
**Introduction to the CAPI Questionnaire
and Codebook**

The newest survey in the NLS program, the National Longitudinal Survey of Youth 1997 (NLSY97) is designed to be representative of the U.S. population born during the years 1980 through 1984. Through the NLSY97, the Bureau of Labor Statistics (BLS) will be able to identify characteristics that define the transition today's youths make from school to the labor market and into adulthood. The NLSY97 cohort includes 8,984 respondents ages 12–16 as of December 31, 1996; the sample contains a cross-sectional subsample and an oversample of black and Hispanic respondents. More information on the selection of respondents for the cohort is available in the *NLSY97 User's Guide*.

This survey was conducted as a computer-assisted personal interview (CAPI). The Round 1 survey was conducted using three different questionnaires: the *Screener, Household Roster, and Non-Resident Roster Questionnaire*; the *Youth Questionnaire*; and the *Parent Questionnaire*. The *Screener, Household Roster, and Non-Resident Roster Questionnaire* was administered to a household resident over the age of 18. The *Youth Questionnaire* was administered to the youth respondent living in the household. The *Parent Questionnaire* was administered to a parent or parent-like figure of the youth residing in the household. For information about how the responding parent was selected, researchers should refer to the *NLSY97 User's Guide*.

The Round 2 survey used a *Youth Questionnaire* similar to Round 1. The interview also included a new instrument, a brief *Household Income Update* administered to one of the respondent's parents. No screener or parent interviews were conducted. The content of these survey instruments is presented in separate questionnaire documents, available from NLS User Services. More information about the administration of the various instruments is provided in the *NLSY97 User's Guide*.

Frequencies from the data collected during the NLSY97 interviews are provided for researchers in the form of a codebook contained on the NLSY97 CD-ROM. For each variable taken directly from the interview, the codebook provides information about the question or check item, the variable reference number, the universe of respondents to whom the question applied, and the distribution of responses to the question. The codebook also contains a number of variables created by survey personnel after the interview, providing information about the content of the variable and the distribution of responses.

Some information relevant to the data contained in the codebook cannot be included on the CD-ROM due to practical constraints. This document, the *NLSY97 Codebook Supplement*, presents additional information which will help researchers to use the data more effectively. The first attachment lists industry and occupation codes and their associated descriptors; this list is too lengthy to be incorporated into the codebook. A number of appendices then provide programs for the created variables; researchers may want to use this information to identify the raw survey data used in a created variable and to understand the rules applied during the creation process. This document also describes the creation of the event history arrays and helps researchers to understand their structure and use.

This introduction to the *Codebook Supplement* is intended to assist researchers in understanding some of the terms and survey methods associated with a CAPI interview. The way in which a CAPI instrument is constructed has important implications for the presentation of the data, and so this discussion will aid in interpretation of the codebook and accompanying documentation.

I. TERMS

Discussions of CAPI surveys include a number of terms that may be unfamiliar to many researchers. The following terms are used throughout the survey documentation; we provide definitions here for easy reference.

Instrument Rosters

A “roster” is a list of one or more items of information pertaining to a specific set of subjects, such as the biological children of the respondent or members of the respondent’s household. For instance, the BIOCHILD roster contains a variety of items (e.g., name, gender, birthdate, etc.) pertaining to each biological child of the respondent. By using the roster format, the CAPI program can gather an inventory of information, tag the data to a specific subject, carry the data along through the interview, and access them when necessary. This format also allows for these items of information to be presented to the interviewer at any time. A listing of rosters, such as the Household Roster, is included in this *Codebook Supplement* as appendix 8. This listing includes the contents of the rosters, the applicable names attached to the rosters that might be encountered in the codebook and/or the questionnaire, and the variable reference number assigned to each piece of data in the codebook.

Interviewer’s Reference Manual

The *Interviewer’s Reference Manual* reproduces the electronic “help screens” that were available to the interviewers for specific questions. These help screens contain definitions, instructions, and other information which interviewers used during the interview to obtain consistent information from respondents, as well as listings of the questions for which they were used during the interview. The *Interviewer’s Reference Manual* should be used in conjunction with the codebook or the questionnaire so that researchers can fully understand the intent of each question in the survey.

Symbols and “Text Fills”

Symbols are reserved fields into which data can be placed, stored and accessed throughout the questionnaire. Each symbol field is assigned a distinct name, often with an index number appended when the same piece of information is stored for a set of subjects. For instance, the symbol “lintdate” contains the date of last interview for the respondent. The symbols “emp.name(1)” through “emp.name(7)” contain the names of the first through the seventh employers of the respondent. Throughout the survey, the last interview date and/or the name of the respondent’s first employer can be accessed by invoking the symbol names “lintdate” and “emp.name(1)” respectively.

Users will encounter these symbol names both in the codebook and in the questionnaire. Sometimes data in a symbol is accessed and used to govern a skip or perform a calculation. At other times, a symbol name may be part of a question text. In this case, the content of the specific symbol content becomes part of the text of the question, and is read as such. This is referred to as a “text substitution” or “text fill.” For instance, a question text reading “When you started working for [emp.name(1)], what kind of work did you do? That is, what was your occupation?” would appear to the interviewer with the appropriate name replacing the symbol “emp.name(1).”

Other types of text fills are also present in the questionnaire. For example, the responding parent may be asked the following question: “What was the reason [he/she] did not live with [his/her] [mother/father/parents]?” In this question, “he/she” and “his/her” refer to the parent’s spouse or partner. The computer automatically fills in the correct gender, and the interviewer reads the question appropriately. “Mother/father/parents” in the question text indicates a set of possible responses to the previous question. The computer automatically fills in the appropriate choice from the three words for the given interview. These automated text fills reduce error and remove the burden of asking the question using the appropriate phrasing from the interviewer.

Loops

Certain sequences of questions in the CAPI instrument are repeated a number of times. For instance, some sets of questions are repeated up to 7 times in the Employment Section of the youth questionnaire, for each of up to 7 jobs. Question names that include a “.01”, “.02”, “.03”, etc. at the end, belong to these repeating sequences of questions. Each repetition of the sequence of questions is referred to as a “loop.” Taking the Employment Section as an example, the sequence of questions asking about the first

job is referred to as the first “loop.” The sequence asking about the second employer is referred to as the second “loop,” and so on. While the codebook generally includes more than one loop in a series (because more than one may contain valid data), the questionnaire includes only the **first loop** in each set of loops.

Hard and Soft Range Restrictions

The “Hard Minimum,” “Hard Maximum,” “Soft Minimum,” and “Soft Maximum” specifications control the allowable range of values that can be entered for a given question. These fields are only active when a question calls for the entry of a time date or amount. Questions that require the interviewer to select one response or to select all that apply do not contain this field, as the range limits are implicit in the distribution code block and are thereby enforced.

Hard minima and maxima are absolute limits that an interviewer- or respondent-generated answer must obey. Entry of values outside the hard range is not allowed. In such cases the interviewer is instructed to enter the maximum or minimum allowable value, as appropriate, enter the actual response in the comment field, and “flag” the case for central office checking. Soft minima and maxima are nested within the hard range. When a response falls outside the soft range but inside the hard range, the computer beeps and asks the interviewer to either confirm or change the response.

In some cases, the hard ranges are themselves determined by a variable. For example, a question may use the respondent’s birth date as a minimum and the current interview date as a maximum, preventing the interviewer from entering a date earlier than the hard minimum birth date or later than the hard maximum interview date. This is a powerful tool for the collection of event histories (among other types of data sequences) and is used extensively in the instrument.

II. APPEARANCE AND PRESENTATION OF DATA

Some CAPI questions generate more than one variable. For example, some questions collect information about the date an event happened and generate three variables: month, day and year. Similarly, questions that ask the respondent to indicate which of several responses are appropriate, and to pick all responses that apply, can generate multiple variables. When a question record generates multiple variables, those variables have decimals in the reference numbers. These two types of questions (date-entry and “code-all-that-apply”) and the examples of the resulting variables, are discussed further below.

Date-Entry Questions

All date questions, whether full dates (month, day and year) or just month and year, are represented in the CAPI codebook with, respectively, three or two variables. Each variable contains the same codeblock displaying the ranges and missing values for all elements of the date. A base reference number ending in “.00”, is assigned to the first variable in the set. The same base reference number, with endings of “.01” and “.02” if necessary, is assigned to the other elements of the date. Thus, if the day is assigned a reference number of R10851.00 for the variable “Date of Birth of HH Member 01 (Scr Ros Item),” the month and year would be assigned reference numbers of R10851.01 and R10851.02 respectively.

Code-All-That-Apply Questions

The NLSY97 includes questions that allow respondents to give multiple responses or code all responses that apply. In the CAPI codebook, each possible response constitutes its own variable. In each codeblock for a given code-all-that-apply question, frequencies for **all** possible responses are represented. However, each specific variable contains **only** the valid data for one specific possible response, as researchers will note when extracting data. Reference numbers are assigned in the same manner as described for date-entry questions. A base reference number is assigned to the first possible response, with a decimal value being appended to that base number for each following possible response. For example, variables R02461.00–R02461.11 (Activities to Find a Job Emp 01) provide responses (coded yes/no) for 12 different methods

of finding a job. Although the codeblock for R02461.01 represents the frequencies for all possible responses to the question, accessing the data for that specific variable will produce only the data for the response “Contacted employment agency.”

Machine-Generated Check Items

Many questions in the NLSY97 data are “machine checks.” In these questions, previously reported information is checked by the computer and computations are made automatically, causing the appropriate skip pattern to be executed without intervention by the interviewer. For example, the survey program may check the respondent’s gender before asking women whether they have ever been pregnant and men whether they have ever fathered a child. In an effort to clarify the skip patterns present in the instrument, a large number of these machine checks appear in the codebook. The text of machine checks is generally in machine language or equation form; the codeblock also includes a note clarifying the purpose of the check. For example, R02678., “Chk R Female (Pregnancy Leave) Emp 01,” includes the following code:

```
([male/female] = 2);
```

The codeblock also contains a translation of the code which is more accessible to users:

```
/* Is respondent a female? This determines whether the paid pregnancy leave questions which follow are asked. */
```