DOCUMENT RESUME

ED 221 593 TM 820 631

AUTHOR Henderson, Louise B.; Allen, Danny R.
TITLE NLS Data Entry Quality Control: The Fourth Follow-Up Survey.
INSTITUTION Research Triangle Inst., Durham, N.C. Center for Educational Research and Evaluation.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO RTI/0884/61/19S
PUB DATE Jun 81
CONTRACT DHEW-OEC-0-73-6666
NOTE 19p.; National Longitudinal Study (NLS) sponsored reports series.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; *Data Processing; High School Graduates; *Information Storage; Input Output Devices; Longitudinal Studies; Mathematical Models; *Quality Control; Questionnaires; *Research Problems; Sampling; Secondary Education
IDENTIFIERS *Error Detection; *National Longitudinal Study High School Class 1972

ABSTRACT

The data entry quality control procedures in discrete data entry tasks in the National Longitudinal Study (NLS) Fourth Follow-Up Survey are examined. Direct data entry terminals were used to key survey questionnaire item responses, telephone interview corrections, respondent background information and supplemental questionnaire responses into computer disk storage. Data entry error rates were computed on the survey questionnaires by selecting a random sample from each batch after initial keying of the data, rekeying the selected questionnaires by two additional operators and determing error rates on the basis of three keyings. In the implementation described, the overall error rate tolerance established for the NLS survey was not exceeded. The variable error rate over samples and operators on the selected supplemental questionnaires was 0.00040; estimated character error rate was 0.00023. The telephone interview additions and corrections, and directory information entry procedures are described. (CM)

NATIONAL LONGITUDINAL STUDY
SPONSORED REPORTS SERIES

NLS DATA ENTRY QUALITY CONTROL: THE FOURTH FOLLOW-UP SURVEY

National Center for Education Statistics

RTI/0884/61-19S

NATIONAL LONGITUDINAL STUDY
SPONSORED REPORTS SERIES

NLS DATA ENTRY QUALITY CONTROL: THE FOURTH FOLLOW-UP SURVEY

by

Louise B. Henderson
Danny R. Allen

Center for Educational Research and Evaluation
Research Triangle Institute
Research Triangle Park, N.C. 27709

Prepared for
National Center for Education Statistics
Education Division
U.S. Department of Health, Education, and Welfare

Andrew J. Kolstad, Project Officer

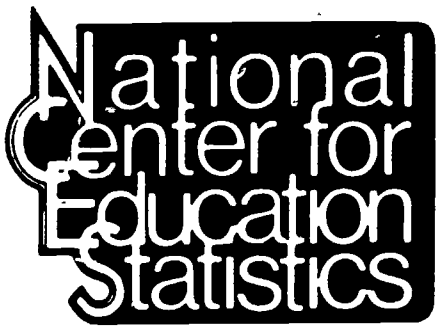
U.S. DEPARTMENT OF EDUCATION
T. H. Bell, *Secretary*

Office of Educational Research and Improvement
Donald J. Senese, *Designate, Assistant Secretary for Educational Research and Improvement*

National Center for Education Statistics
Marie D. Eldridge, *Administrator*

NATIONAL CENTER FOR EDUCATION STATISTICS

"The purpose of the Center shall be to collect and disseminate statistics and other data related to education in the United States and in other nations. The Center shall . . . collect, collate, and, from time to time, report full and complete statistics on the conditions of education in the United States; conduct and publish reports on specialized analyses of the meaning and significance of such statistics; . . . and review and report on educational activities in foreign countries."—Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).



This report was prepared for the National Center for Education Statistics, Education Division, under contract No. OEC-0-73-6666 with the U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects are encouraged to express freely their professional judgment. This report, therefore, does not necessarily represent positions or policies of the Education Division, and no official endorsement should be inferred.

June 1981

ACKNOWLEDGEMENTS

This report has benefited from the close cooperation and combined efforts of several individuals in addition to the authors. Special acknowledgements are due Mr. Mark Watson and Ms. Debra Duncan of RTI's Computer Applications Center for their programming assistance in construction of the fourth follow-up data entry quality control files. Acknowledgements are also due Mr. Peter Stowe of the National Center for Education Statistics, Dr. John Riccobono, NLS Project Director, and Dr. Graham Burkheimer for their significant input and assistance in the preparation of this report. Special thanks are also due Ms. Sarah Aster and Ms. Pam Mikels for their excellent secretarial support.

SUMMARY

The NLS Fourth Follow-up data collection activities began in October 1979 and were completed by May 1980. Data collected were coded, edited, and keyed directly into computer disk storage by operators through programmable direct data entry terminals, as in previous follow-up surveys. Several discrete data entry tasks were involved (follow-up questionnaire, item responses and directory information; telephone interview forms; and Supplemental Questionnaires) and this report describes the data entry quality control procedures implemented for these specific tasks. Data entry errors for fourth follow-up keying operations are estimated to be less than two in one thousand.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Fourth Follow-Up Questionnaire data were keyed directly by operators into computer disk storage through programmable direct data entry terminals. There are several advantages to direct data entry versus standard keypunch operations, the primary advantage being the ability to perform certain data checks at the time of entry. Direct data entry also eliminates the need for most manual coding of data as well as rekey verification required in the standard keypunch-verify approach to recording and transmitting data. Lower error rates also result from direct data entry.

The NLS fourth follow-up survey included several data entry tasks, i.e., Fourth Follow-Up Questionnaire item responses, Fourth Follow-Up Questionnaire telephone interview corrections, respondent background information, and Supplemental Questionnaire responses. The data entry quality control procedures for each of these tasks will be discussed in the following sections.

II. FOURTH FOLLOW-UP AND SUPPLEMENTAL QUESTIONNAIRE DATA ENTRY

In the first NLS follow-up, the overall data entry error rate was determined by sight-verification of a random sample of keyed questionnaire data versus the original hardcopy item responses. Probable biases in error rate calculations using this procedure were due to oversights and fatigue, common problems in the visual comparison of data. To eliminate biases introduced by these inaccuracies, a computer-matching procedure for determining error rates was developed for use in future follow-up surveys. As in second and third follow-up data entry, this procedure was used in calculating error rates for Fourth Follow-Up Questionnaire item response data entry and Supplemental Questionnaire keying. The basic steps in computing error rates for these two data entry tasks are described below.

A. Procedure

1. General

Completed Fourth Follow-Up Questionnaires and Supplemental Questionnaires were separately batched on receipt and routed to direct data entry following initial editing and code assignment. The basic procedure for estimating the data entry error rate for both of these NLS instruments was as follows:

(a) A simple random sample of questionnaires was selected from each batch after initial keying of the data.

(b) The selected questionnaires were rekeyed by two additional operators.

(c) Error rates were determined on the basis of computer matching of the three separate keyings (original and two rekeys).

2. Sampling

By mutual agreement, three questionnaires from each batch of 50 were to be selected for rekey, for a targeted sampling rate of six percent. An automated sampling routine designed to select, at the time of data entry, this six percent sample was implemented at the start of data entry activity. Although not immediately recognized, problems were encountered in computer sampling (machine problems as well as inconsistencies in code) such that in many cases fewer than three questionnaires per batch were automatically selected. Consequently, a manual sampling procedure (using a table of random numbers) was employed subsequently to ensure that exactly three instruments from each batch were selected. Since the exact manual sampling procedure was implemented several weeks after keying began, the realized sampling rate for Fourth Follow-Up Questionnaire data entry quality control was approximately five percent,[1/] which still provided good overall estimates as well as sufficient continued monitoring of the quality of the keying operation. A total of 922 sets of triplicate Fourth Follow-Up Questionnaires and 272 sets of triplicate Supplemental Questionnaires were selected in this manner.

3. Error Model

To estimate the error rate for original keying, let ε_1, ε_2, and ε_3 be the probability of a keying error for the initial data entry operator, the first rekey operator, and the second rekey operator, respectively. (It is not assumed that $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$.) Let N denote the number of elements (either single key-stroke characters or groups of characters defining a particular questionnaire item) involved in the records used for quality check. These N elements were independently keyed by the three operators. Thus, assume that the errors made by data entry operators are independent.

[1/] The problems with sampling by computer were recognized before Supplemental Questionnaire keying began. Thus, the manual sampling procedure was used from the start of Supplemental Questionnaire data entry, resulting in a realized sampling rate of six percent.

Further, let

n_a = number of elements on which operators 1, 2, and 3 matched;

n_b = number of elements on which operators 1 and 2 matched but operator 3 did not;

n_c = number of elements on which operators 1 and 3 matched but operator 2 did not;

n_d = number of elements on which operators 2 and 3 matched but operator 1 did not;

n_e = number of elements on which no two operators matched.

Clearly, $n_a + n_b + n_c + n_d + n_e = N$. An element is assumed to be correctly keyed only when the master or initial keying matches at least one of the two rekeys (n_a, n_b, and n_c each denote numbers of correctly keyed variables).

Let $P_i = n_i/N$, (i = a, b, c, d, e), be the proportion of elements falling into category "i"; then the expected values of these proportions, $E(P_i)$, are given by:

$$E(P_a) = (1-\varepsilon_1)(1-\varepsilon_2)(1-\varepsilon_3)$$

$$E(P_b) = (1-\varepsilon_1)(1-\varepsilon_2)\varepsilon_3$$

$$E(P_c) = (1-\varepsilon_1)(1-\varepsilon_3)\varepsilon_2$$

$$E(P_d) = (1-\varepsilon_2)(1-\varepsilon_3)\varepsilon_1$$

$$E(P_e) = \varepsilon_1\varepsilon_2\varepsilon_3 + (1-\varepsilon_1)\varepsilon_2\varepsilon_3 + (1-\varepsilon_2)\varepsilon_1\varepsilon_3 + (1-\varepsilon_3)\varepsilon_1\varepsilon_2.$$

The empirically established error rate for experienced RTI data entry operators is less than half a percent; therefore, ε_1, ε_2, and ε_3 are assumed to be less than .005. Consequently, as a first approximation terms of the type $\varepsilon_i\varepsilon_j$ and of higher order (i.e., $\varepsilon_i\varepsilon_j\varepsilon_k$) may be omitted. Consequently,

$$E(P_a) \cong 1 - (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

$$E(P_b) \cong \varepsilon_3$$

$$E(P_c) \cong \varepsilon_2$$

$$E(P_d) \cong \varepsilon_1$$

$$E(P_e) \cong 0$$

A first approximation to the estimate can be obtained by equating the sample quantities P_b, P_c, and P_d to their approximate expectations.[2/] The standard error of the error rate estimate can be calculated by first computing the error rate estimate, $\hat{\varepsilon}_1$, for each record and then determining the variance of $\hat{\varepsilon}_1$ over records. Although the errors in elements within a record are likely to be correlated with each other, the assumption of independence between records is more tenable.

4. Implementation

All completed Fourth Follow-Up Questionnaires and Supplemental Questionnaires, returned by mail either from individual sample members or from NLS field interviewers, were separately batched in groups of 50 or less. A Batch Header Sheet was produced containing all ID numbers in a given batch, and questionnaires were subsequently identified and accounted for by this batch control form which detailed the action on each questionnaire within the batch.

Following initial editing and code assignment, the batches of Fourth Follow-Up Questionnaires and Supplemental Questionnaires were assigned to the data entry operators who were responsible for keying all questionnaires in their assigned batches. NLS data entry task leaders randomly selected three questionnaires per batch for quality control purposes, using the procedures previously described.[3/] The three questionnaires selected to be rekeyed were removed from the batch and labeled "REKEY" on the front cover to denote its selection in the quality control sample. The NLS ID numbers for the selected instruments were also circled on the Batch Header Sheet by the task leader. An indicator variable identifying whether or not a particular questionnaire was sampled was keyed into the magnetic data record, for use in constructing the file of sample instruments for quality control purposes.

Questionnaires selected for the quality control sample were then rekeyed by two additional operators; the data entry procedure for rekeying was identical to the initial keying. Problems of interpretation and readability were

---

[2/] More exact estimates of rates and their standard errors may be obtained through maximum likelihood procedures. Since the likelihood equations are nonlinear and computation rather complex, it was decided to use P_d as the estimator of ε_1 or the error rate for original keying.

[3/] Some sample selection by computer was implemented at the beginning of the data entry process.

handled identically for the rekey operation as in the initial keying, constituting a completely "blind" rekey effort to provide more accurate estimates of keying error.

B. Error Rates for Fourth Follow-Up Questionnaire Data

For Fourth Follow-Up Questionnaire data entry quality control purposes, two data entry error rates were computed, one based on the number of variables (questionnaire items) keyed and the other based on the number of individual characters keyed (one or more per variable). For example, "040" hours would be considered one variable consisting of the three characters: "0," "4," and "0." A total of 922 sets of triplicate questionnaires were sampled. The triplicate records were compared variable-by-variable and character-by-character (excluding open-ended questionnaire items) by a computer program which identified the variables (questionnaire items) and characters (within variables) that were not keyed in exactly the same manner. As indicated above, the master keying of a variable or character was considered correct if matched by at least one of the two rekeys. Simple counts of the number of rekeyed variables and characters for which neither rekey matched the initial keying were computed, and these counts were converted to error rates by dividing by the number of keyed variables and the number of keyed characters, respectively. The resulting overall variable and character error rates for individual direct data entry operators are presented in Table 1.

From the start of fourth follow-up data entry operations,[4/] computer reports were generated at various points in the process to indicate the overall variable and character data entry error rates. A computer listing of the variable (questionnaire item) errors that were detected in each report was produced simultaneously. During initial data entry activity, reports generally were produced on a weekly basis and later on a biweekly basis as the number of questionnaires received at RTI decreased. However, the frequency of these quality control reports varied, depending on such factors as the number of

[4/] As new operators were trained for NLS data entry, printouts of at least six test questionnaires keyed by the new operators were manually compared with the respective hard copy instruments by NLS project staff. The new operators were given additional instruction/retraining as necessary before beginning production keying.

Table 1.--Fourth Follow-Up Questionnaire variable and character error rates by operator

NLS operator number	Number of keyed questionnaires sampled[1]	Number of variables keyed	Operator variable error rate	Number of characters keyed	Operator character error rate
1	9	6966	0.00172	18171	0.00088
2	81	62694	0.00085	163539	0.00058
3	65	50310	0.00109	131235	0.00104
4	24	18576	0.00124	48456	0.00186
5	2	1548	0.00065	4038	0.00149
6	2	1548	0.00258	4038	0.00198
7	22	17028	0.00147	44418	0.00122
8	3	2322	0.00000	6057	0.00000
9	20	15480	0.00168	40380	0.00151
10	3	2322	0.00301	6057	0.00528[2]
11	66	51084	0.00057	133254	0.00035
12	50	38700	0.00034	100950	0.00040
13	43	33282	0.00048	86817	0.00046
14	36	27864	0.00032	72684	0.00039
15	36	27864	0.00269	72684	0.00259
16	38	29412	0.00071	76722	0.00042
17	77	59598	0.00305	155463	0.00176
18	10	7740	0.00103	20190	0.00094
19	40	30960	0.00362	80760	0.00300
20	52	40248	0.00186	104988	0.00152
21	1	774	0.01292	2019	0.00941[2]
22	8	6192	0.00113	16152	0.00093
23	47	36378	0.00443	94893	0.00349
24	75	58050	0.00053	151425	0.00038
25	6	4644	0.00409	12114	0.00256
26	50	38700	0.00173	100950	0.00135
27	7	5418	0.00055	14133	0.00042
28	12	9288	0.00603	24228	0.00417
29	7	5418	0.00129	14133	0.00092
30	13	10062	0.00020	26247	0.00011
31	7	5418	0.00757	14133	0.01465[2]
32	7	5418	0.00111	14133	0.00127
33	3	2322	0.00345	6057	0.00495

[1] Although each operator was responsible for one or more batches, the number of sampled questionnaires is not always a multiple of three due to problems with computer sampling discussed earlier.

[2] Although the individual operator error rate is greater than 0.00500, the overall data entry error rate never exceeds the contractually specified tolerance level of .5 percent (see Figure 1). Newly trained operators 10, 21, and 31 keyed NLS data for only a short period of time as indicated by the minimal numbers of keyed questionnaires on which their error rate calculations are based.

NOTE.--There are 774 variables and 2019 characters per Fourth Follow-Up Questionnaire. Open-ended responses and certain variables constant across records, e.g., project number and data entry form number, were not used in determining error rates.

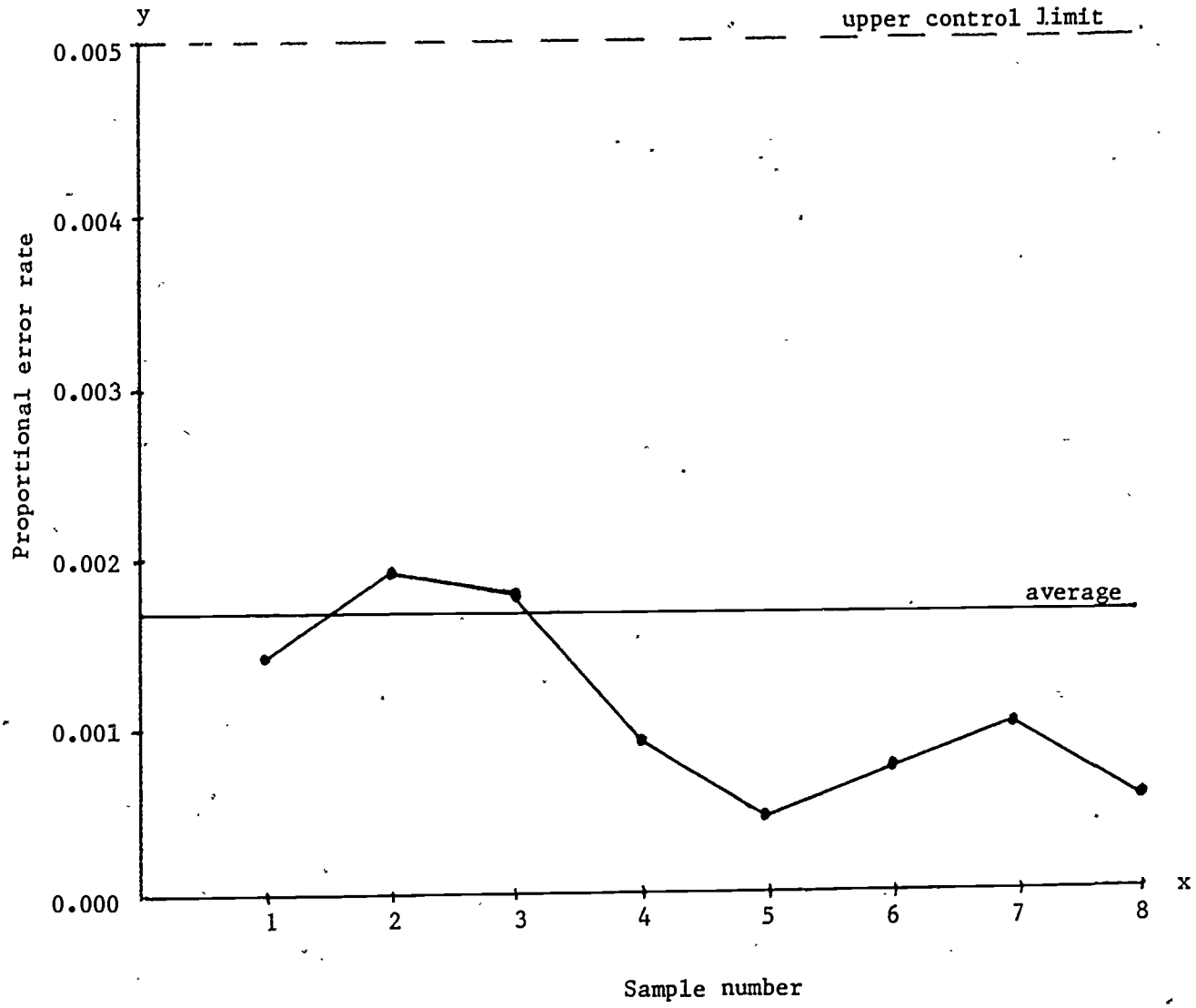
operators keying, the number of questionnaires keyed, and the use of a second shift of data entry operators. Interim quality control reports were generated as necessary for the purpose of keeping close checks on operator performance (e.g., when newly trained operators were first in production mode); however, these interim data were not used for reporting purposes.

Figure 1 presents the overall (over operators) error rate results for variables (questionnaire items) from the eight major data entry quality control reports for Fourth Follow-Up Questionnaire data entry. From the data, it is evident that the 0.005 (.5 percent) overall error rate tolerance established for the NLS survey was not exceeded at any time point. Over time the error rates ranged from a high of 0.00188 (early in the data entry process) to a low of 0.00046. Based on the total sample of 922 selected questionnaires, the estimated variable error rate was 0.00163 (based on 713,628 keyed variables) and the estimated character error rate was 0.00136 (based on 1,861,518 keyed characters).

C. Error Rates for Supplemental Questionnaire Data

The procedure for determining Supplemental Questionnaire data entry error rates also consisted of selecting a six percent random sample of questionnaires from each keyed batch and resulted in a total of 272 sets of triplicate Supplemental Questionnaires. Errors were calculated as described above through variable-by-variable and character-by-character comparison of the triplicate records. The resulting Supplemental Questionnaire variable and character error rates for the individual direct data entry operators are presented in Table 2. Since Supplemental Questionnaire data were keyed primarily by Fourth Follow-Up Questionnaire data entry operators and since a six percent sample of returned instruments resulted in only 272 sets of triplicate questionnaires, only a few interim quality control reports were generated for the purpose of checking each operator's performance. Based on the 272 selected Supplemental Questionnaires, the variable error rate, over samples and operators, was 0.00040 (based on 42,704 keyed variables) and the estimated character error rate was 0.00023 (based on 102,272 keyed characters).

Figure 1.--Fourth follow-up questionnaire variable error rate by sample



x = Computer report number	y = Error rate	Number of questionnaires 1/ on which error rate calculation based
1	0.00139	273
2	0.00188	72
3	0.00175	255
4	0.00085	47
5	0.00046	89
6	0.00074	42
7	0.00104	36
8	0.00057	41

average line: $y = 0.00163$

1/ The total number of records for error rate reports 1-8 does not equal the number of records (922) for which the total error rate was calculated. Each of the eight groups of questionnaires contained incomplete sets of keyings for several sample instruments (e.g., the original keying and first rekey with no second rekey present). No adjustments were made for these cases in the eight individual reports, but many of these incomplete sets of questionnaires were completed for purposes of computing the total error rate.

Table 2.--Supplemental Questionnaire (SQ) variables and character error rates by operator

NLS SQ operator number	Number of questionnaires keyed	Number of variables keyed	Operator variable error rate	Number of characters keyed	Operator character error rate
1	3	471	0.00000	1128	0.00000
2	25	3925	0.00076	9400	0.00032
3	86	13502	0.00022	32336	0.00015
4	57	8949	0.00011	21432	0.00005
5	78	12246	0.00073	29328	0.00044
6	23	3611	0.00028	8648	0.00023

NOTE.--There are 157 variables and 376 characters per Supplemental Questionnaire. As in Fourth Follow-Up Questionnaire data entry, open-ended responses and certain variables constant across records, such as project number and data entry form number, were not used in computing error rates.

III. FOURTH FOLLOW-UP QUESTIONNAIRE TELEPHONE INTERVIEW ADDITIONS AND CORRECTIONS

As in previous follow-up surveys, a set of "key" or critical questionnaire items were defined for fourth follow-up. If any of these key items were indeterminate (omitted or answered partially or inconsistent), then additional data collection procedures were implemented, consisting of attempts to resolve such indeterminacy through a telephone interview. The identification of indeterminacies was accomplished by a computer edit process (replacing the manual editing process used in prior follow-up surveys), which was applied to the set of key items once the data were keyed into machine-readable form.

As data from each questionnaire were computer-edited, a computer-generated problem sheet containing a list of questions and corresponding responses needing clarification or completion was produced for each questionnaire that failed the computer-edit process. The "fail-edit" questionnaires and their problem sheets were routed to telephone interviewers, who were responsible for contacting sample members and clarifying discrepancies, omissions, or inconsistencies in the questionnaire. All item corrections/resolutions were recorded on an answer sheet that provided for correction of all "key" or critical items, as necessary. These "fail-edit" answer sheets (with their associated questionnaire and computer-generated problem sheets) were resubmitted to data entry, following any required manual coding, where only the new data recorded on the answer sheet by telephone interviewers were keyed, transmitted, and merged with the previously keyed questionnaire responses.

Since both the number of key items and the number of respondents failing edit were small, all such additions and corrections obtained from the telephone interview process were 100 percent verified. This verification process involved a rekeying of data recorded on the answer sheet together with identifying information such as batch number, NLS ID number, and a short label (8-character mnemonic) for each questionnaire item with corrections data present. These corrections/additions were verified by a different operator than the original keyer, and the verifying operator corrected, during the key-verification process, any errors found in the initial keying.

IV. FOURTH FOLLOW-UP DIRECTORY INFORMATION ENTRY

One further data entry activity was instituted to ensure additional accuracy in keying directory information (Section G of the Fourth Follow-Up Questionnaire). These data were entered as a separate step after all other questionnaire items were keyed. This information (e.g., name and address, phone number, social security number, driver's license number) was 100 percent verified by a different operator than the original keyer. The verifying operator corrected any errors detected in the initial keying.