ED 151 392 TM 006 850

AUTHOR Levinsohn, Jay R.; Riccobono, John A.

TITLE The National Longitudinal Study of the High School Class of 1972: A Description of the Merged Base-Year/First Follow-Up Data File.

PUB DATE Apr 75

NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Washington, D.C., March 30-April 3, 1975); For related documents, see TM 006 788 and 858

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS Academic Ability; *Data Bases; Data Collection; Data Processing; Editing; *Electronic Data Processing; *Followup Studies; Grade 12; Graduate Surveys; *High School Graduates; *Longitudinal Studies; Magnetic Tapes; National Surveys; Questionnaires; Sampling; School Statistics; Senior High Schools; *Statistical Data; Student Records; Youth

IDENTIFIERS *National Longitudinal Study High School Class 1972

ABSTRACT

The National Longitudinal Study of the High School Class of 1972 (NLS) is a large-scale longitudinal survey effort. The current data tapes contain base-year (1972) survey data integrated with first follow-up (1973) survey data. The base-year data consist of responses to a Student Questionnaire, the results of an ability test, and information from the Student's School Record Information Form. Follow-up data consist of responses to a mailed questionnaire. The pre-machine data processing consisted of coding alphabetical information and "errors", such as partial responses or legitimate non-responses. Subsequent machine editing included range checks, consistency checks, and routing checks. There are 22,398 records in the NLS merged base-year/first follow-up data file--one record for each respondent. Each record consists of 921 variables. The data contained in each record in the file are generally grouped by type and ordered as follows: identification codes, data indicators, test battery data, student's School Record Information Form data, base-year Student Questionnaire data, First Follow-Up Questionnaire data, quality indices sampling weights, school variable data, and analytic indices. These tapes are now available for purchase from the National Center for Education Statistics. (BW)

TM

The National Longitudinal Study of the High School Class of 1972: A Description of the Merged Base-Year/First Follow-up Data File

Jay R. Levinsohn and John A. Riccobono
Center for Educational Research and Evaluation
Research Triangle Institute

A Paper Presented at the
American Educational Research Association Annual Meeting
in Washington, D. C., April 1975

ABSTRACT

The National Longitudinal Study of the High School Class of 1972 is a large-scale longitudinal survey effort. NLS data have been gathered and combined from several sources, coded and edited for analysis purposes, and stored on magnetic computer tapes. The current tape package contains base-year (1972) survey data integrated with first follow-up data (1973). This paper provides a general description of the contents of the tape package which will soon be available for use by the general research community.

I. INTRODUCTION

The National Longitudinal Study (NLS) of the High School Class of 1972 is a large-scale longitudinal survey effort primarily supported by the National Center for Education Statistics (NCES), Office of the Assistant Secretary for Education in the Department of Health, Education, and Welfare (DHEW).

The primary purpose of the NLS is to discover what happens to young people after they leave high school, as measured by their subsequent educational and vocational activities, plans, aspirations, and attitudes, and to relate this information to their prior educational experiences and personal and biographical characteristics. Ultimately, the study will allow a better understanding of the development of students as they pass through the American educational system, and of the complexity of factors associated with individual educational and career outcomes.

NLS data have been gathered and combined from several sources, coded and edited for analysis purposes, and stored on magnetic computer tapes for future access. The current tapes contain base-year (1972) survey data, collected by the Educational Testing Service, integrated with first follow-up (1973) survey data, collected by the Research Triangle Institute. This tape package will be augmented periodically as data from subsequent follow-up surveys become available.

The major purpose of this paper is to describe the contents of the available NLS tapes and, by so doing, stimulate the interest of the educational research community.

II. THE SAMPLE

The merged NLS data file contains data for a total of 22,398 sample members who responded to the base-year Student Questionnaire and/or the First Follow-Up Questionnaire. Comprising this total group are 15,635 individuals who responded to both survey questionnaires, 1,048 who responded to the base-year questionnaire only, and 5,715 who responded to the follow-up questionnaire only (although some base-year information was collected retrospectively for 4,539 of these individuals).

These respondents were sampled as 1972 twelfth graders from a national probability sample of 1,318 public, private, and church-affiliated high schools located throughout the 50 States and the District of Columbia. The school sampling frame was stratified into 600 final strata, based on the following seven variables:

- Type of control (public or nonpublic),
- Geographic division (Northeast, North Central, South, West),
- Grade 12 enrollment (fewer than 300, 300-599, 600 or more),
- Proximity to institutions of higher education,
- Percent minority group enrollment,
- Income level of the community around the school, and
- Degree of urbanization.

Students from low-income families and racial minorities were over-sampled so that sample sizes for these groups in the national sample would be adequate for separate analyses.

The study design excluded schools for the physically or mentally handicapped, schools for legally confined students, or schools that did not enroll students of their own (such as area vocational schools whose students were enrolled in other schools included in the sample). Also excluded were certain classes of students, such as early graduates and adult education students.

III. INSTRUMENTS AND TEST BATTERY

A. Base-Year Instrumentation

Each student in the sample was asked to complete a Student Questionnaire containing 104 questions distributed over 11 major sections. The base-year Student Questionnaire dealt with factors related to the student's personal-family background, educational and work experiences, plans, aspirations, attitudes, and opinions. In addition, each student was asked to complete a 69-minute Test Book designed by the Educational Testing Service to measure both verbal and nonverbal ability. The Test Book consisted of six tests, including vocabulary, picture number, reading, letter groups, mathematics, and mosaic comparisons.

Base-year data were also obtained from a Student's School Record Information Form (SRIF). Items on the SRIF pertained to the student's high school curriculum, grade-point average, and credit hours in major courses. If applicable, positions in ability groupings, remedial-instruction records, involvement in certain federally supported programs, and scores on standardized tests were also included.

Finally, information from a School Questionnaire and one or two Counselor Questionnaires were also collected for each participating high school.

B. First Follow-Up Instrumentation

Two forms (A and B) of a First Follow-Up Questionnaire were developed and designed for self-administration by the student. Form A of the First Follow-Up Questionnaire was mailed to each sample member who responded to the base-year Student Questionnaire. Seniors from the high school class of 1972 who were unable to participate in the base-year survey (usually because of time and scheduling considerations) were mailed Form B of the questionnaire. Questions one through 85 were identical on both questionnaire forms. These questions dealt with information concerning the respondent's activity state (e.g., education, work, etc.) in October 1972 and October 1973, his or her socioeconomic status, work and educational experiences since leaving high school, and future educational and career plans, aspirations, and expectations. Form B of the First Follow-Up Questionnaire also contained an additional 14 questions from the base-year Student Questionnaire. Most of the items of the base-year and first follow-up instruments are of the forced-choice type. Open-ended, or free-response, items were limited to questions involving dates, income, number of hours or weeks worked, and the like.

IV. DATA COLLECTION PROCEDURES

A. Base-Year Data Collection

Of the 1,318 schools, 1,070 were able to participate in the base-year survey, resulting in 16,683 completed Student Questionnaires. The bulk of the student data was collected in April, May, and June of 1972, through group administration in each school. Survey administrators also completed School Record Information Forms (SRIF's) for each participating student, plus approximately 1,000 additional students during the summer of 1972. It was intended that these additional students, who did not complete the Student Questionnaire in spring 1972, would be sent Form B of the First Follow-Up Questionnaire so as to be able to obtain some base-year student information from the respondents.

B. First Follow-Up Data Collection

The first step in data collection for first follow-up involved an extensive tracing operation to update name and address files. The Research Triangle Institute received from the previous contractor a name and address file from the base-year survey. A total of 18,672 of these individuals from 1,043 schools were used for the first follow-up survey. In addition, 257 resurvey schools with 4,450* individuals were added to the base-year lists to give a first follow-up survey sample of 1,300 schools and 23,122 individuals. A newsletter was developed and mailed in July 1973 not only to encourage participation but also to use as a vehicle for updating names and addresses. When mail was returned by the postal service as undeliverable, telephone tracing procedures were used to obtain current addresses where possible.

*The list of resurvey students was provided by the U. S. Office of Education.

Prior to the mailout of the First Follow-Up Questionnaires, 102 individuals were deleted from the mailing list for operational reasons (e.g., duplicate names, bad addresses, deceased).

Questionnaires were mailed to the last known addresses of the resulting sample of 23,020 on 23-24 October 1973. This was followed by a planned sequence of reminder postcards, additional questionnaire mailings, and reminder mailgrams to nonrespondents. Active mail return efforts continued through December 1973; by early January 1974, the questionnaire return rate by mail was 60.9 percent.

The names and addresses of those sample members who failed to mail back their questionnaires were then turned over to the Bureau of the Census for personal interview in accordance with a Bureau arrangement with the U. S. Office of Education. This personal interview phase of first follow-up data collection continued until April 7, 1974, at which time the overall response rate had been increased to 92.8 percent, 21,350 respondents out of 23,020.

V. FIRST FOLLOW-UP DATA PREPARATION

A. Pre-Machine Data Processing

Questionnaires returned by mail, either from individual sample members or from the Bureau of the Census interviewers, were manually edited to ascertain if each questionnaire contained a minimum set of "critical" data. Questionnaires which passed this manual edit stage were then transmitted to the direct data entry section to be transformed into machine-readable format. Those questionnaires that failed the manual edit procedures were routed to the telephone follow-up section, where an attempt was made to contact the respondent and recover missing information or otherwise resolve problems uncovered in manual edit. After resolution, these questionnaires were also transmitted to the direct data entry section for encoding. In general, the encoding process involved transcribing the questionnaire responses onto magnetic tape. However, there were certain categories of questionnaire items that required some manual editing before transcription; i.e., questions dealing with occupations, fields of study, school names, and those items that had an alternative labeled "other" and allowed a write-in response. Manual recoding of alphabetic information was performed prior to transcription into machine-readable form. Questions concerning the respondent's (or his parents') occupation were recoded into the corresponding three-digit codes specified in the Census Index of Industries and Occupations. Postsecondary schools identified by respondents were recoded into six-digit U. S. Office of Education vendor or FICE codes, using a master index provided by OE (these codes have been witheld, however, in order to protect the confidentiality of respondents). Fields of study reported by respondents who attended school during the post-high school period were recoded into HEGIS categories, using both the four-digit academic subdivisions provided by the HEGIS taxonomy and the six-digit HEGIS technological and occupational schemes being developed. Finally, responses indicating a type of license, certificate, or diploma received were similarly converted to numerical codes.

There were 18 questions in the First Follow-Up Questionnaire which had closed or fixed-response alternatives and which included an "other" option with room for the respondent to write in his particular answer. If the "other" category were chosen by the respondent, that response was reclassified whenever possible into the closed-choice options provided. Where reclassification was not an obvious or logical possibility, a code indicating use of the "other" option was inserted, but the alphabetic description was not included on the data file.

There were four questions asking for alphabetic information in the First Follow-Up Questionnaire for which no numeric coding was done. That is, the written replies of the respondents were coded as they appeared on the questionnaire. The user should be aware of these alphabetic fields and use the appropriate computer techniques to process them.

Beyond recoding of alphabetic material, a uniform set of standard numeric "error" codes was defined and applied across the file to indicate certain common classes of erroneous or missing data. These classes of response are listed below:

- Partial Response. This code applied only to those questions in the First Follow-Up Instrument that employ the two-column response format. Each of these questions consists of a set of subitems and requires the respondent to indicate whether each subitem applies or does not apply. If the respondents answered at least one, but not all, of the subitems for a question of this kind, then those unanswered subitems were coded to indicate a partial response to the question.

- Don't Know. (self-explanatory)

- Out-of-Range Response. This code is used when the response or transcription exceeds some specified acceptable range. It is described more fully in the following section.

- Multiple Response. This is used if the respondent gave several answers to a question when the directions called for only one.

- Refusal. (self-explanatory)

- Blank, or Inappropriate Nonresponse. This is used when the respondent should have answered the question and did not.

- Legitimate Nonresponse. This is used when the respondent should not have answered the question and did not. As well, if the respondent did not answer an entire instrument, then all fields representing that instrument were coded 99.

The steps described above were the only manual editing procedures applied to the data. The manual editing was limited to insure, as much as possible, that the data file would accurately reflect the actual response on the questionnaire. It was felt that any further editing should be done by machine to insure uniformity of application and replicability.

VI. MACHINE EDITING PROCEDURES

Some of the major steps taken toward preparing the NLS data tapes for public release included "hard copy" (source document) spot checks and machine editing, which involved recoding all uninterpretable responses and some logical recoding of other responses. As a result, the final data file contains only (a) valid response codes, (b) codes describing type of erroneous or missing data, and (c) "logically recoded values," including, in each case, an indicator for the reasons for such a recode.

Three sequential machine editing programs were employed and are described below. (These programs were not applied to the NLS base-year data, since these data were already subjected to editing procedures employed by the previous contractor. The base-year data were simply reformatted, and in some cases recoded, to achieve consistency with the first follow-up data on the file.)

A. Range Checks

The first program in the editing sequence dealt with out-of-range data. This program checked the responses to each fixed-response item against a specified range of acceptable values, "flagged" any value that fell outside of this range, and recoded each flagged datum to indicate the occurrence of an out-of-range response. It should be noted that this recoding procedure was applied to fixed-choice items only. Acceptable ranges were also specified for 79 free-response items calling for numeric data, and the frequency of responses outside these ranges has been tabulated. In general, these responses were logically possible but were considered highly improbable. It is recommended that the user make his own decisions as to what is acceptable data for these items, using as guidelines the ranges set forth. It was felt that in some cases these outlying responses may provide additional information and that it was best to leave them in the file so as to provide as faithful a transcription of the original records as possible.

B. Consistency Checks

Phase two of the machine editing sequence was concerned with checking the consistency of an individual's responses over the entire questionnaire. A set of 94 internal checks, or response comparisons, were selected on an a priori basis for this purpose. The consistency program read the responses comprising each individual's record and flagged those consistency checks that were failed. As will be described in a later section, an index was subsequently computed for each record based on the number of consistency checks failed by the individual. This index reflects the internal consistency of a record and, therefore, provides the user with a rough indication of the quality of each respondent's data.

C. Routing Checks

The first follow-up instrument contains 33 routing questions. A routing question is one that either implicitly or explicitly directs a respondent around other questions in the instrument. The aim of the routing questions is to quickly move respondents around questionnaire sections that do not apply to them. In order to determine if the respondents correctly followed the routing patterns, a routing check program was developed and implemented. This program read each record and flagged responses to all routing questions that were inconsistent with the subsequent pattern of response, recoding these questions to indicate the type of inconsistency detected. Three types of inconsistency have been identified and recorded in the data file:

1. When the response to a routing item indicates that the questions associated with that item (i.e., contained within the routing pattern) should be skipped but are not.
2. When the response to the routing item indicates that the questions associated with that item should be answered but are not.
3. (Actually a combination of 1. and 2. above.) When the response to the routing item indicates that certain questions should be skipped but are not (type one) and other questions should be answered but are not (type two).

One other function of the routing check program was to differentiate between legitimate nonresponse (coded 99) and illegitimate nonresponse (coded 98). Legitimate nonresponse was defined as nonresponse to questions that the respondent has been routed around. If a respondent was routed into a block, then any nonresponse to those items was considered illegitimate. Also, if the routing pattern was answered in a manner inconsistent with the routing instructions, then the nonresponse was considered illegitimate for these items. The only time that nonresponse would be coded 99 (legitimate nonresponse) was when the respondent had an unflagged response to the routing question that routed him around a group of questions. The effect of this coding for nonresponse was to overestimate the illegitimate nonresponse. That is, in cases where a respondent's pattern of response does not give a clear indication which questions he should answer, then the nonresponse in that pattern was coded 98. In some of the more complex routing patterns, nonresponse will be coded illegitimate (98) for a large section of items due to one inconsistency. This implies that the user should be quite careful in interpreting the 98 and 99 codes. For some analyses, the user [illegible] to redefine legitimate and illegitimate skips.

VII. CONTENTS AND ORGANIZATION OF THE DATA FILE

There are 22,398 records in the NLS merged base-year/first follow-up data file--one record for each respondent. Each record consists of 921 variables. The complete release tape variable and response lists are presented in the NLS Data User's Manual (Levinsohn, Riccobono, and Moore, 1975). These computer generated lists provide a detailed account of the data organization within a record. The variable list contains the name and description of each variable, the field or character positions containing each variable, and a response list reference code for each variable. The response list catalogs the valid response codes for given types of variables in the variable list.

It should be mentioned that a number of variables, or items, from each of the NLS data collection instruments were not included in the release file. Items were excluded primarily for reasons of confidentiality, but there were also a number of items which were deleted or modified because of excessive prior editing or poor response.

The data contained in each record in the file are, <u>in general</u>, grouped by type and ordered as follows: indentification codes, data indicators, test battery data, student's School Record Information Form (SRIF) data, base-year Student Questionnaire data, First Follow-Up

Questionnaire data, quality indices, sampling weights, school-variable data, and analytic indices. Each of these data groups is briefly discussed in the following paragraphs.

Two different identification codes appear at the beginning of each record: a random five-digit student ID number, and a six-digit school code that may be used to group together students from the same high school.

Several indicators have been included in the release file to assist users in processing the data records. These indicators signify the presence or absence of First Follow-Up Questionnaire data, Test Book data, SRIF data, and base-year Student Questionnaire data.

As noted earlier, each student in the base-year survey was asked to complete a battery of six tests, including: vocabulary, picture-number (a two-part test), reading, letter groups, mathematics, and mosaic comparisons (three parts). Total test scores were computed for each of these tests. In addition, subtest scores were computed separately for each part of the picture-number and mosaic comparisons tests. Thus, the test battery data for each record consist of 11 subtotal and total scores in all. For each test or subtest, formula scores were computed from the item response as:

$$FS = R-W/(C-1)$$

where

FS = Formula score

R = Number right

W = Number wrong

C = Number of response to item.

These test scores were also standardized across the sample, and the standardized scores (with a mean of 50 and standard deviation of 10) were stored on the file.

Data from the SRIF and base-year Student Questionnaire were coded and edited according to the previous contractor's specifications. SRIF data in the file include student average, grading system used by the student's high school, lowest and highest grades possible in the school's grading system, and student's percentile rank in class.

Base-year and First Follow-Up Student Questionnaire data were, with few exceptions, coded in a format that was a one-to-one match with the instruments. These data constitute the bulk of the data file.

Unadjusted and adjusted student weights were calculated and stored on the NLS data file. Unadjusted student weights were calculated as the inverse of the sample inclusion probability for all students sampled. A weighting class method was employed to calculate six different sets of student weights adjusted for nonresponse; each weight set is appropriate for analyses involving a particular set of data. The methodology and use of the various weights are described in detail in the NLS Data User's Manual (Levinsohn, Riccobono, and Moore, 1975).

Only three variables from the School Questionnaire were selected for inclusion in each student's data record. They are: region, type of community, and senior class enrollment.

Two kinds of composite indices--quality and analytic--are included in the NLS data file. Four quality indices have been developed in order to quantify the amount and quality of First Follow-Up Questionnaire data present in each record. They are: (1) a consistency index, representing the percentage of the 94 consistency checks failed by an individual; (2) an out-of-range index, representing the percentage of out-of-range responses for an individual's record; (3) a routing index, representing the percentage of routing questions ambiguously answered by an individual, i.e., routing questions that were unanswered or answered in a manner inconsistent with other responses; and (4) a completeness index (for each major section of the questionnaire), representing the precentage of items with valid responses, i.e., responses that are not coded as errors or missing data.

Users should be cautioned that the utility of these quality indices is in the determination of the credibility of individual records. They are of no use in making judgements about data such as item responses when considering more than one respondent, since the real test of item response quality is the over-subjects distribution. These indices should not be considered for discarding subjects unless one's concern is with the entire instrument.

Also included in the data file are two composite indices, ability and socioeconomic status (SES), useful for many analyses. In contrast to the quality indices which focus on a single individual and quantify the amount and quality of information present in an individual's record, the analytic indices are derived from global considerations of the entire file and serve primarily as classificatory variables by which one can group individuals. Thus, each individual was assigned a code of 1, 2, or 3 depending on whether his ability or SES composite score was in the lower, middle two, or upper quartile range.

Both of the analytic indices involved several components and required several steps in their derivation. It should be added, however, that since other procedures and components may be employed in deriving such indices, the ability and SES raw scores were also included in the file, and users are encouraged to decide for themselves whether the derived indices are appropriate for their particular needs or purposes.

Table 1 presents a breakdown of the kinds and amounts of data available on the NLS release file for important subpopulations.

VIII. HOW TO OBTAIN DATA TAPE

A set of NLS tapes containing merged base-year and first follow-up (first level of edit) data is now available for purchase as two 2,400 foot reels of 1600 BPI density at a cost of $156.00 per set. This purchase price also includes a User's Manual. Information on how to obtain a set of these data tapes and file documentation may be obtained from Robert A. Heintze, National Center for Education Statistics, Room 3069, FOB-6, 400 Maryland Avenue, SW, Washington, D. C. 20202.

Table 1. Data availability for subpopulations by instrument

(N = 22,398)

SUBPOPULATION	INSTRUMENT							
	Base-Year (B-Y) Student Quest.	First Follow-Up (FFU) Quest.	SRIF	Test Battery TB	B-Y, FFU	B-Y FFU, SRIF	B-Y FFU, TB	B-Y, FFU SRIF, TB
Sex:								
Male	8,275	10,463	10,233	7,894	7,665	7,658	7,307	7,301
Female	8,397	10,841	10,376	7,953	7,967	7,957	7,553	7,544
Unclassifiable	13	46	42	12	3	3	3	3
Race:								
White	12,656	15,272	14,721	12,111	11,949	11,938	11,442	11,433
Black	2,083	2,739	2,592	1,906	1,920	1,917	1,766	1,763
Other	1,605	1,829	1,816	1,508	1,468	1,465	1,379	1,376
Unclassifiable	339	1,510	1,522	334	298	298	276	276
H.S. Program:								
Academic	6,811	8,511	8,312	6,531	6,468	6,465	6,204	6,201
General	5,665	7,492	7,253	5,363	5,235	5,224	4,955	4,944
Voc/Tech	4,201	5,148	5,063	3,956	3,927	3,924	3,699	3,698
Unclassifiable	6	199	23	9	5	5	5	5
Region:								
North	3,618	4,483	4,316	3,521	3,420	3,420	3,323	3,323
Central	4,568	5,541	5,468	4,122	4,292	4,288	3,875	3,873
South	5,513	7,691	7,242	5,382	5,186	5,178	5,057	5,049
West	2,984	3,635	3,625	2,834	2,737	2,732	2,608	2,603
Unclassifiable	--	--	--	--	--	--	--	--
Ability:								
Low	4,788	4,392	4,783	4,798	4,382	4,374	4,382	4,374
Medium	7,000	6,600	6,997	7,008	6,592	6,585	6,592	6,585
High	4,052	3,890	4,052	4,053	3,889	3,889	3,889	3.889
Unclassifiable	843	6,468	4,819	--	772	770	--	--
SES:								
Low	5,076	6,423	6,227	4,775	4,735	4,729	4,458	4,453
Medium	7,816	9,635	9,393	7,448	7,320	7,310	6,971	6,962
High	3,667	4,686	4,499	3,525	3,506	3,505	3,370	3,369
Unclassifiable	124	606	532	111	74	74	64	64
Total	16,683	21,350	20,651	15,859	15,635	15,618	14,863	14,848

REFERENCE

Levinsohn, J. R., Riccobono, J. A., and Moore, R. P. National longitudinal study of the high school class of 1972: base-year and first follow-up Data File User's Manual. Research Triangle Park, North Carolina: Research Triangle Institute, February, 1975.