tax advantaged plans	\$6,000	\$50,000

*Step 3 - Bracketing Interpolation of Items:* The next step we did to was to impute missing item values using a simple algorithm that takes advantage of the longitudinal aspect of the NLSY79 data. We linearly interpolated any missing value that had a set of bracketing values available, by which we mean known values from any “before” interview and any “after” interview. A “missing value” refers to any situation where the respondents reports holding a particular asset/debt, but does not report its value (directly or via unfolding brackets).

There are two bracketing cases. The first is when bracketing values are available for cash-related asset/debt categories (cash savings, stocks/bonds, trusts, other debt, IRAs, 401ks, CDs). In this case we considered as a valid bracketing value any instance when the respondent reports holding this asset/debt and gives a value, or any instance where the respondent reports not having this asset/debt in which case we assign a value of zero.

The second bracketing case is for property-related asset/debt categories (home value, mortgage, property debt, business assets, business debt, car value, car debt, possessions). Unlike in the first case we only used as a valid bracketing value any instance when the respondent reports holding this asset/debt and gives a value. If the respondent reports *not* having this asset/debt we did *not* assign a value of zero because this would amount to assuming a house (for example) had no value before it entered the respondent's possession.

If the missing value was not centered between two known values, the imputed value is linearly interpolated between the two. This algorithm mirrors the procedure used in the Netherlands Socio-Economic Panel for their asset and debt data.

*Step 4 - Linear Extrapolation of Items:* The primary drawback to the above bracketing interpolation is that it provides no method of estimating an item's value if the item is either the first or last in a series. For example, if a respondent provides information on his car's value in 1985, 1986, and 1987, states he does not know its value in 1988 and then drops out of the survey, there is no bracketing value for 1988.

To estimate these missing starting and ending points we fit the known data using ordinary least squares (OLS) and then extrapolated to determine the missing value. The (respondent-specific) regression we estimated was:

$$Item\ Value_{it} = a_i + b_i Year_{it} + u_{it}$$

using non-missing values for this asset/debt item for respondent *i*. As an example, assume a respondent stated he owned a vehicle in 1985 but did not know its value. Then assume this respondent in 1986, 1987 and 1988 said his vehicle was worth \$14,000, \$10,000 and \$8,000 respectively. The OLS imputation regression would be run with the following values

<i>Y-Value</i>	<i>X-Value</i>
14,000	1986
10,000	1987
8,000	1988

The resulting computation for the missing year (1985) is:

$$Item\ Value = 5,971,666.5 - (3,000 * 1985) = \$16,666.50,$$

so we used \$16,666.50 as the imputed value for 1985. Because the NLSY79 data does not contain any fractional data, all cents values were rounded.

We imposed two types of restrictions on the data used for each respondent-specific regression. First, we require that two or more non-missing values were available. A non-missing value for cash-related asset/debt categories (cash savings, stocks/bonds, trusts, other debt, IRAs, 401ks, CDs) is any value reported by the respondent or a zero value if the respondent states they do not have this asset/debt. For property-related asset/debt categories (home value, mortgage, property debt, business assets, business debt, car value, car debt, possessions) only values reported by the respondent are used. Users should note that this mimics the two types of bracketing strategies

described in step 3. Any values imputed from steps 2-3 are treated as non-missing values and used in the regression.

Second, to run the regression the respondent must also have reported an item value in the next closest wealth interview. For example, if the respondent did not know the value of his vehicle in 2004, we *must* have a known (reported or imputed) vehicle value in 2000 (the closest year, given that asset data were not collected in 2002) for the imputation to occur.

We also imposed additional restrictions on the predicted values that arise from these respondent-specific regressions. If the predicted value arising from the regression is negative, we did not use it as an imputation because respondents cannot report a negative asset or debt. In addition, if the predicted value is more than twice as large as the item value reported in the nearest year, it was not be used as an imputation. These rules are designed to ensure our extrapolated values are conservative estimates.

User Note: It is important to understand that the revised variables will potentially change with the addition of each subsequent round of wealth data. Revisions will occur because additional data will sometimes allow us to impute a missing value via step 3 (bracketing interpolation) rather than step 4 (linear extrapolation). This situation is similar to the NLSY79 work history data, which sometimes change when new information becomes available.

#### *Step 5 - Creating Total Net Worth Variables:*

The new data also include a “created net worth” variable for each survey year in which an asset module was fielded. This series was computed simply by combining the revised asset and debt series using the following equation for each respondent.

$$\begin{aligned} \text{NET WORTH} = & \text{HOME VALUE} - \text{MORTGAGE} - \text{PROPERTY DEBT} + \\ & \text{CASH SAVING} + \text{STOCKS/BONDS} + \text{TRUSTS} + \text{BUSINESS ASSETS} - \\ & \text{BUSINESS DEBT} + \text{CAR VALUE} - \text{CAR DEBT} + \text{POSSESSIONS} - \text{OTHER} \\ & \text{DEBT} + \text{IRAs} + \text{401Ks} + \text{CDs}. \end{aligned}$$

If any of the revised items are missing because they could not be imputed, the computed net worth variable is set to missing. Note that each respondent is asked about 15 types of assets and debts in each round. There might be some types of assets/debts that the respondent reports not holding, some where he gives a value, some where we impute a value, and some where we are unable to impute a value. If any asset/debt falls into the latter category, we do not compute the total net worth variable for that particular respondent-year case. While we do not compute a total net worth in these cases, the revised series are designed to let researchers do it easily, in part because all respondents who do not own an asset or debt have a zero in the revised series, instead of a -4.

*Step 6 – Revision of Top Codes:* The last step calculated new and consistent top codes for the wealth data.

The NLSY79 has used three basic types of top coding algorithms for financial data. In the early years of the survey (up to 1988), every answer to NLSY79 questions that resulted in a response above a specified cutoff value, such as \$100,000 for some variables, is recoded to the truncation value plus one dollar, such as \$100,001. Unfortunately this algorithm results in a sharp downward bias in the sample mean because the right tail of the distribution is truncated. In the middle years (1989 to 1994) a new algorithm was implemented, replacing all values above the hard cutoff with the average of all outlying values. Starting with the 1996 data, a third approach was used. In this approach the hard cutoff was eliminated and the cutoff became the value which would shield the top two percent of respondents. All values in the top two percent were averaged and that averaged value replaced values above the top code.

Because the NLSY79 has used a variety of different methods, because a number of researchers have complained about the lack of information above a hard cut off and because the data cleaning steps dramatically changed a number of the highest values, we retopcoded home and vehicles values because homes and vehicles are clearly identifiable objects which can reidentify respondents. Other asset or debt categories are no longer topcoded because it is difficult to use them to identify a particular respondent.

If the variable was previously top coded we re-top coded it using the top 2% described above. When calculating the top 2% we did not include individuals whose values were set to zero because they did not own the item or have the debt.