
Chapter 3: Guide to the NLSY97 Data

This chapter provides some practical information about how NLS variables are collected, created, and arranged on the CD-ROM. An explanation of the cohort's hard copy and electronic documentation is also included. The first section describes the different survey instruments used to collect the raw NLSY97 data. This section also explains how question names are assigned. Next, the guide discusses the primary types of NLS variables and the process by which each is assigned a reference number and title that serve to identify it throughout the NLS documentation. The third section reviews the codebook—that is, the information about each variable contained on the CD-ROM—and the accompanying paper documentation. This discussion will help users understand how to interpret the various pieces of information presented in the NLS documentation system. Finally, this chapter gives researchers some basic instruction in using the search functions on the CD-ROM software to find variables of interest.

3.1 Survey Instruments & Other Documentation

The primary variables found within the main data set are derived directly from one or more survey instruments. This section explains the conventions used in the NLSY97 documentation to identify questionnaire items from some of the primary survey instruments.

NLSY97 Survey Instruments

The term “survey instrument” is used to refer to the NLSY97 questionnaires that serve as the primary source of information on a given respondent. In round 1, there were separate and distinctly different questionnaires for the household informant (the *Screener, Household Roster, and Nonresident Roster Questionnaire*), the NLSY97 respondent (the *Youth Questionnaire*), and the responding parent (the *Parent Questionnaire*). In rounds 2–4, the *Youth Questionnaire* and the *Household Income Update* were used as survey instruments. Each questionnaire is organized around a set of topical subjects, the titles of which usually appear on either the first page of each section of the questionnaire or as a header. The various survey instruments are described in detail in section 1.4, “Content of the NLSY97.”

User Notes: The questionnaires are critical elements of the NLSY97 documentation system and should be used by researchers to determine the wording of questions, response categories, and the universe of respondents asked a given question.

For each round, NLSY97 questionnaires record (1) interview dates; (2) responses to the topical survey questions; (3) locating information which will assist NORC in finding the respondent for the next interview (not available to users); and (4) interviewer remarks on such topics as the race and gender of the respondent, language in which the interview was conducted, interviewer's impressions, etc. The show card, an interviewing aid used in conjunction with the questionnaire, lists the possible response categories for select questions and helps the respondent keep the more complicated response categories in mind.

Questionnaire Item or Question Name: This generic term identifies the source of data for a given variable. A questionnaire item may be a question, a check item, or an interviewer’s reference item appearing within one of the survey instruments. These items have question names that begin with an abbreviation of the section where each is located. Following the section abbreviation, the question name includes a combination of numbers and letters that identify it within the section. Many questions simply have numbers in numerical order. Some questions, as in the examples in the tables below, have a decimal extension that indicates the question is repeated or looped during the survey. For example, a question about hours worked would be repeated for each employer, with decimal extensions .01 through .09 indicating employers 1–9. Another common extension in question names is _D, _M, or _Y (or -D, -M, -Y), indicating that the variable reports the day, month, or year of a date. If a question is repeated in more than one round, it will have the same question name in each round so that users can easily locate identical questions in the data set across survey years.

3.1 Table 1. Sample Question Names by Youth Questionnaire Section

Section	Rounds 1 and 4 Question Names	Rounds 2 and 3 Question Names
<i>Youth Questionnaire</i>		
Information	YINF-2560	—
Household Information	—	YHHI-50510.04, YHHI-4100.07~M
CPS	YCPS-14400	—
Schooling	YSCH-22800.01, YSCH-26500	YSCH-2857B, YSCH-33900.01
Peers/Opportunity Sets	YPRS-800 (round 1 only)	—
Time Use	YTIM-2200 (round 1 only)	YTIM-300
Employment	YEMP-1800.02, YEMP-103500	YEMP-200A, YEMP-38313.02
Training	YTRN-800, YTRN-9200.01	YTRN-710, YTRN-7725.02
Health	YHEA-1600	YHEA-2050
Self-Administered	YSAQ-006A, YSAQ-447.03	YSAQ-394, YSAQ-503.02
Marriage	YMAR-2100, YMAR-12700.01	YMAR-729E, YMAR-3050.01
Fertility	YFER-700.04, YFER-14600	YFER-7800, YFER-12100A.01
Program Participation	YPRG-1700, YPRG-13500.01_M	YPRG-12700A, YPRG-19400.03~Y
Income / Assets	YINC-2300, YINC-21400.01	YINC-8900, YAST-2696
Expectations	YEXP-900	—
PIAT Math	YPIA-100	YPIA-100
Locator	YLOC-1500	YLOC-350
Interviewer Remarks	YIR-1500	YIR-1740.01
<i>Household Income Update</i>	HIU-5 (round 4 only)	HIU-5

3.1 Table 2. Sample Question Names in Screener and Parent Questionnaires

Section	Round 1 Question Names
<i>Screener, Household Roster, and Nonresident Roster Questionnaire</i>	
Screener	SE-9, SE-31B.01
Household Roster	SH-1B, SH-103.05
Nonresident Roster	SN-225.04, SN-337A.02
<i>Parent Questionnaire</i>	
Information	PINF-015_D, PINF-297.01
Family Background	P2-029, P2-108B.01
Calendars	P3-051.01_M, P3-137
Parent Health	P4-027
Income and Assets	P5-073.02, P5-136
Self-Administered	P6-021B, P6-036
Child Calendar	PC8-009_Y, PC8-025, PC8-086.01
Child Health	PC9-014, PC9-039.04
Child Income	PC10-025
Expectations	PC11-013
Family	PC12-010, PC12-012A
Parent Locator	PLOC-018
Parent Interviewer Remarks	PIR-007, PIR-009K

User Notes: Users should be aware that, while the source of the majority of variables in the main data sets is the questionnaire, certain variables are created either from other NLSY97 variables or from information found in an external data source (see “Types of Variables” below).

3.2 Types of Variables

There are five types of variables present in the NLSY97 data. The type of variable affects the title or variable description of each variable and the physical placement of the variable within the codebook.

Types of variables include:

- (1) Direct (or raw) responses from a questionnaire or other survey instrument.
- (2) Symbols and roster items, which are used to guide the interview.
- (3) Constructed variables based on responses to more than one data item. These items are edited for consistency where necessary.
- (4) Constructed variables from data provided on a non-NLS data set.
- (5) Variables provided by NORC or an outside organization.

User Notes: Users should note that survey personnel do not, in general, impute missing values or perform internal consistency checks across waves. Exceptions will be noted.

Variable Descriptions or Variable Titles

Each variable within NLSY97 main file data sets has been assigned an 80-character summary title that serves as the descriptive representation of that variable throughout the hard copy and electronic documentation system. Variable titles are assigned by CHRR archivists who endeavor, within the limitations described below, to capture the core content of the variable and to incorporate universe identifiers that specify the subset of respondents for which each variable is relevant. Some titles indicate the reference periods (e.g., survey year or calendar year) of the variables as well.

Universe Identifiers: If two ostensibly identical variables differ only in their respondent universes, the variable title will include a reference to the applicable universe. The appropriate universe will either be appended in parentheses or identified before the variable title.

Example 1: R00029. “R Do Any Work for Pay Last Week? (R Does Not Own Bus/Farm)”
R00030. “R Do Any Work for Pay or Profit Last Week? (R Owns Bus/Farm)”

Example 2: R01075. “Compensation Received (Start <16) EMP 01”
R01803. “Compensation Received (Start 16+) EMP 01”

User Notes: Users should not presume that two variables with the same or similar titles necessarily have the same (1) universe of respondents or (2) coding categories or (3) time reference period. While the universe identifier conventions discussed above have been utilized, users are urged to consult the questionnaires for skip patterns and exact time periods for a given variable and to factor in the relevant fielding period(s) for the cohort. In addition, variables with similar content may have completely different titles, depending on the type of variable (raw versus created).

Symbols and Roster Items

There are two main types of variables not necessarily represented by a single item in the questionnaire: symbols and roster items. These items are used by the CAPI system during the interview to organize, display, and store information collected during the interview; to determine which question paths the respondent should follow; and to fill in respondent-specific text in various questions. For example, rather than asking about a respondent’s “current employer,” the CAPI software fills in the actual employer name reported earlier in the interview. Many of these symbols and roster items are provided in the data set for user reference; researchers should be aware of the differences between the two types and the uses of each.

Symbols

The NLSY97 CAPI software generates symbols, which are items containing real-time information provided by the respondent during the survey. Symbols can be used to store data derived from one or more questions. For example, if the youth corrects information from the screener about his or her birth date during the administration of the *Youth Questionnaire*, the corrected information replaces the original data in the symbol item. Symbols are used throughout the questionnaire to determine whether certain groups of questions should be asked. For example, the symbol that states whether the youth is independent (Y12!INDEPEN) is later used to determine whether the youth is asked certain income and asset questions.

All symbol variables have “Symbols” as their primary area of interest. In general, question names for round 1 symbol variables begin with “KEY!”; symbols in rounds 2–4 generally have “SYMBOL!” to start their question names.

Rosters

The NLSY97 uses rosters in various sections in which information is collected on a number of persons, schools, or employers. Rosters are an important part of the NLSY97 data set. These grids of information help researchers to analyze data in an efficient and accurate way. However, the structure and use of rosters may be somewhat confusing, so it is vital that researchers understand how they are constructed.

User Notes: In addition to the detailed discussion in the following paragraphs, the introduction to section 4.3, “Employment,” contains an example that illustrates how to use the employer roster in research. Although that example pertains specifically to employers, the basic concepts apply to other NLSY97 rosters. Researchers using any roster data may find the example helpful.

What is a roster? A roster may be thought of as a list—for example, a list of household members, a list of employers, or a list of children. A respondent with two children will have data on the first two lines of the child list, or child roster. A respondent with four employers will have information on the first four lines of the employer roster. In addition to the name of the person or item (which is not released to the public), the roster contains other basic information, such as the age, race, and labor force status of household members or the start date and stop date for each employer.

In the paper-and-pencil interviews (PAPI) of older NLS cohorts, the questionnaires included a chart or grid listing this type of information, like the one shown in Figure 1. For example, in the household roster grid, each household member’s name was entered in a separate row. The interviewer asked the respondent for each member’s date of birth, enrollment status, employment status, etc., filling in the answers in the

appropriate column. This completed household roster contained all the pertinent information about household residents, and researchers could easily use the variables based on this roster to examine characteristics of household members.

3.2 Figure 1. Sample PAPI Roster Grid

What are the names of all family members who are living in your home?						
Name	What is __'s relationship to you?	How old is __ today?	(Age 4 and older) Is __ enrolled in school?	(Age 16 and older)		
				How many weeks did __ work in the last 12 months?	How many hours did __ usually work per week?	What kind of work was __ doing in the past 12 months?
Susan	Mother	45	No	50	25	Graphic design
John	Father	49	No	50	40	Banking
Jimmy	Brother	17	Yes	35	15	Food service
Sally	Sister	12	Yes	(n/a)	(n/a)	(n/a)
Robert	Brother	3	(n/a)	(n/a)	(n/a)	(n/a)
Jan	Grandmother	77	No	0	(n/a)	(n/a)

When the NLS surveys changed to computer-assisted personal interviewing (CAPI), rosters became a very important way of organizing information during the interview. Instead of using an actual grid, however, CAPI questionnaires include a series of questions that gather the same types of information that would have been included in the grid in a paper-and-pencil interview. The computer then moves the answers to these questions into a grid, creating a roster from the information.

After the roster is created, it can be used to guide subsequent portions of the interview. For example, during the interview the NLSY97 questionnaire gathers the names, dates of attendance, and level of school (secondary school or college) for each of the respondent's schools and organizes them into a roster. The rest of the school section then asks questions about the first school on the roster, followed by questions about the second school, then the third, and so on. The information about the level of the school determines whether the respondent is asked questions that apply to high school or college.

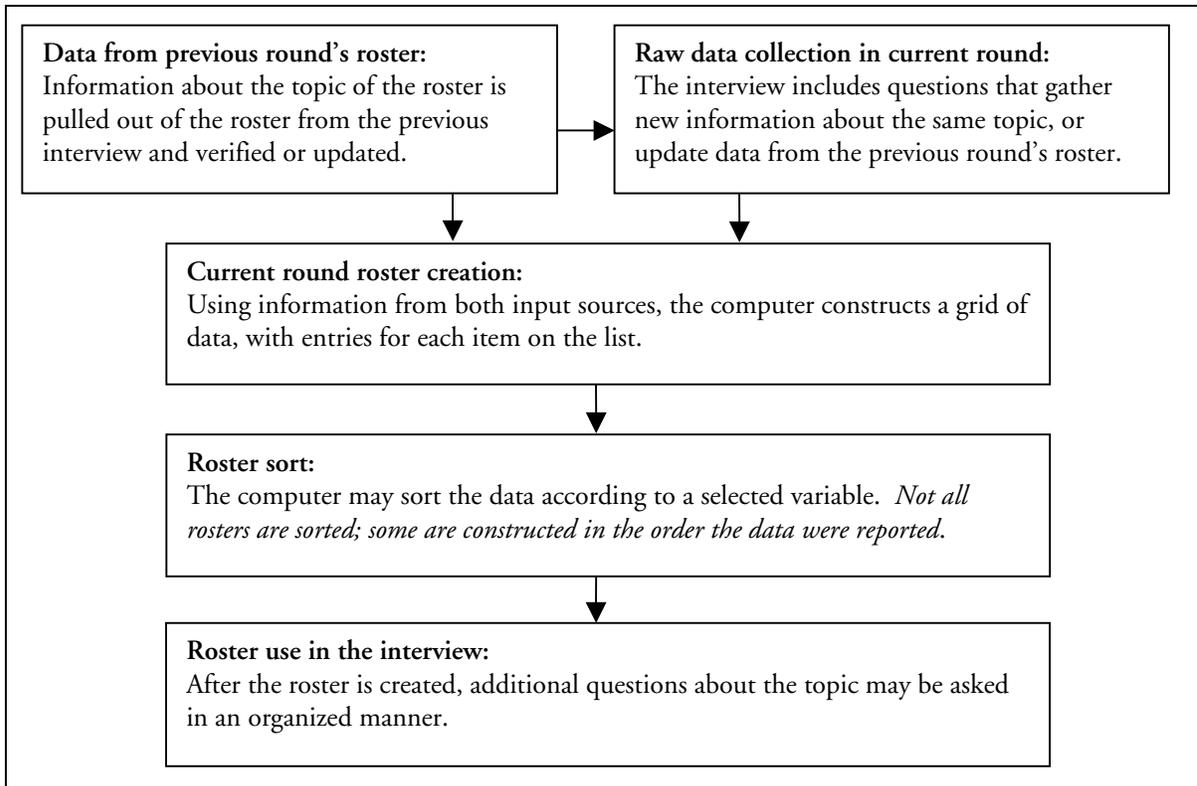
The information from the roster is also presented on the CD as an organized list of data, so that these variables are easy for researchers to access. To the user, the school roster appears as a consolidated block of variables that contains key information such as dates of enrollment, an identification number for the school, and variables indicating the type (private or public) and level (junior high, high school, college) of the school. For example, the variables in the round 2 school roster are listed in Figure 2, along with their reference numbers. Thus, rosters are a way of organizing information both for researchers and for the actual interview so that questions are asked in a logical manner.

3.2 Figure 2. Example: Variables in CAPI-Generated School Roster for Round 2

Question Name	Variable Title	Reference Numbers (one for each school)
NEWSCHOOL_PERIODS.xx	Number of Times R Enrolled in School xx	R24605.–R24610.
NEWSCHOOL_START1.xx	Month/Year R Start 1st Enrollment in School xx	R24611.00–R24616.01
NEWSCHOOL_START2.xx	Month/Year R Start 2nd Enrollment in School xx	R24617.00–R24620.01
NEWSCHOOL_START3.xx	Month/Year R Start 3rd Enrollment in School xx	R24621.00–R24621.01
NEWSCHOOL_STOP1.xx	Month/Year R End 1st Enrollment in School xx	R24622.00–R24627.01
NEWSCHOOL_STOP2.xx	Month/Year R End 2nd Enrollment in School xx	R24628.00–R24631.01
NEWSCHOOL_STOP3.xx	Month/Year R End 3rd Enrollment in School xx	R24632.00–R24632.01
NEWSCHOOL_SCHCODE.xx	School Code Elementary, Middle, High, College	R24633.–R24638.
NEWSCHOOL_INTERVIEW.xx	Which Survey Round School xx Reported in	R24639.–R24644.
NEWSCHOOL_TYPE.xx	Type of School xx R has Attended	R24645.–R24650.
NEWSCHOOL_PUBID.xx	PUBID of School xx R has Attended	R24651.–R24656.

How are rosters created during the interview? This section outlines the process used during the interview to create a roster. Rosters may include data from both previous interviews and the current interview. After the roster is created and sorted, it can be used to guide the rest of the interview. Figure 3 provides a pictorial overview of the creation of a roster.

3.2 Figure 3. How Rosters Are Created



Data from previous interviews: As shown in the figure, creation of a roster for the current round often begins with information found in the roster from the previous round. The appropriate respondent-specific data are saved on the interviewer's laptop before he or she administers the survey. When the interview gets to a point where roster information is collected, the data from the previous round's roster are often used as the base for the current roster. The respondent verifies and updates the information. If no changes have occurred since the last interview—for example, if exactly the same people live in the respondent's household—then the current round's roster will be the same as the one from the previous round.

For example, the interviewer reads a list of all of the people on the household roster from the last interview. The respondent first states whether any of those people have moved out of the household and then reports new household members. If any members remain from the previous year, their information—date of birth, gender, race/ethnicity, etc.—is carried over from the previous interview, and any missing data are collected. This method is more efficient than asking the respondent to report all household members every year.

Raw data collection: After the respondent and interviewer review and update the roster from the previous round, the survey collects current information. For example, new people might have moved into the household, so the interviewer asks the respondent about their characteristics. At this point, the respondent is done answering questions that will fill up the data grid on a particular topic.

Roster creation and roster sort: Using the updated roster from the previous round and the new raw data just collected, the computer creates a new roster for the current round. For example, the employer roster contains the following information for each job: a unique identification number for the employer, employment dates, whether the job was current at the interview date, whether the job was in the military, and whether the job was an internship. If the respondent had held the job at the time of the previous interview, the start date and employer identification number are carried over from the old roster, and the other information is taken from the questions at the beginning of the employment section for the current year. Similarly, the household roster contains information from the previous interview about household members reported at that time and data from the current interview about new household members.

In some cases, the computer also sorts the roster and puts the items in order based on a specified variable. For example, in the round 1 household roster, all youths in the age range of the NLSY97 cohort were listed first, and then all other household members were listed from oldest to youngest. The employer roster is sorted by job end date so that the most recent jobs are listed first.

Roster use in the interview: Finally, the roster is used to determine the order in which the other questions about each topic are asked. In most cases, the survey collects far more information than is stored in the

actual roster, and the answers to these questions remain outside the roster as raw data. So that the interview makes sense to the respondent, these additional questions are asked about the people or things on the roster in the order that the people or things are listed.

For example, the respondent first answers questions about industry, occupation, rate of pay, etc., for the first employer listed on the roster. The same questions are then asked about the second job, then the third job, and so on. Similarly, the first set of questions about household members refers to the first person listed on the roster. When all of those questions have been answered, the same questions are asked about the second person, the third person, etc.

How should researchers use the roster data in analysis? The data set is organized so that rosters can easily be found and used in research. Because rosters present key pieces of information in a structured format, they are the best place to obtain that information. All variables found on rosters have “Roster Item” as their main area of interest. Each roster has a unique name that serves as the beginning of the question name for all variables on the roster; the same name appears at the beginning of the variable title for each item on the roster. Different rosters have been used in different rounds, depending on the topics included in the interview and the type of information collected. The roster names and question names are shown in Figure 4.

3.2 Figure 4. Rosters Included Each Round

Roster	Question name	Round 1	Round 2	Round 3	Round 4
Household Information	HHI2 (rd. 1), HHI (rds. 2–4)	✓	✓	✓	✓
Nonresident Roster	NONHHI	✓	✓	✓	✓
Youth Information	YOUTH	✓			
School Roster	NEWSCHOOL		✓	✓	✓
Employer Roster	YEMP	✓	✓	✓	✓
Freelance Jobs Roster	FREELANCE		✓	✓	✓
Training Roster	TRAINING			✓	✓
Biological Children Roster	BIOCHILD	✓	✓	✓	✓
Parent Household Information	PARHHI	✓			
Parent Youth Information	PARYOUTH	✓			

Data hint →

Researchers can locate rosters on the data CD-ROM by looking at the roster item area of interest, by selecting the appropriate question name, or by searching the any word in context index for variables with “ros” or “roster” and the name of the roster of interest in the title.

User Notes: When the NLSY97 data set was initially created, variables could only be assigned to one area of interest. The newer data extraction software permits variables to be linked to multiple areas of interest. However, additional areas have not been assigned to every variable. Because roster variables were initially located in the roster item area of interest on the CD, they may not be grouped with the rest of the data on a particular topic. For example, the school roster variables will not appear if the user searches for the “School Experience” area of interest. For this reason, it is very important that researchers become familiar with the rosters used in the data set. If a roster is available on the topic of a particular research project, users should always locate that roster using one of the search techniques mentioned above and examine it before using the other variables that relate to their research.

Using rosters in single-round analyses: When looking at the data set, users will notice that many questions are repeated for each person or thing on the roster, and the titles for these repeated questions include a number. This number indicates the line on the roster that corresponds to the person or item being described in that variable. For example, the question “Self-Employed Business/Industry Job 02” indicates the industry of the second job listed on the respondent’s self-employment roster. The researcher may then want to examine information such as the respondent’s start and stop dates or rate of pay for that job. To find this information, he or she can then look at the data for those items contained in the roster for job #02, or the self-employment job that is on the second line of the roster. For all other questions asked after the roster was created in that same survey year, job #02 will refer to the same self-employment job.

Users should be aware that, in some cases, the information contained in the rosters actually appears in the data set more than once. As Figure 1 suggests, data may first be included at the point in the interview when the information was actually collected. For example, the round 1 screener question SE-28 asked the household informant for the date of birth of each household member. After all the raw data had been gathered, the computer sorted all the answers and created the household roster. At this point the date of birth information is also located in the round 1 roster variables named HHI2_DOB. In the case of the round 1 household roster, both the raw data and roster items are included in the data set.

In other cases, the raw answers may be blanked out of the public use data set. If a reference number is not listed for a given question in the questionnaire, then that raw data item may only be represented in roster form. For example, answers to the raw data questions used to create the employer roster are blanked out and do not appear on the CD. In the printed questionnaire, these questions have no reference numbers. However, all of the data collected in these questions (except for confidential information like the name of the employer) appears in the employer roster.

Data hint ➔

Even though the data may appear more than once, **survey staff strongly recommend that researchers use the roster information rather than the raw data whenever possible.** Survey staff are working to eliminate these duplicate sources of information. For example, screener question SE-28 is one of the variables that has already been removed.

For some variables, the roster information may be more accurate because some rosters are updated during the interview if the initial report was inaccurate. When survey staff prepare the data for release, they clean up the rosters if necessary but do not necessarily clean the corresponding raw data. Finally, because many rosters are sorted in a particular order, the number of a person or item on the roster will not match the number in the questions that precede roster creation. For example, in the household screener (the SE questions), person #01 is the first household resident mentioned to the interviewer. In the household roster and all later interview questions, person #01 is the oldest person in the household who was eligible for the NLSY97. Person #01 in the SE questions might be person #05 on the roster. It can be very difficult to determine to which person, school, or job a pre-sort question refers. For all of these reasons, roster data are always preferable to raw data in cases where both are available.

Using rosters from more than one round: Because the NLSY97 is a longitudinal survey, researchers often want to link data across survey rounds. However, household residents, jobs, and so on may move around on the roster in different interviews. That is, a father who was listed third on the roster in round 1 might move to position 2 or 4 in round 2. The unique identification numbers (UIDs) are the key to finding the same person or thing in different rounds. Most of the rosters contain variables assigning a unique number to each person or thing listed. This number never changes and can be used to link roster items across rounds. In some cases, it also makes it possible to link people between two different rosters in the same survey. For example, beginning in round 2 the unique ID listed for a child on the biological children roster is the same one assigned to that child on the household roster. Researchers can therefore examine data on both rosters about the same child.

An additional feature of most unique ID numbers is that they incorporate an indicator of the round in which the person or item was first reported. For example, IDs of roster items reported in round 1 may begin with “1” or “97,” while those first reported in round 2 begin with “2” or “98.” (Beginning with round 3, 4-digit years are used so that IDs begin with “1999” rather than just “99.”) UIDs for people on the household roster are constructed in a slightly different manner; researchers should refer to section 4.6.5, “Household Composition,” for more information.

Created Variables

Created variables generally start with “CV_” in the codebook, as in the ‘Hourly Rate of Pay’ example later in this chapter, with a few exceptions. One major exception is the sampling weight variables, which have question names SAMPLING_WEIGHT and CS_SAMPLING_WEIGHT. In addition, the family process variables constructed by Child Trends (see sections 4.5 and 4.6) have question names beginning with “FP_” in the codebook. In the Event History data, all variables are created and can be located in the “Event History” area of interest (see section 4.4 of this guide for more information and question names).

3.3 NLSY97 Documentation

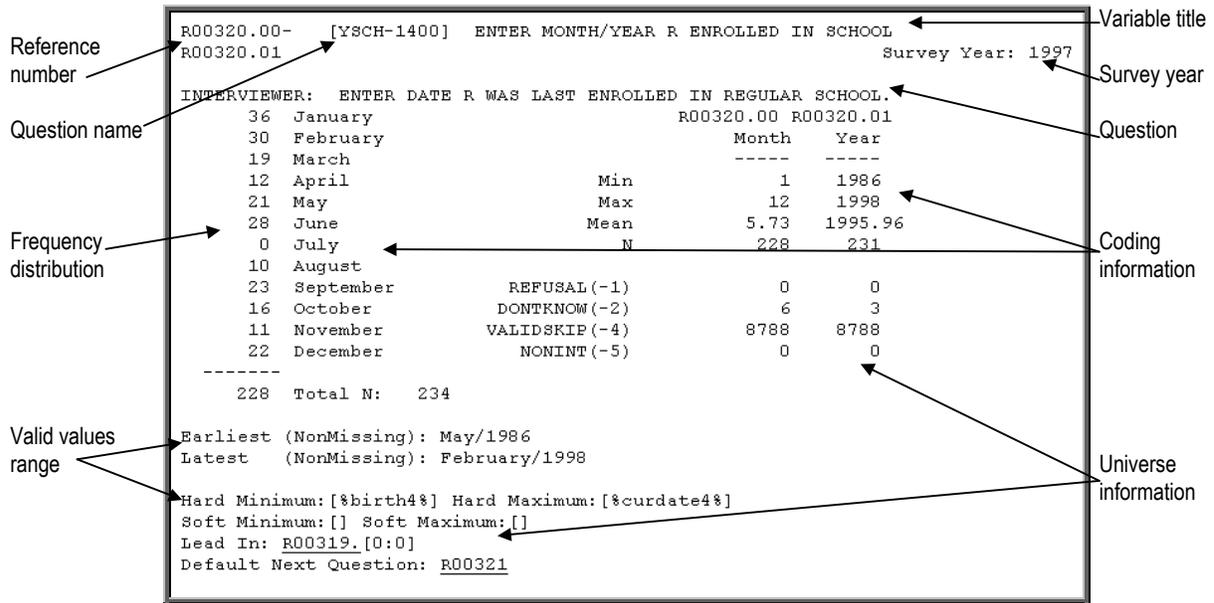
Variables present in the NLSY97 main file are documented via (1) a codebook; (2) accompanying supplemental documentation; and (3) error updates. This section describes the three primary components of the NLSY97 documentation and discusses the important types of information found within each.

Codebook

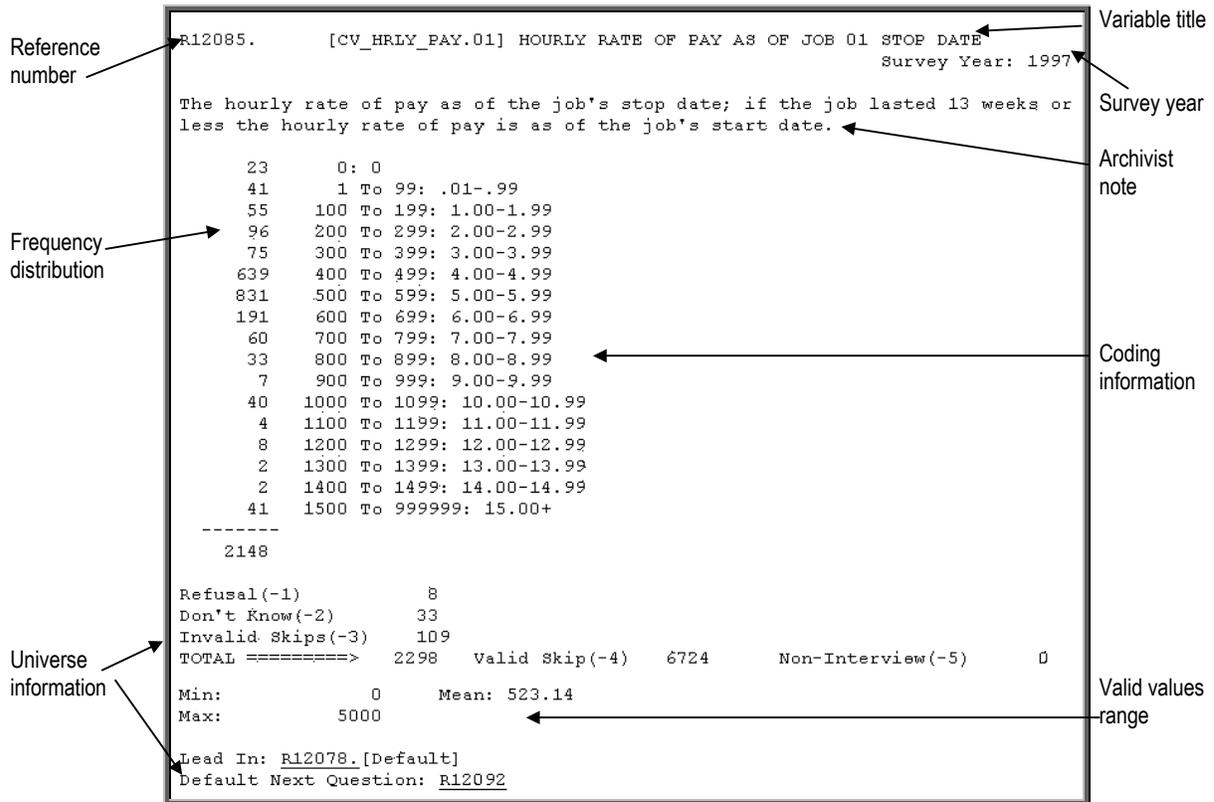
The codebook is the principal element of the NLSY97 documentation system and contains information intended to be complete and self-explanatory for each variable in a data file. The software on the NLSY97 CD-ROMs allows easy access to each variable’s codebook information and permits the user to print a codebook extract for selected variables.

Every variable is presented as a block of information called a “codeblock.” Sample codeblocks are shown in Figures 1 and 2. Codeblock entries depict the following important information: coding information, frequency distribution, questionnaire items, universe information, valid values range, and question text. Each of the above terms is described more completely in the following pages. Codeblocks for many variables also include special notes designed to assist in the accurate use of data.

3.3 Figure 1. NLSY97 Questionnaire Item Codeblock



3.3 Figure 2. NLSY97 Created Variable Codeblock



Coding Information: Each codeblock entry presents the set of legitimate codes that a variable may assume along with a text entry describing the codes. *Users should note that coding information for a given variable in the NLSY97 codeblock is not necessarily consistent with the codes found within the questionnaire. If the two sources are different, the codebook is current and the questionnaire information should not be used in analysis.* For example, an additional code may be added during data processing if a significant number of respondents gave the same answer to the “other—specify” option in an answer list.

The following types of code entries occur in NLSY97 codeblocks:

Dichotomous (or variables answered yes/no), uniformly coded “Yes” = 1 and “No” = 0. Other dichotomous variables have frequently been reformulated to permit this convention to be followed.

Discrete (Categorical), as in the case of ‘Month Enrolled in School.’

January	May	September
February	June	October
March	July	November
April	August	December

Continuous (Quantitative), as in the case of ‘Hourly Rate of Pay’ in the example above. These variables have continuous data but are presented in the codebook using a convenient frequency distribution. Note that rate of pay variables often have two implied decimal points.

Valid data are generally positive numbers. In a small number of variables, negative responses are possible; users should check the minimum values allowed for each question to clarify whether negative numbers are permissible. The following missing value conventions are used throughout the data:

Noninterview	-5
Valid Skip	-4
Invalid Skip	-3
Don’t Know	-2
Refusal	-1

Frequency Distribution: In the case of discrete (categorical) variables, frequency counts are normally shown in the first column to the left of the code categories. In the case of continuous (quantitative) variables, a distribution of the variable is presented using a convenient class interval. The format of these distributions varies.

Some codeblocks show frequency distributions for more than one variable. For example, the codeblock in Figure 1 above includes responses for both the month and the year that the respondent enrolled in school. Note that this codeblock also lists two reference numbers, R00320.00 and R00320.01. These two variables will be listed separately in the data extraction software, but if either is selected the user will see the codebook page shown above. However, if an extraction is performed, the variable for the month of enrollment will have the data shown in the month distribution, and the variable for the year of enrollment will contain the data shown in the year distribution. These combined codeblocks generally occur for date

variables and for variables that permit multiple responses to a single question (for example, see question YEMP-100300, which asks the respondent to identify all fringe benefits made available by an employer).

Questionnaire Item or Question Name: The question name provides the location of the question in the survey instrument or identifies it as a created variable. In the first example, the question name YSCH-1400 shows that the variable is based on a question in the schooling section of the youth instrument. In the second example, the question name CV_HRLY_PAY.01 indicates that the variable is created. For more information on how question names are assigned, refer to section 3.1, “Survey Instruments & Other Documentation.”

Universe Information: The universe information found in the codebook includes:

(1) **Universe Totals:** Two totals are presented: (1) The sum of the frequency counts for each coding category is located below the individual codes. (2) The sum of the valid responses plus missing response counts of “refusals,” “don’t knows,” and “invalid skips” can be found in the TOTAL=====> field. The number of respondents who were not asked a question because it did not apply to them—that is, “valid skips (-4)” —is also depicted.

(2) **Universe Skip Patterns:** The following detailed universe information will enable users to trace the flow of respondents both backward and forward through the CAPI questionnaire:

“Go to # XXXXX,” appended to certain coding categories, indicates that respondents selecting that answer category were routed to the next question specified.

“Lead In(s) # XXXXX” identifies the question or questions immediately preceding the codeblock question through which the universe of respondents was routed. Each lead-in number is followed by the relevant response value indicators, e.g., (Default), (ALL), [1:1], [1:6], etc.

“Default Next Question” specifies the next question that all respondents to the current question will be asked unless some skip condition indicates otherwise.

Valid Values Range: Depicted below the frequency distribution is information relating to the range of valid values for that particular distribution. “MINIMUM” indicates the smallest recorded value exclusive of skips, refusals, and don’t knows. “MAXIMUM” indicates the largest recorded value. As described below, the computer-assisted interview contains internal range checks that limit responses to those between predesignated values, warn interviewers to verify non-normative values, and bolster the information provided by the traditional minimum and maximum fields.

Maximum and Minimum Fields: The MIN and MAX fields define the range of responses, i.e., the minimum and the maximum values, for a data item. The MAX of 5000 (\$50.00) in the ‘Hourly Rate of Pay’ question means that it was the highest value recorded.

Hardmax and Hardmin Fields: Hardmax and Hardmin fields denote the highest and lowest values that were accepted. Dates, e.g., month/day/year of the respondent’s birth (%birth4%) and current interview

(%curdate4%), are often used as Hardmin and Hardmax values in order to restrict responses to certain questions to values within that range, as in the ‘Enter Month/Year R Last Enrolled in School’ example. Responses outside this range must be entered by the interviewer in the comment field; valid numbers are included in the data.

Softmax and Softmin Fields: Softmax and Softmin fields cover ranges where an answer may exceed normal limits yet remain within absolute limits; such answers are accepted after verification. A Softmax set to \$80,000 on an income question will trigger an alert to interviewers that a higher value is unusual.

Income Values: Confidentiality issues restrict release of all income and asset values. To insure respondent confidentiality, the top 2 percent of reported values for many income or asset variables are all converted to one set value. This “topcoded” value is calculated separately for each variable by averaging all the values which exceed the limit for that variable. Calculating topcode values in this way allows statistics such as means to accurately reflect the status of the population under examination without violating respondent privacy.

Verbatim: When an NLSY97 variable is taken directly from the questionnaire, the verbatim of the question or the instructions to interviewers appear beneath the variable title. If a single question is the source for more than one variable, the first variable contains the verbatim, while subsequent variables prompt the user to refer to the variable containing the verbatim.

Archivist information, notes, etc.: Some variables include additional information for users regarding inconsistencies in the data, methods of variable derivation, references to supplemental documentation, and so on. These notes generally appear beneath the variable title or question verbatim.

Supplemental Documentation

Purchasers of the NLSY97 data set must have access to all relevant documentation. Documentation for the NLSY97 includes the following items.

Technical Sampling Report—Youth Survey: This technical manual published by NORC describes the procedures used to select the youth sample. The manual includes weights and standard errors for the initial survey year.

Interviewer Reference Manuals: Accompanying each NLSY97 questionnaire will be an *Interviewer Reference Manual*. In a CAPI survey, interviewers have ready access to general and specific instructions that guide them in the administration of the electronic questionnaire. These “help screens” are physically linked to the appropriate questions throughout the instrument and can be accessed electronically. The

Interviewer Reference Manual reproduces the help screens so that researchers can view the various definitions and other pieces of information used during the interview.

Codebook Supplements: Variable creation procedures and supplemental coding information are provided within the *Codebook Supplement*. This information is **not** available in the hard copy NLSY97 codebooks. The attachments and appendixes in the following list can be found in the *NLSY97 Codebook Supplement*.

Attachment:

1. **1990 Census Industrial and Occupational Classification Codes.** This document lists the 3-digit 1990 Census codes used to classify job and training information (Census Bureau, 1990 Census of Population Alphabetical Index of Industries and Occupations, Washington, DC: U.S. Government Printing Office, 1991).

Appendixes:

1. **Education Variable Creation.** This document provides the programs for several created variables related to education. These include, among others, enrollment status, type of school, date received diploma, highest grade completed, number of schools attended, and math *PIAT* score.
2. **Employment Variable Creation.** This appendix provides programs for created employment variables, including hourly rate of pay, hourly monetary compensation, number of weeks worked, total tenure at job, and number of jobs held.
3. **Family Background Variable Creation.** This appendix of created variable programs contains those dealing with family background, such as household size, marital status, fertility and child status, marriage and cohabitation history, and citizenship status.
4. **Geographic Variable Creation.** Several variables in the main data set provide information about the respondent's area of residence, permitting researchers to identify key characteristics of the area without needing access to the Geocode CD-ROM. Included in this appendix is a summary of the four Census geographic regions, an explanation of the MSA/central city status variable, and the definition for the rural vs. urban variable.
5. **Income and Assets Variable Creation.** This document provides the creation procedures for income and assets created variables. These include household net worth and gross household income, as well as receipt of public assistance.
6. **Event History Creation and Documentation.** This appendix explains the structure of the event history variables and describes the creation process.
7. **Continuous Month Scheme and Crosswalk.** This document explains the structure of the event history month-by-month and week-by-week status arrays and provides crosswalks from continuous month/week numbers to actual month and year dates.
8. **Instrument Rosters.** A number of rosters are used to organize information during various parts of the interview. This appendix identifies these rosters and shows how they were used in different parts of the survey. It also lists the variable names, titles, and reference numbers for the various instrument rosters used in each interview.
9. **Family Process and Adolescent Outcome Measures.** This document, which is provided separately from the *Codebook Supplement*, summarizes the creation procedures for the various scales and indexes created by Child Trends, Inc. The appendix also presents the results of Child Trends' statistical analyses of the scales, indexes, and a number of related attitude and behavior variables.

Geocode Codebook Supplements: Supplemental coding information specific to the Geocode CD-ROM is provided within the *Geocode Codebook Supplement*. Information provided within this document is **not** available in the hard copy NLSY97 codebooks.

Attachments:

100. **1990 Census Bureau State and County Codes.** This attachment provides coding information for the state and county variables on the NLSY97 Geocode CD-ROM. These variables use the current Federal Information Processing Standards (FIPS) codes.
101. **MSA Codes.** This document lists the Metropolitan Statistical Area (MSA) coding scheme used for NLSY97 geocode variables. It also presents Consolidated Metropolitan Statistical Area (CMSA) codes, New England Consolidated Metropolitan Area (NECMA) codes, and Primary Metropolitan Statistical Area (PMSA) codes.
102. **IPEDS Data and College Identification Codes.** This attachment explains the Integrated Postsecondary Education Data System (IPEDS), and how this and other codes are used to identify the colleges reported by NLSY97 respondents.

Error Updates

Prior to working with an NLSY97 data file, users should make every effort to acquire current information on data or documentation errors. A variety of methods are used to notify users of errors in the data files or documentation and to provide corrected information for those persons who acquired an NLSY97 data set from the Center for Human Resource Research. The most up-to-date list of errors is found on the internet by linking to each cohort's page through the <www.bls.gov/nls> web site. Errors discovered after the release of a data file are distributed in hard copy form to current data purchasers along with the data set. Error notices and information on how to acquire the corrected data or documentation also appear in *NLS News*, the quarterly NLS newsletter, available online at <www.bls.gov/nls/nlsnews.htm>.

3.4 CD-ROM Search Functions

NLSY97 variables can be accessed via areas of interest or through a search of variable titles for any word. Both search functions provide users with bridging information to the codebook and survey instruments.

Areas of interest. NLSY97 data files are organized so that variables sharing a common factor are stored in unique groupings called "areas of interest." Users can browse through a given area and examine the variables associated with that topic. NLSY97 areas of interest are listed in Appendix A of this guide.

Any word search. All words, numbers, and symbols found in any variable title form an index on the CD-ROM. The "Any Word in Context" search function in the CD software allows the user to search this index and select NLSY97 variables whose titles contain any single word or combination of words.

User Notes: Any word in context searches for NLSY97 variables are limited by the choice of variable titles. Flexibility in variable title assignment for raw data items is restricted by the wording of the question as it appears in the survey instrument and the maximum allowable length for variable titles.

