DOCUMENT RESUME

ED 151 396 TM 006 856

AUTHOR Conger, Anthony J.; And Others
TITLE National Longitudinal Study of the High School Class of 1972. Reliability and Validity of National Longitudinal Study Measures: An Empirical Reliability Analysis of Selected Data and a Review of the Literature on the Validity and Reliability of Survey Research Questionnaires.
INSTITUTION Research Triangle Inst., Durham, N.C. Center for Educational Research and Evaluation.
SPONS AGENCY National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE Jul 76
CONTRACT OEC-0-73-6666
NOTE 49p.; For related documents, see TM 006 854 and 859

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Academic Ability; *Data Analysis; *Data Collection; Demography; Grade 12; Graduate Surveys; Guidelines; High School Graduates; Item Analysis; Literature Reviews; Longitudinal Studies; *National Surveys; Performance Factors; Post Secondary Education; Question Answer Interviews; *Questionnaires; Racial Differences; Research Methodology; Senior High Schools; Sex Differences; Social Differences; Student Characteristics; Test Items; *Test Reliability; *Test Validity; Youth
IDENTIFIERS Mail Surveys; *National Longitudinal Study High School Class 1972

ABSTRACT

A review of the literature on the validity and reliability of survey data is presented prior to an analysis of the reliability of selected questions in the Second Followup Questionnaire of the National Longitudinal Study of the High School Class of 1972 (NLS). The reliability study includes an evaluation of test-retest reliability as a function of data collection procedures (whether mail-in or personal interview), item characteristics (response format, item content, and item length), respondent characteristics (sex, ethnicity, socioeconomic status, and academic ability), and the interaction of these diverse factors. At issue is not only the quality of NLS questionnaire data, but the consideration of guidelines for analyzing survey data and improving the quality of that data. In lieu of an empirical analysis of validity that might have encroached on the privacy of respondents, the validity check was based on a literature review, focusing on NLS types of items. The validity study also focused on data collection procedures, item characteristics, respondent characteristics, and their interactions. An overview of the purpose of the NLS and the methodologies employed are also included. The Short Form of the Second Follow-up Questionnaire is appended. (Author/EVH)

NATIONAL LONGITUDINAL STUDY of the High School Class of 1972

RELIABILITY AND VALIDITY OF NATIONAL LONGITUDINAL STUDY MEASURES:

AN EMPIRICAL RELIABILITY ANALYSIS OF SELECTED DATA AND A REVIEW OF THE LITERATURE ON THE VALIDITY AND RELIABILITY OF SURVEY RESEARCH QUESTIONNAIRES

Anthony J. Conger
Judith C. Conger
John A. Riccobono

Prepared for the
National Center for Education Statistics
Office of the Assistant Secretary for Education
Department of Health, Education, and Welfare
Contract No. OEC-0-73-6666

by
Center for Educational Research and Evaluation
Research Triangle Institute
Research Triangle Park, North Carolina 27709

July 1976

Foreword

The National Longitudinal Study of the High School Class of 1972 is a large-scale, long-term survey effort supported primarily 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 this effort is to provide statistics on a national sample of students as they move out of the American high school system into the critical years of early adulthood. The base-year survey data, collected by the Educational Testing Service, have been integrated with the first and second follow-up survey data by the Research Triangle Institute (RTI). RTI has processed the data and is presenting major findings in a series of reports, each with a central theme. This report is one of the series and contains empirical information about the reliability of selected NLS Second Follow-Up Questionnaire items and scales, and discusses reliability and validity of NLS type questionnaires and survey research in general. This report evaluates reliability and validity as a function of question characteristics, respondent characteristics, and data collection procedures.

Many people and organizations have contributed generously to the base-year and follow-up surveys, and their efforts are sincerely appreciated. We are especially grateful to the thousands of anonymous sample members who have participated in the surveys, without whose cooperation this continuing study would not have been possible.

J. P. Bailey, Jr.
Director
National Longitudinal Study

Acknowledgment

The authors wish to thank all those people who have contributed to this study and the writing of this report. Dr. Samuel S. Peng was instrumental in the planning and data collection phases of the reliability study, and Ms. Lucia Ward contributed greatly to the literature review. Grateful acknowledgments are also due to Ms. Lois Bressler for her efforts in data processing and to Ms. Diane Brandon for her patient assistance in the preparation of this report. In addition, the helpful editorial comments of Dr. J.P. Bailey, Jr., and George H. Dunteman are sincerely appreciated. Special thanks are also due for the considerable support and input of Mr. Elmer Collins, NCES Project Officer for NLS, Dr. J.A. Davis, Director of the Center for Educational Research and Evaluation at RTI, and Dr. J.P. Bailey, Jr., RTI's Project Director of the NLS study.

Table of Contents

List of Tables

I. INTRODUCTION

This report has two major purposes: a review of the literature on the validity and reliability of survey data; and an analysis of the reliability of selected questions in the Second Follow-Up questionnaire of the National Longitudinal Survey of the High School Class of 1972 (NLS). The key part of the reliability study is an empirical analysis of selected NLS items on a sample of NLS respondents. The reliability study includes an evaluation of test-retest reliability as a function of data collection procedures (mail-in or personal interview), item characteristics (response format, item content, and item length), respondent characteristics (sex, ethnicity, SES, and ability) and the interaction of these diverse factors. The general purpose of the reliability study is to provide information on the quality of NLS questionnaire data; however, a more general discussion of the findings will include guidelines for analyzing survey data and for improving the quality of data in survey studies.

While an empirical analysis of validity would have been desirable, such a study was not undertaken because of concerns about federal policies and pending legislation concerning informed consent and the invasion of privacy. Some concern also existed about the possibility of respondent attrition in reaction to a validity check. In lieu of an empirical analysis of validity, RTI and NCES jointly agreed to an investigation of validity based on a literature review focusing on NLS types of items. This investigation, like the reliability study, considers data collection procedures, item characteristics, respondent characteristics and their interactions.

This report is divided into four major sections. The first section, "A Capsule View of the National Longitudinal Study of the High School Class of 1972," briefly summarizes the purpose of NLS, the sample design, and characteristics of the basic longitudinal questionnaires. The second major section provides a comprehensive review of validity and reliability for NLS type questions and respondents. The third section presents a detailed study of the reliability of a sample of Second Follow-Up questionnaire items on a sample of NLS respondents. The final section integrates the conclusions of the review and reliability study results and discusses the implications of these results for survey research.

II. A CAPSULE VIEW OF THE NATIONAL LONGITUDINAL STUDY OF THE HIGH SCHOOL CLASS OF 1972

The National Longitudinal Study of the High School Class of 1972 (NLS) may be briefly characterized as a wide-ranging longitudinal questionnaire survey on a highly heterogeneous sample. NLS is sponsored primarily by the United States Department of Health, Education, and Welfare and is administered by the National Center for Education Statistics (NCES). Briefly, the overall purpose of the NLS survey is to determine what happens to young adults after they leave high school—as measured by their experiences, plans, aspirations, and attitudes at various points in time. This information is deemed essential for the review and reformulation of federal policies and programs designed to enhance educational opportunity and achievement and to upgrade occupational attainment and career outcomes.

The major vehicle for obtaining data has been mail-administered questionnaires buttressed and augmented by telephone or personal interviews as required. But other information has also been collected. The instruments thus far include: high school records and high school descriptive information for every student and high school comprising the sample. Also, a counselor questionnaire was administered to at least one counselor from every participating high school. Most students in the original base-year sample were administered a test battery covering a range of verbal and nonverbal ability measures; this battery was devised and administered by the Educational Testing Service. Each student has been administered four questionnaires so far: a Base-Year questionnaire administered during the spring of 1972, the First Follow-Up questionnaire mailed in October of 1973, a Second Follow-Up questionnaire mailed in October of 1974, and a Third Follow-Up questionnaire mailed in October 1976. A Fourth Follow-Up is tentatively planned for October 1979. In addition, a replication study of the high school class of 1980 is on the drawing boards.

Most questions (new ones have been added from time to time) have been field-tested and different formats for questions and questionnaires have been evaluated on a 900-member sample of students from the high school class of 1971. In this way, problems with formatting, wording, and item redundancy have been worked out on an independent sample of persons before arriving at the final instrument.

A. Sample Design

The sample design is a deeply stratified two-stage probability sample with schools as first-stage sampling units and students within schools as second-stage units. The population sampled consisted of all 1972 twelfth graders enrolled in *public*, *private*, and *church affiliated* schools in the fifty states and District of Columbia. A variety of strata were used for school selection:

1. Type of control,
2. Geographic region,
3. School size,
4. Percent minority enrollment,
5. Income level of community,
6. Degree of urbanization, and
7. Proximity to institutions of higher learning.

In order to increase the numbers of disadvantaged students in the sample, schools located in low income areas and schools with high proportions of minority group enrollments were sampled at approximately twice the sample rate used for the remaining schools. A variety of other considerations were also employed, and are outlined in detail elsewhere (WESTAT, 1972).

The sample eventually involved about 1,300 primary and back-up schools and over 23,000 students. This breaks down for Base-Year questionnaires into: 1,070 participating schools with 19,144 respondents, and a resurvey during the fall of 1972 provided data from about 4,450 students from 257 schools declining to participate in spring 1972. The resurvey was undertaken to complete the basic sample but some critical data are lacking on these respondents: viz, ability, locus of control, self-esteem; that is, all data which are "soft"—i.e., psychological in nature—were not obtained.

Of the 22,654 students mailed First Follow-Up questionnaires, a total of 21,350 respondents either returned the questionnaire or were personally interviewed (by the Bureau of the Census), yielding the high return rate of 94.2 percent. The Second Follow-Up survey produced an even better return rate—20,876 students or 94.6 percent of the mail-out at that time. Over 90 percent of the Base-Year survey respondents have been retained in the study through two follow-ups.

B. Questionnaire Characteristics

Four kinds of instruments have been used in the NLS survey: school and counselor questionnaires detailing high school and counseling characteristics; a school record information form filled in by the high school for each student in the sample; an ability battery administered to the students surveyed during their senior year in high school; and self-report student questionnaires which provide the longitudinal data. Since this report focuses only on the longitudinal questionnaires, further discussion of the other instruments will not be undertaken. The basic longitudinal questionnaires (administered in the spring of 1972, fall of 1973, fall of 1974, and the fall of 1976) are self-report instruments including complex skip patterns and a variety of formats (simple fill-ins, binary choices, ratings of subjective importance, multiple endorsement among fixed options, single endorsement among fixed options, etc). In addition to a variety of item characteristics, the method of data collection also varied. Three basic data collection methods have been used: self-administration of the questionnaire (handled entirely through the mail); self-administration followed up by telephone interviews to collect critical missing data; and personal interviews by the U.S. Census Bureau (First Follow-Up) and by RTI field staff (Second and Third Follow-Ups) for respondents who failed to mail in their questionnaires by a specified date. This overall data collection strategy has produced a remarkably high return rate, but clearly the quality of data could vary as a function of mode of data collection.

A large variety of topics are covered in the NLS questionnaires. Major sections of the questionnaires are devoted to background information, psychological and social factors, education, work, military, and homemaking. Each of the sections on education, work, military, and homemaking includes questions on activities, plans, and attitudes and opinions. The psychological and sociological topics include consumer and political attitudes and activities, career and life goals, and attitudes about the self. In correspondence with the range of topics and restrictions on questionnaire length, many of the constructs cannot be pursued in depth, thus the instruments may be typified as broad band and low fidelity (Cronbach, 1970).

III. A REVIEW OF SURVEY DATA VALIDITY AND RELIABILITY

A. Introduction

As indicated in the capsule view of the NLS survey, the basic longitudinal questionnaires cover a broad domain of question formats and content. Question types include: factual and subjective information, open-ended and restricted choice formats; and rating scales. The time orientation of the questions varies from retrospective through prospective. Question content includes work, educational, military, and homemaking activities, opinions and plans; family and financial status; and a variety of opinion and attitude questions covering consumer and political arenas, life and career goals, and feelings about the self. Thus, while the NLS questionnaires provide a rich variety of information, the scope of the material covered in conjunction with the variety of data collection procedures and heterogeneity of respondents, plus the novelty or relative novelty of many questions, prohibits an in-depth literature review of validity and reliability. This review of survey question validity and reliability will focus on NLS type data collection procedures, questions, and respondents. However, since it is unlikely that other studies would even approach similarity on all three dimensions, the literature included in this review will be selected for maximum similarity and information on any one or two of these three dimensions. The review will focus primarily on validity and secondarily on reliability as they relate to three major sources of variation:

1. Variation associated with content and item characteristics;
2. Variation associated with the mode of data collection (mail-in questionnaires versus interview obtained data); and,
3. Variation associated with respondent characteristics.

In addition to the problems of obtaining items and respondents maximally similar to NLS items and respondents, other problems should be noted. One major problem in reviewing the literature is the variety of procedures used to obtain information on validity and reliability and the variety of indices used to summarize the results. While this was not unanticipated because of the variety of research disciplines surveyed, it does pose problems for equating findings from various studies. For example, one researcher might conclude on the basis of percent agreement that a particular item is highly valid, while another researcher using a contingency or correlation coefficient might conclude that the same item was relatively invalid. Short of soliciting the original data and recomputing the statistics, there is no way by which such discrepancies could be resolved. We have, however, chosen to emphasize those studies which clearly specify the data collection procedures and which use the most appropriate (for that data set) statistical indices.

In addition, although the classical distinction between validity and reliability is conceptually appealing, the distinction quickly blurs under a variety of conditions. We have attempted to maintain this distinction by referring to validity only when the study compares a respondent's results to data obtained from a factual or independent source of information. The exception to this is the brief review of the validity of psychological variables which generally require a construct validation approach. Reliability is used to refer to results obtained from internal consistency analyses (e.g., coefficient alpha) or test-retest procedures on the same respondents. We have generally not included studies which relate response distributions based on independent samples of respondents.

B. Literature Review

Two unpublished studies have been done on selected NLS data: Ecternacht (1973) carried out a limited study of the reliability and validity of selected NLS Base-Year Student Questionnaire items; and Lyons and Moore (in press) investigated the reliability of selected retrospective Base-Year data collected on the NLS First Follow-up Questionnaire. The research of Ecternacht included a test-retest reliability substudy (similar in design to the reliability study covered in this report) and a proxy validity substudy. While these two substudies had the potential of providing critical information on the quality of NLS data, the analyses were limited to

investigating net and gross difference rates (Hansen, Hurwitz, and Pritzker, 1964) for items for the entire subsample and for males and females separately. For the reliability substudy, Ecternacht concluded that the item reliabilities were generally rather low and recommended that composites be formed to improve data quality. The validity substudy was reported as being methodologically compromised and the responses were sufficiently biased to negate validity estimation. In both substudies by Ecternacht, respondent characteristics (other than sex), item content, and data collection procedures were not investigated. Because of these and other limitations these two substudies are not given further consideration.

The study by Lyons and Moore (in press) also disregarded respondent characteristics and data collection procedures. More importantly, the study focused on the retrospective reliability of a selected set of items in an attempt to specify the degree to which retrospectively collected data could be taken as a valid indicator of an earlier status. In this regard, their results are not particularly relevant to the purpose of this review.

Among others, van Es and Wilkening (1970) assessed the reliability of a standardized interview schedule focusing on item characteristics. Although the major purpose of this study was a comparison of Brazilian and United States data, the results of the U.S. data seem particularly germane. In general, response reliability varied strongly with item content and temporal focus. The most stable items were demographic (70 percent were in the .80-.99 reliability range); next most stable were variables dealing with factually oriented current behaviors (64 percent in the .80-.99 range); third most reliable were variables dealing with factually oriented past behaviors (42 percent in the .80-.99 range); and least stable were evaluations or subjectively oriented items (only 20 percent were in the .80-.99 range).

Astin (1965) obtained similar results from a questionnaire test-retest over a six-week interval based on 107 college students. Questions which had little ambiguity and dealt with important accomplishments (e.g., elected student president) had high rates of stability (95-100 percent). Achievements of a more ambiguous or less important state had slightly lower agreement rates (90-100 percent). Questions dealing with father's occupation, high school grades, and nonacademic accomplishments were subject to greater variability (74-92 percent agreement). Attitudinal items or future plans produced even lower agreement rates (60-70 percent agreement).

The accuracy of survey data shows similar variation. Walsh (1967, 1968) investigated the accuracy of self report on grades, major area of study, number of semester hours completed and high school status (grades, rank in class, etc.). Using only 45 students from three different residence halls, Walsh also attempted to compare accuracy as a function of data collection procedure (questionnaire, interview, or biographical inventory). No differences were detected among the data collection procedures; however, accuracy of information did vary with content. The most accurate information was for number of classes dropped and number of courses in which a D or failing grade was obtained (80-100 percent). Least accurate were self reports on cumulative or recent grade point average (50-80 percent) and retrospective reports of high school grades and class rank (51-53 percent).

A comprehensive study on validity and reliability covering items similar in scope to the NLS questionnaires was carried out by Boruch and Creager (1972). The basic study involved a two-week test-retest by 202 college freshmen on questions dealing with high school experiences and activities, academic and career plans, life goals, and attitudes towards various political and academic issues. In addition, accuracy of grades and age was evaluated on 4,415 respondents.

Because of the similarity of many of their questions to the NLS questions, reliability coefficients of selected variables are reported in detail in Tables 1-3 based on tables presented by Boruch and Creager.

In general, reliability of demographic characteristics, family background variables and high school performance was high (reliabilities were generally above .95). Reliability for items dealing with financial support and concerns about financial support in college were somewhat lower (.85-.90). The items

Table 1

TEST-RETEST RELIABILITIES* FOR ITEMS ON DEMOGRAPHIC CHARACTERISTICS, FAMILY BACKGROUND, HIGH SCHOOL ACHIEVEMENTS, AND COLLEGE ADMISSIONS*

Item	Reliability (r)
Father's Education	.99
Mother's Education	.97
Estimated Parental Income	.98
Concern About Financing College Education	.85
Average High School Grade	.98
Rank in High School	.97
Number of Applications to Other Colleges	.98
Number of Acceptances by Other Colleges	.97
Distance of College From Home	.97

* Based on Boruch and Creager, 1972.

Table 2

TEST-RETEST RELIABILITIES FOR FOUR REPORTED SOURCES OF FINANCIAL SUPPORT DURING FRESHMAN YEAR*

Source of Financial Support During Freshman Year	Reliability (r)
Personal Savings or Employment	.86
Aid From Parents or Family	.88
Repayable Loan	.86
Scholarship/Grant/Other Gift	.85

*For each source, the respondent was asked to check whether it was major (coded 3), minor (coded 2), or not a source (coded 1). Based on Boruch and Creager, 1972.

which dealt more with subjective judgments or opinions were much lower (median for 53 test-retest correlations of .74 in Tables 11 through 15 of Boruch and Creager's report; range .41 to .88). Retrospective reports of high school achievements were highly reliable (range .88 to 1.00).

While Boruch and Creager's results generally support the findings previously reported, several differences are worth noting: the reliability coefficients they obtained were generally higher than others and their retrospective data were about as reliable as their demographic data. The relatively short test-retest interval of two weeks could account for this difference as could other factors (e.g., the ability level of the respondents, or better execution of the study).

In the validity portion of the study, Boruch and Creager gave 4,415 freshmen questionnaires during the fall of 1966 and summer of 1967 (mail questionnaire). The criterion data against which the validity of grades and age was assessed were obtained from college registrars.

The overall correlation between self-reported and registrar-reported grades was .88 and did not vary with the respondent's sex. Validity did however vary with grade level. The reports of students who actually performed at the A/A+ level were most accurate, while students in the B, B-, and C+ strata were least accurate. No major variation in age

Table 3

TEST-RETEST RELIABILITIES FOR REPORTED OBJECTIVE*

Objective†	Reliability (r)
Being Accomplished in a Performing Art	.78
Being an Authority in Field	.73
Obtaining Recognition From Peers	.68
Influencing the Political Structure	.72
Raising a Family	.71
Having an Active Social Life	.67
Having Friends Different From Self	.70
Being an Expert in Finance and Commerce	.74
Having Administrative Responsibility For Work of Others	.66
Being Very Well-off Financially	.81
Helping Others in Difficulty	.65
Becoming a Community Leader	.74
Contributing to a Scientific Theory	.79
Writing Original Works	.80
Not Being Obligated to People	.71
Creating Works of Art	.81
Keeping Up With Political Affairs	.81
Succeeding in Own Business	.67
Developing a Philosophy of Life	.69

* Based on Boruch and Creager, 1972.

† Alternatives and scoring key: essential = 4; very important = 3, somewhat important = 2; not important = 1.

accuracy as a function of age was apparent, but men tended to be more accurate (.91) than women (.85) in reporting age.

Although some differences in validity were associated with race, Boruch and Creager felt that the number of blacks included in their sample (64 men and 86 women) prohibited an accurate estimate of differential responding as a function of grade level. No differences were noted in reporting age as a function of race.

Using data from a national longitudinal survey, Borus and Nestel (1971; 1973) sampled 913 father-son pairs in order to assess the validity of the son's reports of his father's education. A variety of SES and demographic variables were included to determine their relationship to the accuracy of reporting.

Using both univariate and multivariate techniques, Borus and Nestel reached the following conclusions:

1. There are major differences in accuracy as a function of race (68 percent of the white father-son pairs produced identical data compared to 37 percent of black father-son pairs);
2. Father-son pairs from households with 10 or more family members are less likely to be in agreement than father-son pairs from smaller households;
3. Sons who report fathers as "white collar workers" are more likely to be in agreement with their fathers than sons who say their fathers are in service occupations;
4. Males currently in school are more likely to agree with their fathers than males not currently enrolled.

Kerckhoff, Mason, and Poss (1973) investigated the validity of family social status using boys' questionnaire reports of parental educational level and father's occupation. Criterion data were provided by interviews with the parents. The results indicated a moderately high validity level (see Table 4) for reports of parents' education and father's occupational status. The validity of the son's report of father's [illegible]al status did, however, vary wi[illegible]hites provided more valid data than [illegible]cks.

In a similar study on the validity of reports on parental education and occupation, Cohen and Orum (1972) found lower consensus for black parent-child pairs than white parent-child pairs on both occupational *and* educational reports (Table 5).

Differences in the findings of these two studies could, of course, be due to a variety of factors (index of correlation, response categories, inclusion of both males and females in the Cohen and Orum study, and a larger number of subjects in the Cohen and Orum study). In any case, the direction of the results in both studies is consistent: for these variables white parent-child pairs were more in agreement than black parent-child pairs.

In the same study (Cohen and Orum, 1972) validity was also associated with respondent characteristics of child's sex, child's age, and child's school (urban middle-class private school versus inner-city lower

Table 4

VALIDITY INDICES* BY RACE FOR 12th GRADERS' REPORTS OF PARENTAL EDUCATION AND FATHER'S OCCUPATIONAL STATUS

Variable	Blacks (r)	Whites (r)
Mother's Education	.83	.84
Father's Education	.81	.89
Father's Occupation	.74	.93

* Data from Kerckhoff, Mason, and Poss, 1973.

Table 5

GAMMA CORRELATIONS BETWEEN PARENT AND CHILD RESPONSES FOR REPORTS ON PARENTAL EDUCATION AND OCCUPATION*

Variable	Blacks (r)	Whites (r)
Mother's Education	.71	.89
Father's Education	.76	.91
Mother's Occupation	.69	.74
Father's Occupation	.76	.84

* Based on data from Cohen and Orum, 1972.

class predominantly black school versus low [illegible] to working class predominantly white sc[illegible]. Among these groupings there were no c[illegible]ent differences across all four target me[illegible]ures: no group was consistently superior [illegible] inferior although groups did diff[illegible].

Also investigating students' reports of parental SES, Kyaser and Summers (1973) collected student data for each of three successive years and an independent survey of both parents. They concluded that student reports were stable over time with parental education being most reliable (.96 for mothers, .97 for fathers) and father's occupation and income least reliable (.79 for occupation and .56 for income). The validities, corrected for unreliability, were only moderate. Compared to the concurrent mother's report, validity of mother's education was .72, father's education was .83, and father's occupation was .86. Compared to the concurrent father's report, the validity of father's education was .70, father's occupation was .74, and father's income was only .21. Based on the validity findings, Kyaser and Summers concluded that student reports should be utilized with some caution, and, where possible, direct measures of parental SES should be used.

In an investigation of retrospective reporting of own and others' occupational status, Featherman and Hauser (1973) relied on data collected from a variety of sources. Based strictly on adults' reports (age 19 and over) they concluded that retrospective reporting of own occupational status is robust over a five-year period; however, proxy retrospective reports are apparently subject to (memory) decay.

Although Featherman and Hauser (1973) seem guardedly optimistic about the accuracy of retrospective reporting of SES indices, one of the sources on which their conclusions are based (Walsh and Burckholdt, 1970) reached less optimistic conclusions. They concluded:

1. There was a relatively high rate of response error associated with the retrospectively reported data on work status and occupation five years ago. Only 57 percent of the *total* respondents accurately reported both sets of information.
2. A substantial difference was noted between workers and nonworkers in the accuracy of reporting this information. Approximately 70 percent of the nonworkers correctly reported their work status five years ago, whereas only 48 percent of the workers accurately answered the retrospective questions.
3. Errors in reporting work status five years ago was an important factor in the overall level of error. The failure to report having worked five years ago accounted for one-third of the errors made by persons who worked in 1963. This, of course, resulted in a corresponding loss in the data on occupation five years ago.
4. The accuracy rates varied among the major occupation groups. As expected, the accuracy rates were higher for the more skilled occupation groups such as professionals and managers than for the less skilled groups such as farm and nonfarm laborers.
5. Age was a determining factor in the accuracy of reporting this information. Persons in the middle age group—30 to 64 years old—had a significantly higher quality of response than the other two age groups considered (19 to 29 years old and 65 years old and over).
6. The quality of the retrospective data for occupationally mobile persons was substantially lower than that for other workers.
7. The accuracy of the retrospective occupation responses, however, was only about 7 to 9 percentage points lower than the accuracy of reporting current occupation in the 1960 Census. This is due partially to the fact that the nonmobile workers make up the vast majority of respondents and the error rate for nonmobile workers is considerably lower than that for mobile workers.

Keating, Peterson, and Stone (1950) compared interview reports of weekly wages

to employer's wage reports. They found approximately equal validity correlations for males (.90) and females (.93). Duration of employment using the same procedure produced equally high validity correlations for both males and females (.98).

Borus (1966) in a survey of validity studies based on 1950 census data versus other sources concluded that comparisons of census distributions versus other source distributions showed only minor differences in median earnings. In a more direct study of validity, comparing interview data to employer's records, Borus (1966) found a .95 correlation for earnings. A more detailed analysis based on regression techniques showed, however, that the discrepancies between self report and employer's report were related to respondent characteristics. In particular, women compared to men underreported earnings and underreporting varied directly with amount earned, inversely with age, and inversely with education. Borus also found that underreporting decreased as familiarity with the interviewer increased. In a second regression analysis, Borus related response error to characteristics of the respondent's jobs. The additional findings of this analysis were that errors in reporting were greater for new employees and for part-time employees.

Based on these analyses, Borus concluded that response bias exists and is significantly related to respondent characteristics. Thus earnings statistics not only involve sampling error, they also involve response error.

Since the error in Borus's study emerged as bias and not strictly random deviations, the comparison of independent distributions aggregated over different kinds of respondents could obscure this error; yet the majority of studies investigating item formatting and the data collection procedures are based on just this approach (Cannell and Fowler, 1963). Despite the problem of artifactually obscuring differences, Cannell and Fowler note that comparisons of distributions generated by questionnaire versus interview procedures do show differences. The most relevant findings are:

1. There are few differences between the procedures but those that occur are thought to arise from the respondent's desire to present a socially desirable or nonthreatening self-image;
2. When questions are ambiguous, the presence of an interviewer results in fewer nonresponses;
3. People with a high education and income are more likely to cooperate in a mail survey; and
4. People who consult records tend to report more accurately.

Cannell and Fowler also undertook their own study of interview versus questionnaire procedures on hospitalization variables. Using 462 interviews and 465 questionnaires soliciting information on length of stay, month of discharge, diagnosis, type of surgery (if any), and whether or not surgery was performed, they concluded that (1) the self-enumerative procedure is more accurate when records are available (i.e., length of stay and month of discharge); (2) anonymity rather than interviewer presence or absence is the relevant variable in the finding that self-enumerative procedures tend to reduce social desirability; (3) education of the respondent is more important in the interview than in the self-enumerative procedure; (4) motivation is an important factor, particularly in the questionnaire procedure, with education seemingly only important within the "better motivated" group of respondents; and (5) given minimal interest on the part of many respondents, they recommend that questionnaires be self-explanatory and as short as possible.

The above review has been primarily focused on the validity and reliability of factually oriented data; the NLS questionnaires do, however, involve a number of attitude and opinion questions. Chief among these are self-esteem, locus of control and work, and community and family orientations. These scales cannot, of course, be verified in the same way that income, grades, or other factually based data can be verified: there simply are no objective measures of self-esteem, locus of control, etc. To further complicate matters, the majority of attitudinal or psychological scales are novel to NLS and no literature can be surveyed to weave a construct validity net.

The best approximation to the construct validity of these scales is to be found in Conger, Peng, and Dunteman (1976) using precisely these scales on NLS respondents. In this report, the variable of self-esteem based

on Rosenberg's self-esteem scale (Rosenberg, 1965) did not emerge as a potent variable for discriminating among demographic groups; however, the differences which did emerge were generally consistent with previous and anticipated differences.

The locus-of-control scale borrowed from Coleman, Campbell, Hobson, McPartland, Mood, Weinfield, and York (1966), while deviating in content and form from other locus of control scales—for example, Rotter's scale (Rotter, 1966) and the James-Phares' scale (James, 1957)—generally showed similar patterns of between-group discrimination. In particular, whites were more internal than nonwhites, and low SES persons were more external than middle or high SES persons. The strongest relationship was with ability, a finding that, at first glance, was not entirely consistent with the literature; however, as Conger, Peng, and Dunteman noted, this finding actually integrated previous discrepancies in the literature. While no possibility existed for investigating sources of contamination of bias in the self report of locus of control, two major sources of bias have been investigated for other locus-of-control scales. In particular, anxiety is purported to have an unduly high correlation with externality (cf. Butterfield, 1964) but socially desirable responding is thought to play at most a minor role (Lefcourt, 1966) To the extent that the NLS locus-of-control scale is similar to these other scales, one could anticipate that these sources of bias operate in the NLS data in a similar way.

The life goal orientation composites generally showed weak but interpretable relationships. Since no prior literature exists on these scales, and since they had fairly modest or low internal consistency reliabilities, it is difficult to formulate a statement about their construct validity.

Reliability of the NLS self-esteem, locus-of-control, and life goal scales has been obtained by both internal consistency methods and test-retest methods (see Chapter III). Overall, the reliability of those short (three- or four-item) psychological scales is reasonable via either approach for self-esteem ($\alpha = .66$, $r = .66$), and test-retest reliability is generally high for locus of control ($r = .71$) and work ($r = .68$), community ($r = .67$) and family ($r = .68$) life goals. The internal consistencies, however, of all but the self-esteem scale are generally low: locus of control ($\alpha = .56$); work goals ($\alpha = .53$); community goals ($\alpha = .44$); and family goals ($\alpha = .30$). The reliability level, of course, puts constraints on the maximum values for the validity coefficients.

C. Conclusions

In terms of the three sources of variation outlined in the introduction, i.e., item, data collection, and respondent characteristics, the following represent the general conclusions which can be drawn from this review:

1. Variability of responses is very much affected by the "content" of the items. Content in this instance refers to the dimension of objectivity-subjectivity. Demographic characteristics and factual information about present behavior yield the highest validity (and reliability) coefficients, respectively. Factual information on past behavior and evaluative or judgmental behavior yields the least stable data, with the latter representing lowest response stability. Furthermore, validity of reports of past behavior may be moderated by the "importance of the accomplishment." That is, past events which have low ambiguity and are significant to the respondent in terms of accomplishment, e.g., elected student president, tend to have high rates of validity.

   The reliability of report of future events and for goals yields moderate to moderately high coefficients ($.70 \leq r \leq .88$), in most cases. However, as one would expect, these reports are affected by the objectivity-subjectivity dimension of the query. Those future events and/or goals requiring evaluation or judgment may yield coefficients in the .50's and .60's.

2. Variation of response due to data collection mode can produce a significant effect when one is using proxy reports and mixed effects when one is [illegible] to an interview versus a questio[illegible]e method.

Proxy reports by high school age children yield reasonably reliable data when measuring parental education; however, reliability decreases when measuring father's occupation or income. Validities for all SES measures were moderate, but the validity of income reports is particularly low. Proxy reports are further affected by respondent characteristics. Reliability and validity are decreased if the respondent is black, is not enrolled in school or has low educational attainment, or has a blue collar or service worker for a father. These characteristics by no means exhaust the domain of variables which affect validity and reliability of reporting, but do represent some of the major factors. Proxy reports of father's current status on some variables, i.e., education and possibly occupation, may be acceptable for research purposes; however, retrospective proxy reports are in general unacceptable.

The most detailed study of interview versus questionnaire procedures yields mixed results. When the respondent has records to consult, questionnaire procedures produce more accurate results. According to Cannell and Fowler (1963), education of the respondent is more important in the interview procedure than in the questionnaire procedure. Motivation of the respondent comes into play but only within educational groups. On the other hand, motivation is more important in the questionnaire procedure with education becoming a factor within "better motivated" groups of respondents. There is little information on content or respondent characteristics by data mode interactions.

3. Respondent characteristics play an ubiquitous role in affecting the variability of response, perhaps because they encompass such a limitless domain. By and large, white, highly educated, well-to-do, employed respondents from small families produce the most accurate and stable responses. Some studies find males slightly more reliable and accurate than females, while others find no sex differences. There are exceptions to the conclusions above, of course, but as a general rule they hold.

IV. A RELIABILITY STUDY OF NLS DATA

The basic longitudinal questionnaires of the NLS study have been described (above) as broad in scope and diverse in item characteristics. The review of the available literature indicates that the NLS questionnaire items should vary in reliability as a function of item content and, to a lesser degree, as a function of item formatting. In a similar fashion, the diverse characteristics of the NLS respondents can be expected to be associated with differential reliability. In particular, the literature review indicates that reliability varies with respondent's race and sex, with whites and males generally being the more reliable. Only limited information has been collected on reliability as a function of SES and ability variables; however, the results indicate that high SES and high ability respondents are most reliable. Studies on data collection procedures do not strictly favor interview or mail-in approaches; rather an interaction appears to exist between data collection procedures and the respondent's ability and SES level, and between data collection procedures and item content.

It would thus appear that the very factors which make the NLS data base unique and desirable (i.e., heterogeneity of content and respondent characteristics and data collection procedures designed to provide a high return rate) also introduce the possibility of highly differing data quality as a function of item content, respondent characteristics, and data collection procedure. Interactions among these three sources of variability are also likely. Obviously a study of these factors would provide useful information to users of NLS data and to survey research in general.

To date, however, only a small amount of effort has been devoted to investigating the quality of NLS data (Esternacht, 1973 and 1974; and Lyons and Moore, in press).* In the course of instrument development (via field tests on 900 students from the high school class of 1971) decisions about instruments have primarily focused on questionnaire format, the reduction of item redundancy and the development of composites: generally, problems of validity and reliability have not been investigated. Exceptions to this include the determination of composite score internal consistency indices (coefficient alpha) for certain psychological measures (e.g., self-esteem, locus of control, and life goals, Conger, Dunteman, and Peng, 1976), and an investigation of the quality of retrospective data (Lyons and Moore, in press).

Information on reliability is generally desirable and often critical depending on how the data are to be used. One primary purpose of reliability information is to establish subjective confidence in the manifest responses and statistics summarizing these responses. Unreliability inflates variances and proportionally broadens confidence intervals. Unreliability also weakens the power of statistical tests (i.e., the probability of detecting true among-group differences) and attenuates the magnitude of relationships. These are problems of "degree" and not of "kind." However, a number of studies have been done or are comtemplated which use path analyses or structural modeling. The consequences of unknown or incorrectly estimated measurement errors for these models can be profound: conclusions about determination and contribution in path analyses or structural modelings could well be misleading and could be directly opposite to the true relationships if errors of measurement are ignored or are incorrectly specified (cf. Duncan, 1975).

This chapter reports the results of a study designed to provide partial answers to the above issues. This chapter is oriented around the following questions:

1. How reliable are NLS data?
2. How does reliability vary as a function of item characteristics (e.g., subjectiveness, item format, item length, and item content)?
3. How does reliability vary as function of data collection procedures (mail versus interview)?
4. Does reliability vary with respondent characteristics?
5. What interactions exist among data collection procedures, item characteristics, and respondent characteristics?

* See Chapter III for a discussion of these studies.

Answers to these questions will provide useful information on the quality of data in the NLS survey; however, limitations in the design and execution of this study prohibit a comprehensive or definitive conclusion. Consequently, generalizations to all NLS data and to surveys in general can be made only in a tentative fashion. The limitations are: only a subset of items (17) from the Second Follow-Up Questionnaire have been included; only a subset (600) of NLS respondents were targeted as participants; and the limited number of actual participants prohibits detailed comparisons of reliability among subgroups formed by cross-classifying respondent characteristics and respondent characteristics by data collection modes.

A. Sampling Procedures

A probability sample of 600 students was selected for the reliability study. The sampling frame for the reliability study sample consisted of 22,398 individuals who participated in either the base-year or first follow-up surveys. This sampling frame was formed by merging the original base-year student file with a preliminary edition of the first follow-up file. (The final edited version of the merged base-year and first follow-up data file was not available for use when this sample was selected.)

The sampling frame was stratified by sex, race, and planned activity state for October 1974 (item 16 of the First Follow-Up Questionnaire). A "not ascertained" category was included for each stratification variable. The categories used for each of the stratification variables were as follows:

1. *Sex*: Male, Female, Not Ascertained;
2. *Race*: White, Nonwhite, Not Ascertained;
3. *October 1974 Plans*: Four-Year College, Other Postsecondary Education, Work, Other (e.g., active military duty, homemaker), Not Ascertained.

Table 6 shows the numbers of persons in the sampling frame sorted by the cross-classification of the three stratification variables. Of the 45 cells defined by crossing the three variables, two cells were empty and 14 additional cells each contained fewer than 56 persons. Since strata with fewer than 56 individuals would be allocated fewer than two sample persons, these cells were combined with other similar cells to form 29 final strata. First, persons in small cells (i.e., $n < 56$) with sex not ascertained (NA) were combined with females of the same race and activity classification. Then males with race NA and activity state NA were combined with white males with activity state NA. The 29 final strata constructed in this manner are listed in Table 7.

The allocation of the total sample of 600 persons to the 29 final strata was determined in several steps. First, a preliminary allocation in proportion to the stratum counts was calculated as

$$n_i = \frac{600\ N_i}{22{,}239}$$

where n_i = allocation to stratum-i,

and N_i = frame size for stratum-i.

Next, the total sample allocations for certain analysis categories were checked. Specifically, samples of at least 100 persons were desired for each of the following groups: males, females, persons attending four-year colleges, persons attending other postsecondary institutions, and persons working. Closer examination indicated only one of these groups—other postsecondary education—would fall short of 100 sample individuals. To insure a sample size of 100 for this category, the allocations to all strata involving 1974 plans of "other postsecondary education" were oversampled (i.e., multiplied by 1.535) and the allocations to all strata involving 1974 plans of "four-year college" were undersampled (i.e., multiplied by .846). These adjusted allocations are shown in Table 7 in the column headed "adjusted allocation." The final allocations shown as the last column in Table 7 were determined by rounding the adjusted allocations to integer values and by reducing the allocations to the largest strata to force a total sample size of 600 persons.

The last step in drawing the reliability study sample involved the selection of the number of persons in the final allocation column for each stratum from the total

Table 6

NUMBER OF PERSONS IN THE SAMPLING FRAME, BY SEX, RACE, AND 1974 PLANS

Sex and Race	1974 Plans: College	Other Education	Work	Other	NA	Total
Whites:						
Males	3,611	916	2,560	432	402	7,921
Females	3,047	761	3,015	772	250	7,845
NA	8	1	8	3	2	22
Total	6,666	1,678	5,583	1,207	654	15,788
Nonwhites:						
Males	859	351	701	169	175	2,255
Females	973	457	834	144	121	2,532
NA	1	3	2	6	3	9
Total	1,833	811	1,540	313	299	4,796
Race NA:						
Males	360	93	242	58	29	782
Females	298	65	288	73	21	745
NA	7	0	7	3	111	128
Total	665	158	537	134	161	1,655
All Races:						
Males	4,830	1,360	3,503	659	606	10,958
Females	4,318	1,283	4,140	989	392	11,122
NA	16	4	17	6	116	159
Total	9,164	2,647	7,660	1,654	1,114	22,239

number in the stratum sampling frame. These selections were made with equal probabilities and without replacement using random numbers generated by a computer program.

B. Instruments

The questions used for the reliability analyses are a subset of items from the Second Follow-Up questionnaire. This subset was extracted from the Second Follow-Up questionnaire and compiled into a separate "short form" questionnaire (the first set of data for the "test-retest" design were the responses to the selected questions embedded in the Second Follow-Up questionnaire; the second set of data were the responses to the short form questionnaire).

The decision as to how many and which items to include in the reliability study was made primarily on the basis of the following criteria: (1) respondent burden, i.e., the questionnaire should be brief and require no more than 15 minutes to complete; (2) amenability to analysis, i.e., the stability or consistency of the items should be capable of estimation by questioning the same respondents at two close points in time; (3) criticality, i.e., the items should be important or central to the basic NLS analysis; and (4) representativeness, i.e., the items should represent the variety of formats, content areas, and reliance on fact versus subjective opinion.

Based on these criteria, 17 items (Table 8) were selected for inclusion in the reliability

Table 7

DESCRIPTION OF STRATA USED FOR RELIABILITY SAMPLE

Stratum	Sex	Race	1974 Plans	Frame Size	Preliminary Allocation	Adjusted Allocation	Final Sample Allocation
1	Male	White	College	3,611	97.4	82.4	81
2	Male	NA	College	360	9.7	8.2	8
3	Male	White	Other Education	916	24.7	37.9	37
4	Male	NA	Other Education	93	2.5	3.8	4
5	Male	White	Work	2,560	69.1	69.1	68
6	Male	NA	Work	242	6.5	6.5	7
7	Male	White	Other	432	11.7	11.7	12
8	Male	NA	Other	58	1.6	1.6	2
9	Male	White	NA	431	11.6	11.6	12
	Male	NA	NA				
10	Male	Nonwhite	College	859	23.2	19.6	20
11	Male	Nonwhite	Other Education	351	9.5	14.6	15
12	Male	Nonwhite	Work	701	18.9	18.9	19
13	Male	Nonwhite	Other	169	4.6	4.6	5
14	Male	Nonwhite	NA	175	4.7	4.7	5
15	Female	White	College	3,055	82.4	69.7	69
	NA	White	College				
16	Female	NA	College	305	8.2	6.9	7
	NA	NA	College				
17	Female	White	Other Education	762	20.6	31.6	32
	NA	White	Other Education				
18	Female	NA	Other Education	65	1.8	2.8	3
19	Female	White	Work	3,023	81.6	81.6	81
	NA	White	Work				
20	Female	NA	Work	295	8.0	8.0	8
	NA	NA	Work				
21	Female	White	Other	775	20.9	20.9	21
	NA	White	Other				
22	Female	NA	Other	76	2.1	2.1	2
	NA	NA	Other				
23	Female	White	NA	252	6.8	6.8	7
	NA	White	NA				
24	Female	NA	NA	132	3.6	3.6	4
	NA	NA	NA				
25	Female	Nonwhite	College	974	26.3	22.2	22
	NA	Nonwhite	College				
26	Female	Nonwhite	Other Education	460	12.4	19.0	19
	NA	Nonwhite	Other Education				
27	Female	Nonwhite	Work	839	22.6	22.6	23
	NA	Nonwhite	Work				
28	Female	Nonwhite	Other	144	3.9	3.9	4
29	Female	Nonwhite	NA	124	3.3	3.3	3
	NA	Nonwhite	NA				
Total				22,239	600.2	600.2	600
	Subtotal		College	(9,164)	(247.2)	(209.0)	(207)
	Subtotal		Other Education	(2,647)	(71.4)	(109.7)	(110)

Table 8

DESCRIPTION OF ITEMS EXTRACTED FOR RELIABILITY STUDY

Questionnaire Number		
Short Form	Long Form	Description
1	8	Race or Ethnicity
2	10	Educational Activity
3	12	Kind of School Attended
4	39	Grades From Oct. '73 to Oct. '74
5	75	Work Activity in First Week of Oct. '74
6	76	Description of Job Held First Week of Oct. '74
7	77	Average Hours Per Week Worked at This Job
8	105	Marital Status as of First Week of Oct. '74
9	111	Number of Dependents as of First Week of Oct. '74
10	113,114	Income (self, spouse, other, and total)
11	132	Self-Esteem and Locus of Control
12	133	Consumer Behavior
13	139	Expected Activities in Oct. '75
14	140	Educational Expectations
15	141	Factors Important in Determining Life's Work
16	148	Life Goal Orientations (work, community, and family)
17	142	Career Expectations at the Age of 30

study. These 17 items are identical in wording and format to those of the Second Follow-Up questionnaire. (A copy of the Short Form Second Follow-Up questionnaire is provided in Appendix A.)

C. Data Collection Procedures

The data were collected for this study, and for NLS as a whole, through a combination of mail, field interview, and telephone efforts. Data collection activity for the reliability study actually began the second week of October 1974 with the initial mailing of Second Follow-Up questionnaires to all NLS sample members. All incoming Second Follow-up questionnaires completed by mail or by personal interview were event-coded into a computerized Real Time Control System. A computer printout identifying reliability study sample members whose long-form questionnaires had been received was generated on a daily basis.

Short-form questionnaires, with a cover letter (see Appendix A), were then mailed to reliability study members who returned their long-form questionnaires by mail. This event occurred no earlier than ten days after the completion date denoted on the background information page of the Second Follow-Up questionnaire. Two weeks after the mailing of the short-form questionnaire, a prompting telephone call was made to the nonrespondent encouraging him/her to return the questionnaire. If the nonrespondent indicated that he/she had either lost or had never received a short-form questionnaire, or if the nonrespondent could not be contacted for prompting, then a second mailout occurred immediately. No further attempts were made to obtain a response.

Reliability sample members who completed a long-form questionnaire by personal interview were recontacted two weeks after the first interview, at which time an interview with the short-form questionnaire was completed.

Each returned short-form questionnaire underwent an initial editing process to determine whether or not it contained adequate information for acceptance and entry onto the data file. Generally, the editing process required cross-checking a respondent's answers to each of 12 "key" questions on the short form with his or her answers to the same questions on the long form. The decision rule for determining whether or not a key question (and, therefore, the short-form questionnaire) should fail edit may be stated as follows: The fail-edit condition results if and only if a key question was appropriately answered on the long-form questionnaire but was unanswered or inappropriately answered on the corresponding short-form questionnaire.

If a respondent's short-form questionnaire failed edit, a telephone call was initiated and an attempt was made to obtain information for the key items that the respondent failed to answer. Edited short-form questionnaires

and the relevant portions of the corresponding Second Follow-Up (long-form) questionnaire were then coded and keypunched. All data collection activities were completed by 30 April 1975.

D. Data Analysis Procedures

The variety of research questions, data collection procedures, items and item uses (e.g., composites) requires a variety of analytic procedures. First of all, the items have been classified (Table 9) as either categorical or continuous in nature. Reliability estimates for categorical items are based on the percent agreement in responses (including item nonresponses) across the two time points, and the degree of association is additionally summarized by Cramer's V (cf. Statistical Package

Table 9

VARIABLE SPECIFICATION AND DESCRIPTION

Short Form Item Number	Description
Categorical Variables	
1	Ethnic self-descriptions (8) plus missing clustered into 3 categories: white, nonwhite, and missing
3	School types (4) plus missing
5	Work activity categories (3) plus missing
6	Census code job description-analyzed major classifications
6e	Employer types (4) plus missing
8	Marital statuses (4) plus missing
14	Educational (7) expectancies (7) plus missing
17	Career goals (17) plus missing
Continuous Variables	
2	School attendance (dichptomous variable)
4	School performance (seven levels of self-reported grade)
6f	Date of employment: for job held in October 1974: scored as number of months from December 1971
7	Hours worked per week (write in response)
9	Number of dependents: 0, 1, 2, 3, 4
10a	Own income (write in)
10b	Spouse income (write in)
10c	Other income (write in)
10d	Total income (write in)
11	Self-esteem composite* (items a, c, d, and h)
11	Locus of control composite* (items b, e, f, and g)
12 SQ12, LQ	Consumer composite* 1 (items a, b, and d)
12 SQ12, LQ	Consumer composite* 2 (items e and f)
12 SQ12, LQ	Consumer composite 3 (item c)
13	Plans (6)—each of a-f handled as a dichotomous variable (applies or does not apply) with responses assumed missing only if all are blank
15	Work composite* 1 (items a and b)
15	Work composite* 2 (items c, e, f, and i)
15	Work composite* 3 (items g, h, and j)
15	Work composite* 4 (item d)
16	Work orientation* (items a, c, and e)
16	Community orientation* (items f, g, and j)
16	Family orientation* (items b, h, and i)

* Composite scores are computed by averaging available responses. Composites used are based on factor analyses from previous NLS studies.

for the Social Sciences, 1975). Reliability estimates for the continuous and dichotomous variables are provided by product-moment correlations on the available (test-retest) responses.

In order to address the various research questions regarding data collection procedures and respondent characteristics, the following procedures were employed:

1. *Categorical variables*: reliability estimates for subgroups were calculated as percentage agreement and Cramer's V; differences in reliability among subgroups were determined by comparing percent disagreement and agreement among subgroups using a χ^2 procedure. That is, for each respondent on each categorical variable, responses across the two time points were compared and scored as either "disagreement" or "agreement." These scores were then cast into a subgroup by agreement cross-tabulation table and analyzed by a χ^2 statistic. For each variable designed as categorical the following analyses were done:
   a. Data Collection Mode by Agreement,
   b. Sex by Agreement,
   c. Ethnicity by Agreement,
   d. SES by Agreement,
   e. Ability by Agreement,
   f. Sex Within a Data Collection Mode by Agreement,
   g. Ethnicity Within a Data Collection Mode by Agreement,
   h. SES Within a Data Collection Mode by Agreement,
   i. Ability Within a Data Collection Mode by Agreement.
2. *Continuous variables*: reliability estimates for subgroups were computed by product-moment correlations; differences in reliability among subgroups were determined by comparing these correlations using χ^2 tests on Fisher log transformations of the correlations. The correlations were calculated on available test-retest data with no imputation for missing observations.

E. Results

As part of the full-scale Second Follow-Up survey, long-form questionnaires data were obtained from 558 of the 600 designated participants. This return rate of 93 percent is very similar to that obtained for the entire sample. The second questionnaire (short form) was administered to these 558 initial respondents, but only 462 short-form questionnaires (82.8 percent) were completed. The larger than normal attrition from long form to short form obviously introduces some potential bias. In order to further investigate this problem, an analysis was done to see if a differential response rate was associated with demographic characteristics of sex, ethnicity, SES, and ability. These analyses indicated that no major or statistically significant differences in attrition could be associated with these demographic characteristics.

A second preliminary analysis was done on method of data collection. Of the 462 respondents providing data on both the long and short form, 133 (28.8 percent) were interviewed and 329 (71.2 percent) provided data by mail or mail plus telephone solicitation for critical data. These analyses comparing mode of response among classification subgroups defined by sex, ethnicity, SES, and ability showed that nonwhites were more likely to be interviewed than whites (44 versus 22 percent), low SES persons were more likely (36.1 percent) to be interviewed than either middle (24.3 percent) or high (22.0 percent) SES persons, and low ability persons were more likely to be interviewed (39.2 percent) than middle ability persons (24.3 percent) and this group in turn was more likely to be interviewed than high ability persons (11.4 percent). These differences in data collection procedures as a function of ethnicity, SES, and ability could introduce bias into subgroup reliability estimates to the extent that reliability is differentially associated with data collection procedures. For example, if interview data are generally more reliable than mail-in data, nonwhites, low SES, or low ability persons could show higher reliability values than their counterparts. While this

would not necessarily introduce bias into the overall reliability indices for the entire NLS sample, it definitely would limit generalizations to survey research as a whole. As a result of the differences in data collection procedures among subgroups, it was decided that mode of response needed to be controlled for in subgroup analyses. This has the unfortunate consequence of severely limiting sample size for cross-classifications (e.g., sex by ethnicity) among the major classification variables and even among subgroups within a classification variable within a data collection mode (i.e., only eight high-ability persons were interviewed).

Tables 10 and 11 present the total sample reliability indices for the categorical and continuous variables, respectively. The reliability of the items included in this study is generally quite reasonable. Based on the correlation coefficients and Cramer's V coefficient for categorical data, the median reliability is .67. The reliability of some items, however, is quite low (e.g., .36 for other income, .41 for nonacademic educational plans, and .48 for "other" plans). The highest reliability observed for the total sample was .92 for school attendance.

Table 10 also presents percents agreement for the total sample across the two time points for the eight categorical variables. These percents present a more favorable

Table 10

OVERALL RELIABILITY AND RELIABILITY AS A FUNCTION OF DATA COLLECTION PROCEDURES: (CATEGORICAL VARIABLES)

	Total Sample		Data Collection Groups	
	Percent Agreement	Cramer's V	Percent Agreement	
			Mail	Interview
Ethnic self-description	97.2	.67	96.4	99.2
Type of school	94.4	.83	93.3	97.0
Work activity	90.7	.75	89.4	94.0
Job description	88.5	.83	86.9	92.5
Type of employer	87.7	.76	88.4	85.7
Marital status	90.0	.72	90.9	88.0
Educational expectations	70.6	.59	72.0	66.9
Career goals at 30	67.7	.64	62.9*	79.7
N	462	462	329	133

* Difference between subgroup percent agreements is significant at $p < .001$

Table 11

OVERALL RELIABILITY AND RELIABILITY AS A FUNCTION OF DATA COLLECTION PROCEDURES: CONTINUOUS VARIABLES

Description	Total Sample (r)	Data Collection Group	
		Mail (r)	Interview (r)
School attendance	.92 (453)	.92 (320)	.93 (133)
School performance	.81 (211)	.78[a] (161)	.89 (50)
Date of employment	.66 (288)	.64 (198)	.75 (90)
Hours worked/week	.81 (293)	.78[b] (202)	.90 (91)
Number of dependents	.84 (448)	.78[c] (318)	.94 (130)
Own income	.62 (369)	.57[b] (252)	.75 (117)
Spouse income	.67 (228)	.94[c] (132)	.35 (96)
Other income	.36 (221)	.34 (124)	.50 (97)
Total income	.74 (363)	.70[c] (247)	.95 (116)
Self-esteem	.66 (454)	.67 (324)	.60 (130)
Locus of control	.71 (454)	.68 (324)	.73 (130)
Consumer behavior 1	.63 (455)	.58[c] (325)	.75 (130)
Consumer behavior 2	.58 (454)	.51[c] (324)	.72 (130)
Consumer behavior 3	.50 (447)	.50 (319)	.52 (128)
Plans: Working	.77 (459)	.79 (326)	.71 (133)
Plans: Academic education	.85 (459)	.84 (326)	.87 (133)
Plans: Other schooling	.41 (459)	.31[c] (326)	.60 (133)
Plans: Military	.86 (459)	.81[c] (326)	1.00 (133)
Plans: Homemaker	.84 (459)	.82[a] (326)	.89 (133)
Plans: Other	.48 (459)	.43[c] (326)	.66 (133)
Work 1	.56 (449)	.55 (319)	.58 (130)
Work 2	.66 (447)	.65 (317)	.69 (130)
Work 3	.54 (447)	.47[b] (317)	.69 (130)
Work 4	.56 (441)	.51[a] (311)	.65 (130)
Work goals	.68 (457)	.65 (327)	.74 (130)
Community goals	.67 (457)	.65 (327)	.73 (130)
Family goals	.68 (457)	.68 (327)	.68 (130)

Numbers in parentheses are sample sizes. Numbers differ across variables due to respondent nonresponse.

Letters a, b, c, refer to significance levels for between-group comparisons of product-moment correlations a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$ The significance of differences between correlations is based on a χ^2 test on Fisher log transformations of the correlations.

picture than do the coefficients of association.* Four of the categorical variables had percents in the 90's: 97.2 percent for ethnic self-description; 94.4 percent for type of

* This discrepancy results from the inherent characteristics of these two statistics. The percent agreement index is a measure of agreement for the average respondent ignoring categories of response while Cramer's V is related to the agreement of the average response category. Differences between the two statistics indicate that some response categories are likely to be highly unreliable. For example, ethnic self-description had a 92.7 percent agreement index but a .67 coefficient of association. A closer examination of that crosstab table revealed that the major frequency categories of white and nonwhite had high percent agreements but the category "Not Ascertained" had 0 percent agreement. Each approach is legitimate depending on the use; however, percent agreement indices tend to result in higher values. In fact, a high percent agreement could result from totally unreliable data as measured by coefficients of association. For example, a high endorsement rate for a single response category on both testing occasions imposes a high percent agreement overall. If, for example, 99 percent of the respondents claim to have a dictionary in the home, the test-retest percent agreement must be at least 98 percent. But if the overall agreement were 98 percent, a ϕ coefficient calculated on the same data would be -.01.

school attended; 90.7 percent for work status and 90.0 percent for marital status. Educational expectations and career goals at the age of 30 had much lower percents (70.6 and 67.7 percent, respectively).

1. Results: Content and Format

Some variation in the reliabilities can be associated with a dimension of objectivity-subjectivity. Among the categorical variables both indices of reliability (Cramer's V and percent agreement) show that the two most subjective items (educational expectations and career goals) had the lowest reliabilities (Table 10); however, the percent agreement indices provide the sharpest delineation between the subjective and objective items. Among the 27 variables handled as continuous or dichotomous measures there was a larger degree of variation in the reliabilities. The median reliability of the five factually oriented items (Table 11) was higher than the median reliability of the remaining subjective items (.81 versus .67) with only one factual item (hours worked per week) not being superior to the subjective items. Based on these results one can assert that subjective variables are per se generally less reliable than variables with a factual bias. One must also consider that a number of the subjective variables are composites and thus have a higher reliability than the single items making up the components of the composites. Thus, if corrections were made for this factor, the factually oriented items would emerge as even more reliable than the subjective items.

Systematic variations with other item characteristics are not apparent. Seven of the items are open-ended (job description, date of employment, hours worked per week, and the four income items). These items cluster around the median reliability of .67. One item (expected activity in October 1975) was a multiple response item requiring respondents to circle all of the appropriate responses. This item had an overall median reliability of about .80; however, two of the three lowest correlations were associated with components of this item (.41 for noncollege educational plans and .48 for miscellaneous plans). Items embedded in skip patterns (short form items 3, 4, 6, and 7) which were all factual in nature were neither better nor worse than factual items not so embedded.

Three of the items (ethnic self-description, kind of school attended, plans for October 1975) included a waste basket response category termed "other." The percent agreement across the two time-points was quite low in each case. The variable "other income" also manifested a low test-retest correlation. Taken in conjunction, these results would tend to indicate that the use of catch-all or miscellaneous response categories does not provide reliable data.

2. Results: Data Collection Procedures

Variation in reliability was also associated with the mode of data collection, mail-in or interview. Fifteen of the thirty-five variables investigated showed significant and often substantial differences in reliability as a function of data collection mode (Tables 10 and 11). One difference (spouse's income) strongly favored the mail-in procedure, but this difference should not be interpreted without reference to the sex by mode interaction discussed below. The remaining fourteen differences favored the interview procedure. These differences cut across the item characteristic differences previously described, but most of them were also involved in interaction with other respondent characteristics. Despite these interactions, however, it seems fairly safe to conclude that interview-obtained data is more reliable than mail-in data. The previously described relationship between respondent characteristics and probability of being interviewed would indicate that substantial differences in data quality probably exist across individuals both as a function of mode of data collection (a direct effect) and as a function of respondent characteristics (an indirect effect).

3. Results: Variations Associated with Respondent Characteristics

Reliability variations as a function of respondent characteristics will be discussed separately for the categorical variables and for the continuous variables.

Among the categorical variables there were no differences associated with ethnicity or SES and small differences associated with sex and ability or their interactions with data collection procedures.

Table 12 shows the percent agreement indices for males and females overall and within data collection modes. Overall, the comparisons indicate that females are more reliable than males when there is a difference; however, these differences emerge primarily in the mail-in data collection procedure. Four of the five differences show that females are more reliable than males for the mail mode with no differences for the interview mode.

Among the three ability groups there was only one overall difference (marital status, Table 13), but this is qualified by an ability by mode interaction. There were three significant differences among the ability groups for the mail-in questionnaires and no significant differences among groups for interview-collected data. All of the significant differences were of the same type: low ability persons were less reliable than either middle or high ability persons. The latter two groups were generally comparable.

For continuous variables, a quick glimpse at Tables 14 through 17 indicates that a large number of differences exist among groups differentiated on the basis of sex, ethnicity, ability, and SES, and their interaction with data collection procedures. Only 4 of the 27 items classified as continuous variables showed no difference among any groups (consumer composite 3, plans for education other than college, and community and family life goals); the first two of these, however, had differences in reliability as a function of data collection procedures.

Altogether, 16 of the 27 variables (Table 14) handled as continuous or dichotomous data produced sex by data collection mode interactions and an additional 2 items only had an overall sex difference. Twelve items, however, manifested a difference in reliability between males and females when the data collection procedure is ignored. The items for which there was no apparent interaction are: own income (males substantially more reliable than females), self-esteem, and locus of control (females more reliable than males). For the remaining items involving interactions, differences occurred between males and females within both data collection procedures. Within the mail mode four of the differences in reliability favored males and four favored females. Males more reliably reported total expected income, plans for college, and military and homemaking activities for the following year. Females more reliably reported whether or not they were attending school, school performance the past year, date of employment for their current job, and plans for nonacademic schooling in the subsequent year.

Table 12

TEST-RETEST PERCENT AGREEMENT FOR SEX GROUPS: CATEGORICAL VARIABLES

Description	Mail Mode *Sex Groups*		Interview Mode *Sex Groups*		All Respondents *Sex Groups*	
	Males	Females	Males	Females	Males	Females
Type of school	88.9[b]	97.2	97.1	96.9	91.4[a]	97.1
Work activity	84.3[b]	93.8	94.1	93.8	87.3	93.8
Job description	81.0[b]	92.0	92.6	92.3	84.6[a]	92.1
Type of employer	—	—	—	—	84.2[a]	90.9
Educational expectations	65.4[a]	77.8	70.6	63.1	67.0	73.9
N	153	176	68	65	221	241

Letters a, b, c, refer to significance levels for between-group comparisons of percent agreement using a χ^2 statistic. a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$.

Table 13

TEST-RETEST PERCENT AGREEMENT FOR ABILITY GROUPS: CATEGORICAL VARIABLES

Description	Mail Mode Ability Groups			Interview Mode Ability Groups			All Respondents Ability Groups		
	Low	Middle	High	Low	Middle	High	Low	Middle	High
Job description	78.0[a]	93.0	83.9	94.7	87.8	100.0	84.5	91.7	85.7
Type of employer	81.4[a]	92.2	93.5	89.5	80.5	75.0	84.5	89.3	91.4
Marital Status	79.7[b]	92.2	95.2	94.7	95.1	100.0	85.6[a]	92.9	95.7
N	59	128	62	38	41	8	97	169	70

Letters a, b, c, refer to significance levels for between-group comparisons of percent agreement using a χ^2 statistic: a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$.

Within the interview mode, five differences favored males and three favored females. Males more reliably reported date of employment (reversing the relationship observed in the mail mode), number of dependents, spouse's income, miscellaneous plans, and were more reliable for consumer composite 2. Females were more reliable on hours worked per week, other income, and work composite 4.

These results cannot be easily interpreted, but some may be dismissed. In particular, little meaning should be associated with differences between males and females on plans for the military or plans for homemaking activities. In the former case, about one percent of the females indicated a likely military involvement and in the latter around one percent of the males indicated a likely role as a homemaker.

The difference in reliability between males and females on spouse's income in the interview procedures is, however, interpretable. Females were generally less reliable than males in reporting their own expected income and total expected income, perhaps indicating that they are not as attuned to financial matters as males. Consequently the highly unreliable report of spouse's income in the interview mode ($r = .28$) relative to the mail mode ($r = .91$) could indicate that in the mail mode, the females requested this information from their spouses (who were not available in the interview mode). One must also consider, however, that females in the interview mode very reliably reported other income—the only group across modes to do so.

In general, while a large number of reliability differences between males and females do exist, they do not lend themselves to a comprehensive interpretation, nor do they consistently favor one group over another.

Numerous differences (Table 15) in reliability occurred between whites and nonwhites. Nine differences (school attendance, number of dependents, own income, spouse's income, locus of control, military and homemaking plans, and work composites 1 and 3) were interactive in nature (i.e., occurred only under one data collection procedure) and four differences occurred independent of data collection mode (date of employment, other income, total income, and self-esteem). Of the 20 significant differences in Table 15, 14 favored nonwhites over whites, a finding not indicated by prior research on the reliability or validity of survey data (see review chapter). Five of the six differences for which whites were more reliable than nonwhites occurred on subjectively oriented data (self-esteem, locus of control, and work composites 1 and 3). By comparison, only 3 of the 14 differences favoring nonwhites were subjective in nature (military and homemaking plans).

Of particular interest are the six white-nonwhite differences in reliability for income.

Table 14

RELIABILITY FOR SEX GROUPS: CONTINUOUS VARIABLES

Description	Mail Mode *Sex Groups*		Interview Mode *Sex Groups*		All Respondents *Sex Groups*	
	Males (r)	Females (r)	Males (r)	Females (r)	Males (r)	Females (r)
School attendance	.89[c] (149)	.95 (171)	.93 (68)	.91 (65)	.90[b] (217)	.94 (236)
School performance	.68[b] (79)	.86 (82)	.92 (28)	.86 (22)	.75[a] (107)	.86 (104)
Date of employment	.51[c] (99)	.87 (99)	.94[c] (50)	.67 (40)	.61[a] (149)	.75 (139)
Hours worked/week	.79 (100)	.76 (102)	.86[c] (51)	.99 (40)	.81 (151)	.81 (142)
No. of dependents	.78 (148)	.76 (170)	[illegible] (66)	.85 (64)	.85[a] (214)	.78 (234)
Own income	.79[c] (119)	.42 (133)	.95[c] (62)	.56 (55)	.85[c] (181)	.46 (188)
Spouse income	.91 (49)	.93 (83)	.99[c] (51)	.28 (45)	.95[c] (100)	.62 (128)
Other income	.49 (52)	.30 (72)	.30[c] (53)	.91 (44)	.35 (105)	.51 (116)
Total income	.89[c] (114)	.64 (133)	.94 (62)	.96 (54)	.91[c] (176)	.67 (187)
Self-esteem	.60[b] (151)	.73 (173)	.41[a] (66)	.70 (64)	.56[b] (217)	.72 (237)
Locus of control	—	—	—	—	.66[a] (217)	.75 (237)
Consumer behavior 2	.48 (151)	.54 (173)	.83[b] (66)	.60 (64)	.62 (217)	.56 (237)
Plans: Working	.72[b] (151)	.85 (175)	.70 (68)	.71 (65)	.72[b] (219)	.82 (240)
Plans: Academic education	.89[b] (151)	.80 (175)	.87 (68)	.88 (65)	.88[b] (219)	.82 (240)
Plans: Military	[illegible][c] [illegible]	.57 (175)	1.00 (68)	1.00 (65)	.88[c] (219)	.70 (240)
Plans: Homemaking	1.00[c] (151)	.78 (175)	x (68)	.88 (65)	.82 (219)	.81 (240)
Plans: Other	.38 (151)	.48 (175)	.81[b] (68)	.57 (65)	.48 (219)	.49 (240)
Work composite 4	.55 (147)	.48 (164)	.55[a] (66)	.76 (64)	.55 (213)	.56 (228)

Numbers in parentheses indicate sample sizes.

Letters a, b, c, refer to significance levels for between-group comparisons of correlations: a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$. The significance of differences between correlations is based on a χ^2 test on Fisher log transformations of the correlations.

Table 15

RELIABILITY FOR ETHNIC GROUPS: CONTINUOUS VARIABLES

Description	Mail Mode Ethnic Groups		Interview Mode Ethnic Groups		All Respondents Ethnic Groups	
	White (r)	Nonwhite (r)	White (r)	Nonwhite (r)	White (r)	Nonwhite (r)
School attendance	.92 (266)	.92 (54)	.95[b] (84)	.86 (49)	.93 (350)	.90 (103)
Date of employment	.61[b] (164)	.84 (34)	.70[c] (62)	.98 (28)	.61[c] (226)	.89 (62)
No. of independents	.71[c] (267)	.90 (51)	.95 (82)	.93 (48)	.80[c] (349)	.92 (99)
Own income	.60 (219)	.41 (33)	.68[c] (76)	.93 (41)	.62 (295)	.70 (74)
Spouse income	.94 (116)	.90 (16)	.28[c] (60)	.995 (36)	.65[c] (176)	.95 (52)
Other income	.33 (108)	.68 (16)	.29[c] (59)	.86 (38)	.30[c] (167)	.80 (54)
Total income	—	—	—	—	.72[c] (288)	.88 (75)
Self-esteem	—	—	—	—	.70[b] (350)	.52 (104)
Locus of control	.67 (268)	.69 (56)	.81[b] (82)	.54 (48)	.71 (350)	.64 (104)
Plans: Military	.76[c] (270)	.92 (56)	1.00 (84)	1.00 (49)	.80[c] (354)	.95 (105)
Plans: Homemaking	.83 (270)	.77 (56)	.86[b] (84)	.94 (49)	.84 (354)	.87 (105)
Work composite 1	.60[b] (265)	.24 (54)	.64 (82)	.46 (48)	.61[b] (347)	.37 (102)
Work composite 3	.47 (262)	.46 (55)	.73[a] (82)	.46 (48)	.56 (344)	.48 (103)

Numbers in parentheses indicate sample sizes.

Letters a, b, c, refer to significance levels for between-group comparisons of correlations: a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$. The significance of differences between correlations is based on a χ^2 test on Fisher log transformations of the correlations.

The greater reliability for nonwhites stands in direct contradiction to other studies and is internally contradictory with ability group results (below) but supported by the observed SES differences (below).

Some difference in reliability among the three ability groups occurred on 18 of the 27 continuous variables (Table 16); however, there were 36 significant differences overall. Of these, the lowest ability group was the least reliable 30 times, and the highest ability group was the least reliable only once (working plans). This large number of differences occurred despite a substantially smaller

Table 16

RELIABILITY FOR ABILITY GROUPS: CONTINUOUS VARIABLES

Description	Mail Mode Ability Groups			Interview Mode Ability Groups			All Respondents Ability Groups		
	Low (r)	Middle (r)	High (r)	Low (r)	Middle (r)	High (r)	Low (r)	Middle (r)	High (r)
School attendance	.90	.92	.93	.82[c]	1.00	1.00	.88[a]	.94	.94
	(56)	(125)	(61)	(38)	(41)	(8)	(94)	(166)	(69)
School performance	.55[a]	.71	.86	-.17[a]	.90	.88	.49[a]	.78	.86
	(18)	(56)	(49)	(6)	(19)	(7)	(24)	(75)	(56)
Date of employment	.69[c]	.47	.97	.998[c]	.90	.92	.81[c]	.53	.96
	(32)	(81)	(35)	(22)	(29)	(6)	(54)	(110)	(41)
Hours worked/week	.51[a]	.83	.73	.71[c]	.84	1.00	.56[b]	.83	.79
	(33)	(81)	(37)	(22)	(31)	(6)	(55)	(112)	(43)
No. of dependents	.42[c]	.77	.92	.93[c]	.90	1.00	.70[c]	.80	.95
	(54)	(126)	(62)	(37)	(41)	(8)	(91)	(167)	(70)
Own income	.32[c]	.91	.92	.58[c]	.98	.98	.41[c]	.93	.92
	(37)	(102)	(57)	(32)	(40)	(7)	(69)	(142)	(64)
Spouse income	.83[c]	.98	.91	.16[c]	1.00	1.00	.24[c]	.98	.91
	(19)	(51)	(34)	(28)	(30)	(5)	(47)	(81)	(39)
Other income	.66	.53	.38	.86[c]	.26	1.00	.68	.35	.42
	(17)	(43)	(36)	(26)	(29)	(5)	(43)	(72)	(41)
Total income	.29[c]	.90	.96	.95	.97	.99	.38[c]	.91	.96
	(36)	(100)	(53)	(32)	(38)	(7)	(68)	(138)	(60)
Self-esteem	.46[a]	.64	.77	.70	.66	.50	.56	.64	.75
	(56)	(128)	(62)	(37)	(41)	(8)	(93)	(169)	(70)
Consumer behavior 2	.25[b]	.52	.70	.72	.64	.56	.46	.54	.68
	(56)	(127)	(62)	(37)	(41)	(8)	(93)	(168)	(70)
Plans: Working	.75[a]	.87	.73	.55	.74	.74	.67[b]	.84	.73
	(57)	(128)	(62)	(38)	(41)	(8)	(95)	(169)	(70)
Plans: Academic education	.89	.80	.80	.82[c]	.84	1.00	.86	.81	.83
	(57)	(128)	(62)	(38)	(41)	(8)	(95)	(169)	(70)
Plans: Military	.80[c]	.84	1.00	1.00	1.00	1.00	.84[c]	.88	1.00
	(57)	(128)	(62)	(38)	(41)	(8)	(95)	(169)	(70)
Plans: Homemaking	.48[c]	.89	.85	.93	.94	.66	.70[c]	.90	.82
	(57)	(128)	(62)	(38)	(41)	(8)	(95)	(169)	(70)
Plans: Other	.57[a]	.35	.65	.47	x	x	.50[a]	.33	.65
	(57)	(128)	(62)	(38)	(41)	(8)	(95)	(169)	(70)
Work composite 1	—	—	—	—	—	—	.35[a]	.59	.58
	—	—	—	—	—	—	(94)	(163)	(70)
Work composite 2	.49[a]	.68	.77	.77	.53	.37	.61	.65	.74
	(55)	(123)	(62)	(37)	(41)	(8)	(92)	(164)	(70)

Numbers in parentheses indicate sample sizes.

Letters a, b, c, refer to significance levels for between-group comparisons of correlations: a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$. The significance of differences between correlations is based on a χ^2 test on Fisher log transformations of the correlations.

x indicates 0 variance on one or both questionnaires; thus no correlation was computed.

Table 17

RELIABILITY FOR SES GROUPS: CONTINUOUS VARIABLES

Description	Mail Mode SES Groups			Interview Mode SES Groups			All Respondents SES Groups		
	Low (r)	Middle (r)	High (r)	Low (r)	Middle (r)	High (r)	Low (r)	Middle (r)	High (r)
Date of employment	.85[c]	.51	.84	.98[c]	.94	.57	[illegible][c]	[illegible]	.63
	(63)	(102)	(33)	(31)	(43)	(16)	(94)	(145)	(49)
Hours worked/week	.80	.7[illegible]	.80	.93[c]	.83	.99	.84	.79	.85
	(64)	(105)	(33)	(30)	(45)	(16)	(94)	(150)	(49)
No. of dependents	.79[c]	.76	.92	.86[c]	.99	.84	.82	.84	.89
	(90)	(159)	(69)	(49)	(59)	(20)	(139)	(218)	(89)
Own income	.24[c]	.91	.97	.97[c]	.62	.91	.40[c]	.77	.96
	(76)	(122)	(54)	(45)	(52)	(20)	(121)	(174)	(74)
Spouse income	.89	.95	.96	.96[c]	.25	1.00	.92[c]	.58	.97
	(36)	(70)	(26)	(36)	(43)	(16)	(72)	(113)	(42)
Other income	.68[a]	.32	.29	.96[c]	.30	.26	.84[c]	.32	.24
	(36)	(60)	(28)	(38)	(39)	(19)	(74)	(99)	(47)
Total income	.35[c]	.93	.94	.98[b]	.94	.91	.46[c]	.94	.94
	(74)	(1[illegible])	(56)	(43)	(54)	(19)	(117)	(171)	(75)
Locus of control	.80[b]	.65	.54	.70	.77	.70	.77[a]	.70	.58
	(96)	(159)	(69)	(49)	(59)	(20)	(145)	(218)	(89)
Consumer behavior 1	.61	.56	.54	.74[a]	.68	.92	.66	.59	.66
	(95)	(160)	(70)	(40)	(59)	(20)	(144)	(219)	(90)
Plans: Academic education	.83	.83	.82	.85[c]	.81	1.00	.84	.82	.86
	(96)	(160)	(70)	(50)	(60)	(21)	(146)	(220)	(91)
Plans: Military	.84[b]	.72	.89	1.00	1.00	x	.88	.84	.89
	(96)	(160)	(70)	(50)	(60)	(21)	(146)	(220)	(91)
Plans: Homemaking	.85	.82	.80	.86[b]	.95	.80	.85	.85	.80
	(96)	(160)	(70)	(50)	(60)	(21)	(146)	(220)	(91)
Plans: Other	.22[b]	.46	.65	.81[b]	.48	x	.44[a]	.45	.65
	(96)	(160)	(70)	(50)	(60)	(21)	(146)	(220)	(91)
Work composite 1	.42[a]	.56	.72	.57	.50	.79	.48[b]	.53	.74
	(91)	(159)	(69)	(49)	(59)	(20)	(140)	(218)	(89)
Work composite 3	.44	.44	.55	.82[a]	.62	.47	.58	.50	.54
	(90)	(158)	(69)	(49)	(59)	(20)	(139)	(217)	(89)
Work goals	.59	.64	.73	.88[b]	.59	.77	.70	.63	.74
	(97)	(161)	(69)	(49)	(59)	(20)	(146)	(220)	(89)

Numbers in parentheses indicate sample sizes.

Letters a, b, c, refer to significance levels for between-group comparisons of correlations: a indicates $p < .05$, b indicates $p < .01$, and c indicates $p < .001$. The significance of differences between correlations is based on a χ^2 test on Fisher log transformations of the correlations.

x indicates 0 variance on one or both questionnaires; thus no correlation was computed.

sample size due to the lack of ability data on 126 subjects, and, in particular, significant differences can be noted among ability groups in the interview response mode for as few as 32 respondents (e.g., school performance). Although there are a few exceptions to the general trend of reliability being positively associated with ability, it seems reasonable to conclude that low ability persons do not provide data as reliable as that of middle or high ability persons. This trend exists across item characteristics and data collection procedures.

Among the more critical differences are the low reliabilities (for low ability persons) for school performance (r = .49), hours worked per week (r = .56), and own income, spouse's income, and total income (r's of .41, .24, and .38). Taking the data collection procedure into account, mail-in questionnaire data for low ability persons is particularly unreliable for number of dependents (r = .42), own income (r = .32), total income (r = .29), and consumer composite 2 (r = .25). In the interview mode, low ability persons were unreliable on school performance (r = -.17) and spouse's income (r = .16).

Table 17 provides data on SES differences in reliability. Like the preceding results on sex, ethnicity, and ability, there are many differences which do not form a readily interpretable picture. Most of the differences are interactive with data collection procedures—sometimes with striking results (e.g., on own income, low SES persons are the least reliable in the mail mode, r = .24, but the most reliable in the interview mode, r = .97). Table 17 indicates a total of 30 significant differences: 8 overall, 9 in the mail-in mode, and 13 in the interview mode. Only two of the eight overall differences (other income and work composite 1) are apparently independent of the data collection procedure. Generally, the least reliable group was the middle SES group.

Reliabilities below .50 are numerous and occur frequently on income data. Own income and total income are quite unreliable for the low SES group overall and in the mail mode. These results may of course be related to the previously mentioned ability differences in reliability.

F. Discussion

The vast array of differences in reliability associated with item and respondent characteristics, data collection procedures, and interactions among these factors prohibits any singular conclusion about the reliability of NLS data. While the average item (of those analyzed) is reasonably reliable (median reliability = .67) as measured by correlation coefficients for continuous or dichotomous variables and by Cramer's V for categorical variables, the range is fairly broad. The highest total sample reliability was .92 (school attendance), the only reliability coefficient in the .90's. By contrast, three items had reliability indices below .50: .36 for other income, .41 for educational plans other than college and .48 for "other" miscellaneous plans. When respondent characteristics and data collection procedures are considered the range is much greater: various subgroups had reliabilities of 1.00 but there were many reliabilities below .50 (the lowest observed test-retest correlation was -.17 for low ability respondents on school performance collected by interview).

Based strictly on the total sample, factually oriented items were substantially more reliable than items dealing with expectations and self-evaluations. This result is highly consistent with previous studies of survey questionnaire reliability (see review chapter) and stands as the best substantiated conclusion of this study.

1. Reliability as a Function of Data Collection Procedure

Fairly clear differences also exist between the reliability of interview and mail-in responses. Interview data were, with only one exception, as reliable or more reliable than mail-in data. The single exception was for spouse's income; however, this exception could well have been due to females who were not well informed about their husbands' earnings but who could seek out accurate (or consistent) results from their husbands for the mail-in procedure. Also, numerous interactions with respondent characteristics qualify the main results of interviews being more reliable than mail-solicited data. Despite these interactions, it seems safe to conclude that

the interview procedure *generally* produces more reliable data than mail-in procedure.

2. Reliability as a Function of Respondent Characteristics

Reliability also varied with respondent characteristics; however, many of the differences in reliability associated with respondent characteristics were qualified by interactions of respondent characteristics with data collection procedures. Since many of the interactions between respondent characteristics and data collection procedures were not consistent across items, it seems necessary to conclude that there is at least a three-way interaction (respondent characteristics by data collection made by item content).

Differences in reliability between males and females exist but neither group was consistently more reliable than the other. Males were more reliable than females for items involving numerical judgments (e.g., income); otherwise females generally were more reliable than males. When males and females were compared within a data collection procedure, differences were more frequent but did not consistently favor either group.

The ethnic group comparisons showed a tendency for nonwhites to be slightly more reliable than whites. The items favoring nonwhites involved factually oriented data (e.g., anticipated income, number of dependents, date of employment) while those favoring whites were more subjective in nature (e.g., self-esteem, work factors). There was no tendency for the pattern of differences to be associated with data collection procedures. While it is comforting that there was no consistent bias in reliability associated with ethnic/racial groupings, this finding does run counter to previous research and deserves further scrutiny. While test-retest questionnaire nonresponse was not associated at conventional statistical levels ($p < .05$) with any comparison among demographic groups, there was a tendency for whites to have a higher return rate than nonwhites. If this trend has been operating since base year and if it also occurs at an item nonresponse level, the current findings could be attributed to different questionnaire-taking behaviors. That is, if unreliable nonwhite respondents tend to drop out at a greater rate than reliable nonwhite respondents and whites in general, the current results could be artifactual. Only an evaluation of questionnaire and item nonresponse would provide the necessary data to resolve this.

The vast majority of items showed variation in reliability as a function of ability. While there were a few minor exceptions (e.g., date of employment and other income in the interview mode), it seems safe to conclude that low ability persons provide less reliable data than middle or high ability respondents. Generally, the high ability respondents were more reliable.

A large number of differences also occurred among SES groups, but unlike the ability group differences the pattern did not consistently favor any single SES level. Some items (e.g., date of employment) favored the lowest SES groups, others (e.g., number of dependents-interview mode) the middle SES group, but most favored the highest SES group. While there are numerous exceptions, the overall trend was for the highest SES group to be most reliable and for the middle SES group to be least reliable.

It should be apparent that the association of reliability with respondent characteristics forms a complex and almost paradoxical pattern. The general trend is of the least reliable respondents being low ability, middle SES, and white and the most reliable being high ability, high SES, and nonwhite. The paradoxical nature of these results is to be found in the fact that nonwhites have lower ability scores and SES indices than whites. Thus one would expect that if high ability and high SES persons were generally most reliable, so too would whites be more reliable than nonwhites. Unfortunately, the already small sample sizes prohibit any meaningful comparison at a cross-classification level (e.g., ethnicity by ability by SES), particularly when controlling for mode of data collection.

3. Integration of Results

On balance, it would appear that complex multiway interactions are operating among the demographic factors. These interactions are further complicated by interactions with data collection procedures and item characteristics (particularly the objectivity-subjectivity dimension). Despite the existence of these

interactions, there are fairly strong main effects, most of which are supported by previous research. These findings in order of strength of substantiation are:

a. Factually oriented items are more reliable than subjectively oriented items;

b. Interview collected data is more reliable than mail-in data;

c. Low ability respondents are less reliable than middle or high ability respondents;

d. Middle SES respondents are less reliable than low or high SES respondents;

e. Females are more reliable than males on nonquantitative items;

f. Nonwhites are more reliable than whites;

g. Response categories with an ambiguous referent (i.e., "other") are generally unreliable.

4. Interpretation and Implications

The above conclusions are, of course, qualified by the frequent interactions and they need to be interpreted with caution due to the unknown effects of instrument and item nonresponse biases. Generalization of these results to the entire NLS data base also needs to be done with care. While the reliability of the average item included in the study was respectable (.67), there was a lot of variation. This level of reliability clearly indicates that the item or composite data are not totally reliable and for some purposes they are not sufficiently reliable. For example, the overall level of reliability is not sufficiently high for path analyses even using a liberal assumption that if the reliability is in the .90's analytic work may proceed. Similarly, construct interpretations of correlations and regression analyses and comparisons of effects among various multiple classification groups all need to be done with caution. The failure to obtain significant relationships between dependent variables and independent variables could be due to an actual absence of a relationship or to poor data quality.

The existence of numerous and perhaps complex interactions among demographic groups alone or with data collection procedures further complicates interpretation. For many data analyses, only a subset of the NLS sample is used. For example, the investigation of work activity and attitudes would generally involve more unreliable respondents than would an investigation of postsecondary education and related factors. Thus, even if a researcher used only those variables included in this study, adequate estimates of reliability for structural modeling might not be available if the subsample did not correspond to one of the demographic subgroups included in this study.

Generalization to the entire NLS sample is also complicated by the higher than usual attrition rate for the short form. There could be a further confound also if the relationship between demographic groups and data collection procedures differs for this sample versus the entire NLS sample. Clearly, generalizations based solely on this study to the entire NLS data base can only be provisionally and cautiously offered.

V. IMPLICATIONS AND CONCLUSIONS

The review of available literature on the reliability of NLS type questions, respondents, and data collection procedures generated findings which are generally consistent with those obtained in the reliability study. One of the goals of this chapter is to integrate the findings of the reliability study with the existing literature so that the quality of NLS data, and survey research data in general, can be better understood. Since a validity study was not carried out, a comparison of validity considerations cannot be similarly offered. Consequently, generalizations about validity can only be made from previous research. The validity results generally are in agreement with the findings on reliability; however, some discrepancies do exist and require resolution.

A. Reliability and Validity as a Function of Item Characteristics

The literature and reliability study are unequivocally consistent in the finding that contemporaneous, objective, factually oriented items are more reliable than subjective, temporally remote, or ambiguous items. In particular, the reliability study results indicated that the reliability of contemporaneous factually based items was in the range (.67 - .92) and the remaining items were in the range (.36 - 86) While the level of reliability was perhaps lower than that obtained by van Es and Wilkening (1970), Boruch and Creager (1972), or Kyaser and Summers (1973), the consistency of the findings is unambiguous.

Boruch and Creager (1972) and van Es and Wilkening (1970) also found that items with a future or retrospective orientation were less reliable than contemporaneous items. The reliability study indicated similar findings, although a comparison on this basis was limited. Other dimensions of item characteristics generally did not produce consistent differences, but the limited sampling of items in the reliability study could be responsible. From the literature, however, even factually oriented items differ in reliability depending on the importance (Astin, 1965), with personally important items being more reliable. The well detailed tables of Boruch and Creager (1972) also indicate that personally sensitive items (e.g., income) are less reliable than other factually oriented items. In the reliability study the income items were not strictly factual since they required earnings for the current year; however, these items were generally of low reliability.

The validity results are similarly consistent. Both Walsh (1967 and 1968) and Kyaser and Summers (1973) found that highly factually oriented items were more valid than less factually oriented items. Of particular interest are the Kyaser and Summers results on income data; even corrected for attenuation due to unreliability, proxy reports of income were highly invalid.

The attitudinal and psychological variables included in the sample were typically of moderate reliability. Most of these variables, either as items or composites, have no prior history; hence their construct or predictive validity is unknown. Although the results of the study by Conger, Peng, and Dunteman (1976) indicated a reasonable pattern of relationships for the self-esteem, locus of control, and work, community, and family life-goal orientations, the relationships were generally weak. Reviews of the validity of even highly refined personality and psychological measures produce a similar result (cf. Fiske, 1974; or Wiggins, 1972).

Based on the differences in reliability and accuracy or construct validity as a function of item content, research involving combinations of content needs to be done cautiously. Path analyses or structural models with incorrectly assumed reliability levels would generally produce erroneous conclusions about the relative importance of these different kinds of variables. Even simple regression studies involving statements about constructs would a priori be more likely to conclude that factually oriented constructs are more important than subjectively based constructs; however, one could probably trust results which indicated the opposite pattern (e.g., Coleman et al. (1966), who found locus of control to be the most important predictor of college achievement among minority persons).

Research focusing on strictly empirical relationships would, however, not be subject to the same problems.

B. Reliability and Validity as a Function of Data Collection Procedures

Data collection procedures have not been exhaustively studied as they bear on validity and reliability. The prime studies by Walsh (1967, 1968), and Cannell and Fowler (1963) and the review by Borus (1970) indicate that neither approach is consistently superior. The Borus review and the Cannell and Fowler study do, however, indicate some important interactions. In particular, high ability and high SES persons are less influenced by data collection procedures than low ability or low SES persons. The latter groups are more cooperative and produce more accurate or valid data in the interview procedure.

Both researchers also found a content by data collection procedure interaction. That is, questions which could be answered by consulting records are more accurately answered in a mail-in procedure.

The results of the reliability study indicated that the interview procedure generally produced more reliable results with one major exception: the reporting of the spouse's income. On this variable, however, the discrepancy was resolved by noting that the lower reliability in the interview procedure could be attributed to women. In the mail-in approach, males and females were equally reliable and were not different from males who were interviewed. Apparently males need not consult records in reporting the wife's income but the females may have consulted their husbands or records in reporting the husband's income.

This mixed state of affairs means that there is no singularly best procedure for collecting reliable and valid survey data. Interviews would lose information on factual archival data and mail-in questionnaires would lose information on other variables. Other considerations, such as response rates and cost should perhaps take precedence, particularly if the respondents are heterogeneous on ability and SES.

C. Reliability and Validity as a Function of Respondent Characteristics

Few of the studies reviewed looked directly at reliability as a function of respondent characteristics; however, most of the studies on validity did consider this as a factor. The validity studies indicated that race is an important variable, with blacks providing less accurate information than whites (Borus and Nestel, 1971; Kerckhoff, Mason, and Poss, 1973; and Cohen and Orum, 1972). Also, high SES respondents were found to be more accurate than low SES respondents (Borus and Nestel, 1971; Cohen and Orum, 1972; and Walsh and Burkhold, 1970).

Boruch and Creager (1972) found no male-female differences in validity and Cohen and Orum (1972) found differences, but neither group was consistently superior. Ability was not directly investigated; however, persons with high obtained grades or higher educational levels produced more valid data than their counterparts (Boruch and Creager, 1972; and Borus and Nestel, 1971).

The reliability study had results generally in agreement with the validity review with one major exception: the role of race. The validity studies consistently demonstrated that blacks provided less valid data than whites; however, the reliability study indicates a balanced set of differences with a tendency for nonwhites to be more reliable than whites. Unfortunately, possible problems in nonresponse rates in the reliability study in particular or in the sequential loss in respondents over time could account for this discrepancy. Other alternatives are also possible: the validity studies might have been biased in their criterion or sample; or, while blacks produce less valid data, errors in response might be consistent over time. No definitive statement can be offered at this time; however, a careful study of this problem clearly is warranted.

The remaining relationships between respondent characteristics and reliability or validity are consistent. No clear edge exists for either males or females, but a trend exists of males producing higher quality data on

financial or numerical questions. High ability and high SES respondents are substantially more reliable and to a lesser extent more accurate than lower ability or lower SES respondents.

These differences in reliability and validity as a function of race, sex, SES, and ability can have profound consequences for research. In particular, generalizing results across populations differing on respondent characteristics would be highly problematic. One could not, for example, make reasonable estimates of measurement errors for structural modeling purposes if the reliability estimates came from a population differing in ability level or SES. Similarly, the path coefficients of causal models could well vary with respondent characteristics. In short, demographic variables need to be considered not only as control variables but also as moderator variables.

D. Reliability and Validity as a Function of Interactions Among Item, Data Collection, and Respondent Characteristics

The above discussion has generally focused on main effects associated with item, data collection, and respondent characteristics. The literature review and the reliability study both indicate that the factors interact. For example, the reliability study indicated that females unreliably report spouse's income in an interview but are as reliable as males in the mail-in procedure. In a similar fashion, Cannell and Fowler (1963) found that mail-in questionnaires produce more valid data than interviews if records can be consulted. The existence of such interactions not only requires that the main effects need to be qualified, but also indicates that the design of surveys and analyses based on survey data should carefully consider the content, data collection procedure, and respondent population. If the survey is directed primarily at high ability respondents, the method is probably irrelevant; however, if the respondent population is of low ability, interview procedures would be more appropriate—unless the content emphasized data for which records could be consulted.

Path analyses can apparently be safely undertaken with factually oriented data (excluding income) on high ability or high SES respondents, but otherwise structural modeling with error of measurement estimates based on a similar population is required. Overall, the safest approach would be to carry out reliability and validity pilot studies prior to the main survey. Only in this way could the most appropriate adjustments be made.

VI. REFERENCES

Astin, A.W. *Who Goes Where to College?* Chicago: Science Research Associates, 1965.

Boruch, Robert F., and Creager, John A. Measurement error in social and educational survey research. *ACE Research Reports*, May 1972, 7 (2).

Borus, Michael E. Response error in survey reports of earnings information. *American Statistical Association Journal*, 1966, *61*, 729-738.

Borus, Michael E., and Nestel, Gilbert. Response bias in reports of father's education. *Proceedings of the Social Statistics Section*, 1971, Washington, D.C.: American Statistical Association, 1972, 263-268.

Borus, Michael E., and Nestel, Gilbert. Response bias in reports of father's education and socioeconomic status. *Journal of the American Statistical Association*, 1973, *68*, 816-820.

Bureau of the Census. *The post-enumeration survey: 1950.* Technical Paper No. 4. Washington, D.C.: U.S. Government Printing Office, 1960.

Butterfield, E.C. Locus of control, test anxiety, reactions to frustrations, and achievement attitudes. *Journal of Personality*, 1964, *32*, 355-370.

Cannell, Charles F., and Fowler, Floyd J. Comparison of a self-enumerative procedure and a personal interview: A validity study. *Public Opinion Quarterly*, Summer 1963, *27*, 250-264.

Cohen, Roberta S., and Orum, Anthony M. Parent-child consensus on socioeconomic data obtained from sample surveys. *Public Opinion Quarterly*, Spring 1971, *36*, 95-98.

Coleman, J.S., Campbell, E.A., Hobson, C.J., Partland, J., Modd, A.M., Wernfield, F.D., and York, R.L. Equality of educational opportunity. Superintendent of Documents Catalog No. FS5.238.38001. Washington, D.C.: Government Printing Office, 1966.

Conger, A.J., Peng, S.S., and Dunteman, G.H. National Longitudinal Study of the High School Class of 1972: Group Profiles on Self-Esteem, Locus of Control and Life Goals. Research Triangle Park: Research Triangle Institute, 1976.

Cronbach, L.J. *Essentials of Psychological Testing.* New York: Harper & Row, 1970.

Duncan, O.D. "Introduction to Structural Equation Models." Academic Press, 1975.

Ecternacht, Gary J. The Response Stability of the Base-Year National Longitudinal Study Data and the Validity of the Base-Year National Longitudinal Study Data. Appendix F of the final report, "The Base-Year Survey of the National Longitudinal Study of the High School Class of 1972." Princeton: Educational Testing Service, 1973.

Ecternacht, Gary J. Reliability and validity of questionnaire items-status, social-psychological, and educational plans and aspirations variables. *Research Bulletin No. RB-74-12*, Princeton, N.J.: Educational Testing Service, April 1974.

Featherman, David L., and Hauser, Robert M. On the measurement of occupation in social surveys. *Sociological Methods and Research* 1973, *2* (2), 239-251.

Fiske, D.W. The limits for the conventional science of personality. *Journal of Personality*, 1974, 1-11.

Hansen, M., Hurwitz, W., and Pritzker, L. The estimation and interpretation of gross difference and the simple response variance. In *Contributions to Statistics*. Calcutta: Statistical Publishing Society, 1964.

James, W.H. *Internal versus external control of reinforcement as a basic variable in learning theory.* Unpublished doctoral dissertation, Ohio State University, 1957.

Keating, E., Peterson, P.G., and Stone, C.H. Validity of work histories obtained by interview. *Journal of Applied Psychology*, 1950, *34* 6-11.

Kerckhoff, Alan C., Mason, William M., and Poss, Sharon Sandomirsky. On the accuracy of children's reports of family social status. *Sociology of Education*, Spring 1973, *46* 219-247.

Kyaser, Brian D., and Summers, Gene F. The adequacy of student reports of parental SES characteristics. *Sociological Methods and Research*, 1973, *1* (3), 303-315.

Lefcourt, H.M. Internal *vs.* external control of reinforcement: A review. *Psychological Bulletin*, 1966, *65*, 206-220.

Lyons, N.I., and Moore, R.P. National Longitudinal Study of the High School Class of 1972: Reliability of Retrospective Data. Research Triangle Park: Research Triangle Institute, in press.

Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton, N.J.: Princeton University Press, 1965.

Rotter, J.B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, *80* No. 1 (whole no. 609).

Statistical Package for the Social Sciences. New York: McGraw-Hill, 1975.

van Es, J.C., and Wilkening, Eugene A. Response stability in survey research: A cross-cultural comparison. *Rural Sociology*, June 1970, *35* (2), 191-205.

Walsh, T.C., and Buckholdt. *Accuracy of retrospectively reporting work status and occupation five years ago*. U.S. Bureau of the Census working paper, Economic Statistics Branch, Population Division, Vol. 15, No. 3, 1970.

Walsh, W.B. Validity of the self-report. *Journal of Counseling Psychology*, 1967, *14* 18-26.

Walsh, W.B. Validity of the self-report: Another look. *Journal of Counseling Psychology*, 1968, *15*, 180-186.

WESTAT, Inc. *Sample Design in the Selection of a Sample of Schools with Twelfth-Graders in a Longitudinal Study*. Rockville, Md., 1972.

Wiggins, J.S. *Personality and Prediction: Principles of Personality Assessment*. Reading, Mass.: Addison-Wesley, 1973.

Appendix

SHORT-FORM QUESTIONNAIRE AND COVER LETTER

RESEARCH TRIANGLE INSTITUTE
POST OFFICE BOX 12194
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

CENTER FOR EDUCATIONAL RESEARCH AND EVALUATION

Dear Follow-Up Participant:

We appreciate your completing the second follow-up questionnaire for the National Longitudinal Study of the High School Class of 1972. Your cooperation greatly helps in the continuing effort to collect information for planning better programs to enhance work and educational activities of young people like yourself.

We are always trying to improve the questionnaire. One thing we wonder about is whether a question has the same meaning to a person when asked at different times. To determine this, we have selected a few questions from the second follow-up questionnaire, which you have already completed, and we are asking you to answer them again. The results of this study will help us improve future questionnaires.

Please read carefully each question in the short questionnaire. It is important that you follow the directions for responding. Sometimes you are asked to fill in a blank--in these cases, simply write your response. Where you are asked to circle a number, make a heavy circle. Here is an example:

Did you complete high school?

(Circle one.)

No, still in high school	1
No, left high school without completing	2
Yes, graduated	(3)

The entire questionnaire will take only a few minutes of your time. When you complete the questionnaire, please seal it in the postpaid envelope provided and return it to:

OPERATION FOLLOW-UP
Research Triangle Institute
Post Office Box 12036
Research Triangle Park, North Carolina 27709

Thank you again for your help.

Sincerely,

J. A. Davis
RTI Project Director

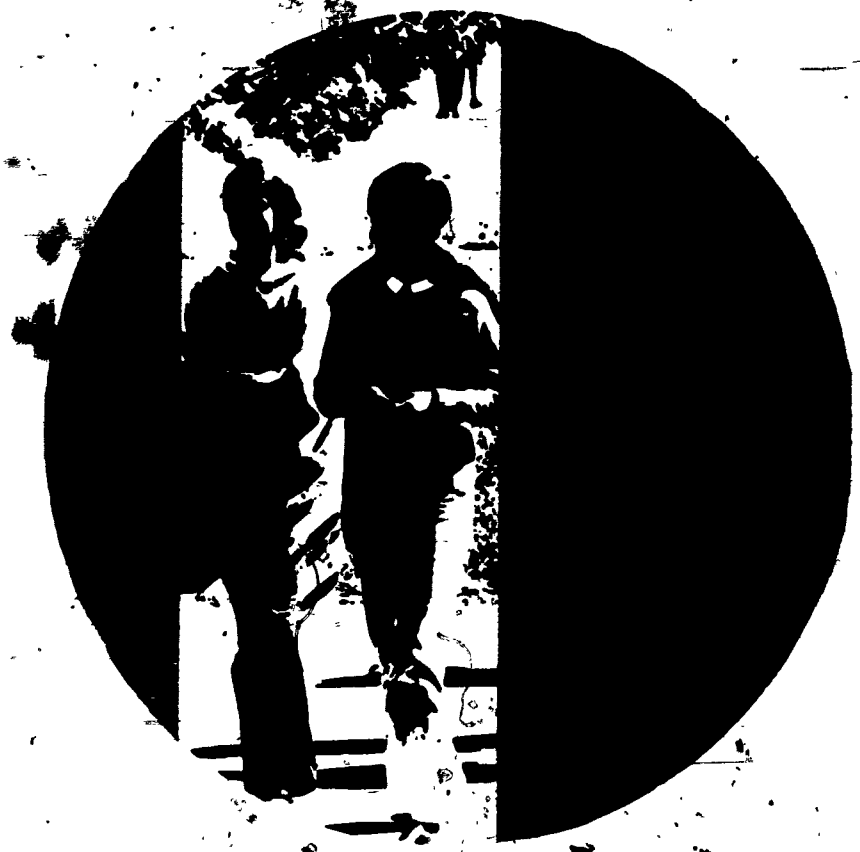
JAD:fh

Enclosure

NOTICE—All information which would permit identification of the individual will be held in strict confidence, will be used only by persons engaged in and for the purposes of the survey, and will not be disclosed or released to others for any purposes.

O.M.B. No. 51-S-74047
APPROVAL EXPIRES SEPT. 1975

OPERATION FOLLOW-UP



NATIONAL LONGITUDINAL STUDY OF THE HIGH SCHOOL CLASS OF 1972

SHORT FORM
Second Follow-Up Questionnaire

Prepared for the
DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
BY RESEARCH TRIANGLE INSTITUTE □ RESEARCH TRIANGLE PARK, NORTH CAROLINA
FALL 1974

1. **How do you describe yourself?**

	(Circle one.)
American Indian	1
Black or Afro-American or Negro	2
Mexican-American or Chicano	3
Puerto Rican	4
Other Latin-American origin	5
Oriental or Asian-American	6
White or Caucasian	7
Other	8

2. **Did you attend school in the first week of October 1974?**

No	1	*GO TO Q. 5*
Yes	2	*GO TO Q. 3*

3. **What kind of school is this?**

	(Circle one.)
Vocational, trade, business, or other career training school	1
Junior or community college (two-year)	2
Four-year college or university	3
Other (describe _____)	4

4. **Which of the following best describes how well you have done in all of your coursework or program from October 1973 through October 1974? If your school(s) or program(s) do not use letter grades, please choose the letter grade that comes closest to describing your progress.**

	(Circle one.)
Mostly A	1
About half A and half B	2
Mostly B	3
About half B and half C	4
Mostly C	5
About half C and half D	6
Mostly D or below	7

5. **Were you working during the first week of October 1974?**

No	1	*GO TO Q. 9*
Yes, full-time	2	*GO TO Q. 6*
Yes, part-time	3	*GO TO Q. 6*

6. **Please describe below the job you held during the first week of October 1974. (If you held more than one job at that time, describe the one at which you worked the most hours.)**

a For whom did you work? (Name of company, business organization, or other employer)
(Write in) _____

b What kind of business or industry was this? (For example, retail shoe store, restaurant, etc.)
(Write in) _____

c What kind of job or occupation did you have in this business or industry? (For example, salesperson, waitress, secretary, etc.)
(Write in) _____

d What were your most frequent activities or duties on this job? (For example, selling shoes, waiting on tables, typing and filing, etc.)
(Write in) _____

e Were you

(Circle one.)

An employee of a PRIVATE company, bank, business, school, or individual working for wages, salary, or commissions?	1
A GOVERNMENT employee (Federal, State, county, or local institution or school)	2
Self-employed in your OWN business, professional practice, or farm	3
Working WITHOUT PAY in family business or farm?	4

f. When did you start working at this job? _____(month) _____(year)

g Are you currently working at this job?

Yes 1

No 2 Date left _____(month) _____(year)

7. **How many hours did you usually work at this job in an average week?**

_____ Hours per week

8. **What was your marital status, as of the first week of October 1974?**

(Circle one.)

Never married, but plan to be married within the next 12 months	1
Never married, and don't plan to be married within the next 12 months	2
Divorced, widowed, separated	3
Married	4

9. **Not including yourself, how many persons were dependent upon YOU for more than one half of their financial support as of the first week of October 1974.**

(Circle one.)

0 1 2 3 4 or more

10. **What is the best estimate of your income before taxes for ALL OF 1974? If you are married, include your spouse's income in the total, but do not include loans and gifts. Please make an entry on each line, either a dollar amount, or if you will receive no income from a source during 1974, write in the word "none".**

Source	Amount Will Receive
Your own wages, salaries, commissions, and net income from a business or farm	$
Your spouse's (husband or wife) wages, salaries, commissions, and net income from a business or farm	$
All other income you and your spouse will receive (include interest, dividends, rental property income, public assistance, unemployment compensation, cash gifts scholarships fellowships, etc.)	$
TOTAL INCOME YOU AND YOUR SPOUSE WILL RECEIVE	$

11. **How do you feel about each of the following statements?**

(Circle one number on each line.)

	Agree Strongly	Agree	Disagree	Disagree Strongly	No Opinion
a I take a positive attitude toward myself	1	2	3	4	5
b Good luck is more important than hard work for success	1	2	3	4	5
c I feel I am a person of worth, on an equal plane with others	1	2	3	4	5
d I am able to do things as well as most other people	1	2	3	4	5
e Every time I try to get ahead, something or somebody stops me	1	2	3	4	5
f Planning only makes a person unhappy since plans hardly ever work out anyway	1	2	3	4	5
g People who accept their condition in life are happier than those who try to change things	1	2	3	4	5
h On the whole, I'm satisfied with myself	1	2	3	4	5

12. **What ways do you assure yourself of a good buy for your money?**

(Circle one number on each line.)

	Regularly	Sometimes	Never
a I compare prices and label information of similar products or services	1	2	3
b I return merchandise that is unsatisfactory to the store where I bought it	1	2	3
c I rely on brands or companies I know well even if they cost more	1	2	3
d I follow leads in articles from Consumer Reports, Changing Times, or other such magazines	1	2	3
e I check a company's reputation with the Better Business Bureau or consumer protection agency before agreeing to an expensive service or repair	1	2	3
f I write to the manufacturer about the quality of the product if I'm unsatisfied	1	2	3

13. **What do you expect to be doing in October 1975?**

(Mark as many as apply.)

- Working for pay at a full-time or part-time job 1
- Taking academic courses at a two- or four-year college 2
- Taking vocational or technical courses at any kind of school or college (for example, vocational, trade, business, or other career training school) 3
- On active duty in the Armed Forces (or service academy) 4
- Homemaker 5
- Other (describe _____) 6

14. **As things stand now, how far in school do you think you actually will get?**

(Circle one.)

- High school only 1
- Vocational, trade, or business school
  - Less than two years 2
  - Two years or more 3
- College program
  - Some college (including two-year degree) 4
  - Finished college (four or five-year degree) 5
  - Master's degree or equivalent 6
  - Ph.D., M.D., or equivalent 7

15. **How important is each of the following factors in determining the kind of work you plan to be doing for most of your life?**

(Circle one number on each line.)

		Very Important	Somewhat Important	Not Important
a	Previous work experience in the area	1	2	3
b	Relative or friend in the same line of work	1	2	3
c	Job openings available in the occupation	1	2	3
d	Work matches a hobby interest of mine	1	2	3
e	Good income to start or within a few years	1	2	3
f.	Job security and permanence	1	2	3
g	Work that seems important and interesting to me	1	2	3
h	Freedom to make my own decisions	1	2	3
i	Opportunity for promotion and advancement in the long run	1	2	3
j	Meeting and working with sociable, friendly people	1	2	3

16. **How important is each of the following to you in your life?**

(Circle one number on each line.)

		Very Important	Somewhat Important	Not Important
a.	Being successful in my line of work	1	2	3
b.	Finding the right person to marry and having a happy family life	1	2	3
c	Having lots of money	1	2	3
d	Having strong friendships	1	2	3
e	Being able to find steady work	1	2	3
f	Being a leader in my community	1	2	3
g	Being able to give my children better opportunities than I've had	1	2	3
h	Living close to parents and relatives	1	2	3
i	Getting away from this area of the country	1	2	3
j	Working to correct social and economic inequalities	1	2	3
k	Having leisure time to enjoy my own interests	1	2	3
l.	Having a good education	1	2	3

17. **What kind of work will you be doing when you are 30 years old? (Circle the one that comes closest to what you expect to be doing.)**

		(Circle one.)
a.	CLERICAL such as bank teller, bookkeeper, secretary, typist, mail carrier, ticket agent	1
b.	CRAFTSMAN such as baker, automobile mechanic, machinist, painter, plumber, telephone installer, carpenter	2
c.	FARMER, FARM MANAGER	3
d.	HOMEMAKER OR HOUSEWIFE ONLY	4
e.	LABORER such as construction worker, car washer, sanitary worker, farm laborer	5
f.	MANAGER, ADMINISTRATOR such as sales manager, office manager, school administrator, buyer, restaurant manager, government official	6
g.	MILITARY such as career officer, enlisted man or woman in the Armed Forces	7
h.	OPERATIVE such as meat cutter, assembler, machine operator, welder, taxicab, bus, or truck driver, gas station attendant	8
i.	PROFESSIONAL such as accountant, artist, registered nurse, engineer, librarian, writer, social worker, actor, actress, athlete, politician, but not including public school teacher	9
j	PROFESSIONAL such as clergyman, dentist, physician, lawyer, scientist, college teacher	10
k.	PROPRIETOR OR OWNER such as owner of a small business, contractor, restaurant owner	11
l	PROTECTIVE SERVICE such as detective, police officer or guard, sheriff, fire fighter	12
m	SALES such as salesperson, advertising or insurance agent, real estate broker	13
n.	SCHOOL TEACHER such as elementary or secondary	14
o	SERVICE such as barber, beautician, practical nurse, private household worker, janitor, waiter	15
p	TECHNICAL such as draftsman, medical or dental technician, computer programmer	16
q.	NOT WORKING	17

18. **When did you complete this questionnaire?**

_____ (month) _____ (day) _____ (year)

THANK YOU FOR YOUR COOPERATION

THIS INFORMATION WILL BE KEPT IN STRICT CONFIDENCE AND WILL BE USED ONLY FOR FUTURE FOLLOW-UPS IN THE NATIONAL LONGITUDINAL STUDY OF THE HIGH SCHOOL CLASS OF 1972

NAT

JU