DOCUMENT RESUME

ED 410 313 TM 027 663

AUTHOR	Salvucci, Sameena; Walter, Elizabeth; Conley, Valerie; Fink, Steven; Saba, Mehrdad
TITLE	Measurement Error Studies at the National Center for Education Statistics.
INSTITUTION	Synectics for Management Decision, Inc., Arlington, VA.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-97-464
PUB DATE	Jul 97
NOTE	234p.
AVAILABLE FROM	National Data Resource Center; fax: 703-820-7465, phone: 703-845-3151 (free single copies).
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Data Analysis; Data Collection; Elementary Secondary Education; *Error of Measurement; Evaluation Methods; *Information Dissemination; National Surveys; *Reliability; *Research Methodology; *Sampling; Validity
IDENTIFIERS	Educational Indicators; Multiple Measures Approach; *National Center for Education Statistics

ABSTRACT

This report provides an overview of a selection of measurement error studies conducted on National Center for Education Statistics (NCES) surveys. Its intent is not to offer new analyses of program data, but to summarize information from internal memoranda, working papers, and adjudicated reports about errors of measurement that occur during five phases of survey operations: sample selection; data collection; data processing; estimation and analysis; and dissemination of results and postsurvey evaluation. The report illustrates the diversity of NCES efforts in this area and focuses on the major national surveys NCES conducts. The emphasis of the review is on reinterview studies, but other types of empirical studies of measurement error are discussed, including multiple indicators studies, record check studies, and cognitive studies. The following chapters are included: (1) "Introduction and Overview"; (2) "Profile of NCES Reinterview Studies"; (3) "Reinterview Studies: Simple Response Variance and Response Bias"; (4) "Reinterview Studies: Reliability and Validity"; (5) "Multiple Indicators' Studies"; (6) "Record Check Studies"; (7) "Cognitive Studies"; and (8) "Summary." An appendix presents a summary of a reinterview study on mode effects for the 1990-91 Schools and Staffing Survey. (Contains 60 tables, 1 exhibit, 31 figures, 4 tables in the appendix, and 102 references.) (SLD)

NATIONAL CENTER FOR EDUCATION STATISTICS

Measurement Error Studies

at the National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement
NCES 97-464

NATIONAL CENTER FOR EDUCATION STATISTICS

Measurement Error Studies

at the National Center for Education Statistics

Sameena Salvucci
Elizabeth Walter
Valerie Conley
Steven Fink
Mehrdad Saba
Synectics for Management Decisions, Inc.

Steven Kaufman, Project Officer
National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement **NCES 97-464**

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ramon C. Cortines
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208–5574

July 1997

The NCES World Wide Web Home Page address is
http://www.ed.gov/NCES/

Suggested Citation
U.S. Department of Education. National Center for Education Statistics. *Measurement Error Studies at the National Center for Education Statistics,* NCES 97-464, by Sameena Salvucci, Elizabeth Walter, Valerie Conley, Steven Fink, and Mehrdad Saba. Project Officer: Steven Kaufman. Washington, DC: 1997.

Contact:
Steven Kaufman
(202) 219–1337

For free single copies of this publication, call the National Data Resource Center at (703) 845–3151 or send a FAX request to (703) 820–7465.

Preface

This report provides an overview of a selection of measurement error studies conducted on National Center for Education Statistics (NCES) surveys. Its intent is not to offer new analysis of program data, rather it summarizes information from internal memoranda, working papers, and adjudicated reports. The purpose of this report is to illustrate the diversity of NCES efforts in this area.

The emphasis of our review is on reinterview studies, but other types of empirical studies of measurement error including "multiple indicators" studies, record check studies, and cognitive studies will also be described. The report is not meant to be comprehensive, but can be used as a reference for types of items within surveys at NCES that have been examined, the techniques used for measuring and controlling measurement error, as well as the results of the examinations.

Acknowledgments

The authors wish to extend grateful thanks to all of those who contributed time, knowledge, and efforts to this report. National Center for Education Statistics (NCES) staff were enormously supportive. Our project officer, Steve Kaufman, provided guidance from beginning to end, but Dan Kasprzyk and Marilyn McMillen, Senior Technical Advisor, also reviewed several versions of the report. Project officers (past and current) for all of the surveys we discuss in this report helped us locate material, answered our questions, and reviewed their respective sections: from the Data Development and Longitudinal Studies Group, C. Dennis Carroll, Aurora D'Amico, Paula Knepper, and Andrew G. Malizio, and from the Surveys and Cooperative System Group, Dan Kasprzyk, Kathryn Chandler, and Peggy Quinn, Peter Stowe, and Linda Zimbler. Sue Ahmed, Chief Statistician of NCES, Statistical Standards and Services Group, chaired the adjudication committee. Ralph Lee, Data Development and Longitudinal Studies Group, reviewed the entire report. Irwin Schreiner, Bureau of the Census, Michael Brick, Westat, and Sylvia Fisher, Bureau of Labor Statistics, also reviewed the entire report, and all provided extensive and useful comments. The work was sponsored by the Surveys and Cooperative Systems Group.

Among Synectics staff, principal thanks goes to Dhiren Ghosh, who spent many hours working through the statistical models and formulas. Mindy Reiser, Steven Wenck, and George Carter also read and criticized various portions of the report. Kevin Jones created the figures.

Table of Contents

List of Tables

List of Exhibits

List of Figures

Acronyms

Statistics

Simple Response Variance and Response Bias (for definitions see chapter 3)

GDR	Gross Difference Rate
IOI	Index of Inconsistency
NDR	Net Difference Rate
SRV	Simple Response Variance
SV	Sampling Variance

Reliability and Validity (for definitions see chapter 4)

ϕ	Phi Coefficient (coefficient of association/agreement)
κ	Cohen's Kappa (coefficient of association/agreement)
ρ	Spearman's Rho (coefficient of rank correlation)
r	Pearson's Coefficient of Correlation
τ	Kendall's Tau (coefficient of rank correlation)

Other Acronyms

AE	Adult Education Component (NHES:95)
B&B	Baccalaureate and Beyond Longitudinal Study
BPS	Beginning Postsecondary Students Longitudinal Study
CAO	Chief Administrative Officer
CAPI	Computer-Assisted Personal Interview
CATI	Computer-Assisted Telephone Interview
ECE	Early Childhood Education Component (NHES:91)
FTB	First-Time Beginning Student (NPSAS)
HS&B	High School and Beyond
IPEDS	Integrated Postsecondary Education Data System
K	Kindergarten
LEA	Local Education Agency
LEP	Limited English Proficiency

NCES	National Center for Education Statistics
NELS:88	National Education Longitudinal Study of 1988
NHES	National Household Education Survey
NPSAS	National Postsecondary Student Aid Study
NSOPF	National Survey of Postsecondary Faculty
PK	Prekindergarten
RCG	Recent College Graduates Study
SASS	Schools and Staffing Survey
S.E.	Standard Error
SES	Socioeconomic Status
SR	School Readiness Component (NHES:93)
SS&D	School Safety & Discipline Component (NHES:93)
SSU	Secondary Sampling Units
TDS	Teacher Demand and Shortage Survey (SASS)
TFS	Teacher Followup Survey
TLF	Teacher Listing Form
TLR	Teacher Listing Record
TVLS	Teacher List Validation Study (a 1993 study)

CHAPTER 1
Introduction and Overview

Monitoring survey data quality involves learning as much as possible about errors[1], both in a descriptive sense for statistical correction strategies, and in a causal sense for improving the process. The National Center for Education Statistics (NCES) has shown its commitment to the evaluation of the quality of its survey data through systematic, ongoing efforts to monitor the components of error in its data products, making after-the-fact corrections as necessary, and constantly improving the survey process designs to eliminate errors before they occur.

This report focuses on illustrating an important part of this NCES commitment: the study of measurement errors that occur during the data collection stage of its surveys. It reviews a sample of past and current measurement error studies conducted by NCES and summarizes the results of each study, drawing upon relevant NCES publications. The report does not provide a comprehensive review of all its measurement error programs, but rather indicates the range of techniques used within these programs across its surveys. More importantly, by uniting results of measurement error studies previously available in widely disparate sources, from internal memorandums and working papers to adjudicated NCES reports, we hope that the information can now be more easily accessible for use by managers during survey planning and by users of NCES data.

This first chapter, a general introduction to survey errors, including measurement error, will define the terms and concepts used in the subsequent sections of the report. This discussion is not a complete description of the topic, but is intended to provide the reader with a framework for interpreting the information presented.

Measurement Errors in the Context of Other Survey Errors

A survey consists of a number of survey operations and in large national surveys, such as those conducted by NCES, the operations may extend over a considerable period of time, from the planning stage to the ultimate publication of results. Each phase of the operations affects the quality of survey estimates, and with each phase we can associate sources of errors in the estimate. Sarndal, Swensson, and Wretman (1992) distinguish five phases of survey operations and associated errors as follows:

a. Sample Selection

This phase consists of the execution of a preconceived sampling design using a suitable sampling frame. The sample size necessary to obtain the desired precision is determined.

[1] In the context of this report "error" refers to deviations of obtained survey results from those that are true reflections of the population.

Errors in estimates associated with this phase are (1) *frame errors*, of which undercoverage is particularly serious, and (2) *sampling error*, which arises because a sample, not the whole population, is observed.

b. Data Collection

There is a preconceived measurement plan with a specified mode of data collection (personal interview, telephone interview, mail questionnaire, or other). The field work is organized, interviewers are selected, and interviewer assignments are determined. Data are collected, according to the measurement plan, for the elements in the sample. Errors in estimates resulting from this phase include (1) *measurement errors* when, for instance, the respondent gives (intentionally or unintentionally) incorrect answers, the interviewer misunderstands or records incorrectly, the interviewer influences the responses, the questionnaire is misinterpreted, etc. and (2) *error due to nonresponse* (i.e., missing observations).

c. Data Processing

During this phase collected data are prepared for estimation and analysis. It includes the following elements

- Coding and data entry
- Editing
- Renewed contact with respondents to get clarification if necessary
- Imputation

Errors in estimates associated with this phase include *transcription error (keying errors), coding errors, error in imputed values, errors introduced by or not corrected by edit.*

d. Estimation and Analysis

This phase entails the calculation of survey estimates according to the specified point estimator formula, with appropriate use of auxiliary information and adjustment for nonresponse, as well as a calculation of measures of precision in the estimates (e.g., variance estimate, coefficient of variation of the estimate, confidence interval). Statistical analyses may be carried out, such as comparison of subgroups of the population, correlation and regression analyses, etc. All error from phases (a) to (c) above will affect the point estimates, and they should ideally be accounted for in the calculation of the measures of precision.

e. Dissemination of Results and Postsurvey Evaluation

This phase includes the publication of the survey results, including a general declaration of the conditions surrounding the survey. This declaration often follows a set of specified guidelines for quality declaration which traditionally include two major categories: sampling and nonsampling errors.

Conceptual Structure of Survey Errors

The field of measurement of survey error components has evolved through the somewhat independent, and uncoordinated, contributions of researchers trained as statisticians, psychologists, political scientists, and sociologists. Therefore, it lacks a common language and a common set of principles for evaluating new ideas. According to Groves (1989) at least three major languages of error appear to be applied to survey data. They are associated with three different academic disciplines and exemplify the consequences of groups addressing similar problems in isolation of one another. The three disciplines are statistics (especially statistical sampling theory), psychology (especially psychometric test and measurement theory), and economics (especially econometrics). Although other disciplines use survey data (e.g., sociology and political science), they appear to employ languages similar to one of those three.

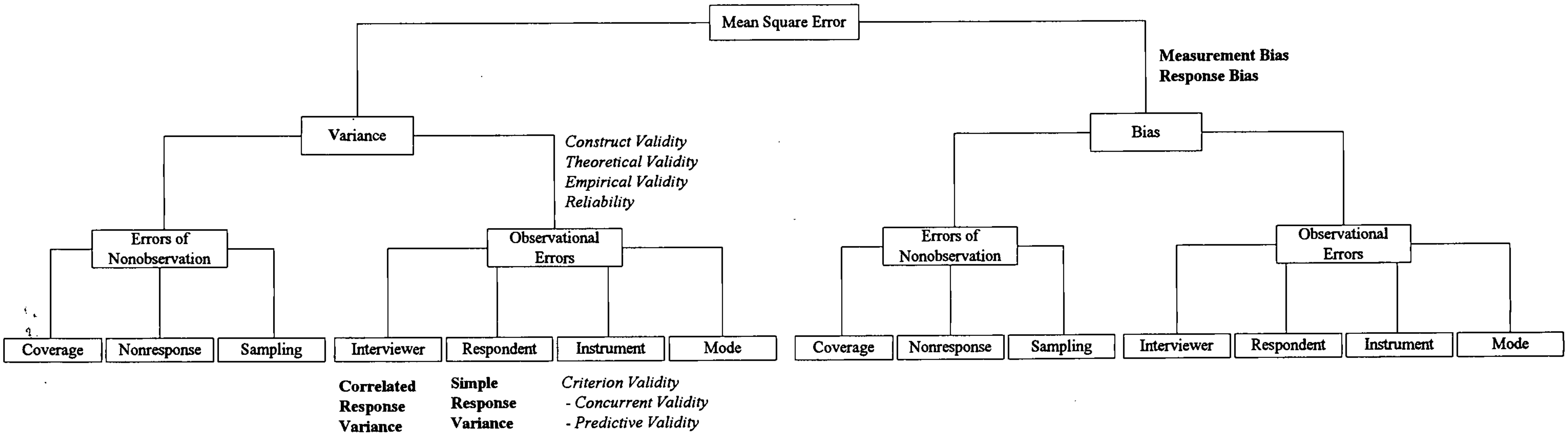
Some attention to these terminology differences is necessary to define *measurement error* unambiguously. Groves uses four nested levels of concepts to classify errors (see exhibit 1). "The total error of a survey statistic is labeled the *mean squared error*; it is the sum of all variable errors and all biases. *Bias* is a systematic error that affects the statistic in all implementations of a survey design; in that sense it is a constant error (e.g., all possible surveys using the same design might overestimate the mean years of education per person in the population). A variable error, measured by the *variance* of a statistic, arises because achieved values differ over the units (e.g., sampled persons, interviewers used, questions asked) that are the sources of the errors. The concept of variable errors inherently requires the possibility of repeating the survey, with changes of units in the replications (e.g., different sample persons, different interviewers). Variable errors and biases are therefore connected; bias is the part of error common to all implementations of the survey design, and variable error is the part that is specific to each trial" (1989, p. 8).

"There are two types of error under both 'variance' and 'bias' in exhibit 1. *Errors of nonobservation* are those arising because measurements were not taken on part of the population. *Observational errors* are deviations of the answers of respondents from their true values on the measure; for our purposes, these are *measurement errors*" (Groves, 1989, p. 11).

"The final level of conceptual structure concerns the alternative sources of the particular error. Errors of nonobservation are viewed as arising from three sources—coverage, nonresponse, and sampling" (Groves, 1989, p. 11). Sources of *observational errors* are categorized into four principal sources

- the interviewer
- the respondent
- the questionnaire
- the mode of data collection, that is, whether telephone, personal interview, self-administered questionnaire, or other medium is used

Exhibit 1. -- A conceptual structure and languages of error sources in surveys



Notes:
All terms in **bold** refer to error terminology from survey statistics literature
All terms in *italics* refer to error terminology from psychometric literature

SOURCE: Derived from figure 1.1, figure 1.3, and figure 1.5, Groves, (1989), *Survey Errors and Survey Costs*, p.10.

Note that exhibit 1 "is not a complete enumeration of all sources of error in survey data. The most notable omissions are those observational errors arising after the answers to the survey questions have been obtained by the interviewers—the coding, editing, imputation, and other data processing activities that follow the data collection phase" (Groves, 1989, p. 12). These will not be considered measurement errors, but *processing errors.*

Languages of Measurement Error

Measurement error was defined above as any error arising in the data collection phase of the survey operations. Having presented the full conceptual structure of error terms above, we will now discuss two alternative "measurement error" notions from different disciplines in order to offer some definitions of the terms used throughout this report. It will become clear that differences across these disciplines are not merely a matter of different labels for similar concepts of error, but real differences in the set of factors that are seen to influence survey estimates. Groves (1989) asserts that answering the following three questions will avoid any misunderstanding between the two concepts of measurement error.

- What is the statistic of interest when errors are being considered?
- Which features of the data collection are viewed to be variable over replications and which are fixed?
- What assumptions are being made about the nature of those persons not measured, or about properties of the observational errors?

We will keep these questions in mind while reviewing notions of measurement error in the two disciplines.

a. Measurement Error Terminology in Survey Statistics (see exhibit 1)

Groves (1989, p. 15) says "a more elaborated view of survey error held by some survey statisticians comes from those interested in *total survey error* (e.g., Fellegi, 1964; Hansen, Hurwitz and Pritzker, 1964; Bailar and Dalenius, 1969; Koch, 1973; Bailey, Moore and Bailar, 1978; Lessler, Kalsbeek and Folsom, 1981). Underlying this perspective is the notion that the survey at hand is only one of an infinite number of possible trials or replications of the survey design. Respondents are assumed to vary in their answers to a survey question over trials, leading to *simple response variance* (Hansen, Hurwitz and Pritzker, 1964). The interviewer is often treated as a source of error in this perspective, and is most often conceptualized as a source of variable error. The variable effects that interviewers have on respondent answers are sometimes labeled *correlated response variance* in this perspective (Bailey, Moore and Bailar, 1978)." *Measurement bias or response bias* refers to systematic errors that have a discernible pattern compared to the "true response." "For example, if respondents tend to omit certain types of income, say interest income from savings, then the estimated income would be expected to be lower than the true income" (Brick, Kim, Nolin and Collins, 1996, p. 3).

b. Measurement Error Terminology in Psychological Measurement (see exhibit 1)

Groves (1989) states that "when moving from survey statistics to psychometrics, the most important change is the notion of an unobservable characteristic the researcher is attempting to measure with a survey indicator (i.e., a question). In contrast, within survey statistics, the measurement problem lies in the operationalization of the question (indicator, in psychometric terms). That is, in psychometrics the problem is not the impossibility of measuring the characteristic, but the weakness of the measure.

There are two influential measurement models. In the first, *classical true score* theory, all observational errors are viewed as joint characteristics of a particular measure and the person to whom it is administered. Errors in responses are acknowledged. In such measurement, however, the expected value (over repeated administrations) of an indicator is the true value it is attempting to measure. That is, there is no measurement bias possible, only variable errors over repeated administrations.

Although classical true scores provide the basis for much of the language of errors in psychometrics, it is found to be overly restrictive for most survey applications. The need to acknowledge possible biases in survey measurements is strong. Therefore, in psychometrics, most survey measures will be labeled as sets of *congeneric measures* or indicators in a *multiple factor model*, where measurement errors can yield biases in indicators of underlying constructs (characteristics) and indicators can be influenced by various methods of measurement" (Groves, 1989, p. 18).

"An additional change when moving to the field of psychometric measurement is the explicit use of models as part of the definition of errors. That is, error terms are defined assuming certain characteristics of the measurement apply. In classical true score theory (a model), the most important assumption is that if an indicator were administered to a person repeatedly (and amnesia induced between trials), the mean of the errors in the respondent's answers would be zero. That is, the indicator is an 'unbiased' measure of the respondent's characteristic, in the sense used by survey statisticians. (Here the parameter of interest is the single respondent's value on the construct). This is not as strong an assumption as it may appear to be because psychometricians often view the scale on which their measurements are made as rather arbitrary. This fact arises because the statistics of interest to psychometricians are not generally means or totals for persons studied, but rather correlation coefficients, relative sizes of variance components, factor loadings, and standardized regression coefficients. All of these statistics are functions of variance and covariance properties of measures, not of means (expected values).

In this perspective, expectations of the measures are taken over trials of administration of the measurement of a person. That is, each asking of a question is one sample from an infinite population (of trials) of such askings. The *propensity distribution* describes the variability over trials of the error for the particular person. Under the classical true score assumption the mean of that distribution is zero. The only concept of error akin to those of the survey statistician is the variance of the error term, the *error variance* (Lord and Novick, 1968). This is the dispersion of the propensity distribution. When there is interest in a population of persons, the expected value of the indicator is taken both over the many propensity distributions of the persons in the population *and* the different persons. It is only within this context (measurement of a set of persons) that other concepts of error are defined" (Groves, 1989, p. 19).

Two terms in the psychometric perspective, *validity* and *reliability*, are frequently used to label two kinds of variable errors.

Messick (1989) states that "The major concern of *validity* is not to explain any isolated event, behavior, or item response, because these almost certainly reflect a confounding of multiple determinants. Rather, the intent is to account for *consistency* in behaviors or item responses, which frequently reflects distinguishable determinants. In contrast with treating the item responses separately as a conglomeration of specifics, these response consistencies are typically summarized in the form of total scores or subscores. The term *scores* is used here in the most general sense of any coding or summarization of observed consistencies on a test, questionnaire, observation procedure, or other assessment device. The emphasis is on scores and measurements as opposed to tests or instruments because the properties that signify adequate assessment are properties of scores, not tests or questionnaires. Questionnaires do not have reliabilities and validities, only questionnaire responses do. This is an important point because responses are a function not only of the items, tasks, or stimulus conditions but of the *persons* responding and the *context* of measurement" (p. 14).

The notion of *theoretical validity,* sometimes called *construct validity*, "is based on an integration of any evidence that bears on the interpretation or meaning of the scores or measurement. The measurement or score is not equated with the construct it attempts to tap, nor is it considered to define the construct. This is in stark contrast with strict operationism, in which each construct is defined in terms of a narrowly specified set of operations that becomes its sole empirical referent. Rather, the measure is taken to be one of an extensible set of indicators of the construct. Indeed, the construct is invoked as a latent variable or 'causal' factor to account for the relationships among its indicators. Because the set of indicators is extensible and because indicators are often probabilistically related to the construct as well as to each other, constructs are not explicitly defined, but, rather, are more like 'open concepts' (Pap, 1953, 1958).

Construct validity also subsumes *content relevance and representativeness* as well as *criterion-relatedness*, because such information about the content domain of reference and about specific criterion behaviors predicted by the scores or measurements clearly contributes to score interpretation. In the latter instance, correlations between test scores and criterion measures, viewed in the broader context of other evidence supportive of score meaning, contribute to the joint construct validity of both predictor and criterion. In other words, empirical relationships between the predictor scores and criterion measures should make theoretical sense in terms of what the predictor test is interpreted to measure and what the criterion is presumed to embody (Gulliksen, 1950).

Thus, construct validity embraces almost all forms of validity evidence. The only source of evidence not yet explicitly incorporated in a type of validity is the appraisal of social consequences" (Messick, 1989, p. 17). More details on the components of construct validity and the data and analyses relevant to construct validation can be found in Messick (1989).

Note well that "validity" is not to be simply equated with "unbiasedness," as used by survey statisticians because it is defined only on a population of persons (who vary on the true values), not on a single person. That is, there is no concept of a valid measure of a single person's attribute.

The other.error concept used in psychometrics is *reliability,* the ratio of the true score variance to the observed variance (Bohrnstedt, 1983, p. 73). Groves (1989) says "*variance* refers to variability over persons in the population and over trials within a person. With this definition of reliability, it can be noted that the concept is not defined for measurements on a single person, only on a population of persons and reliability has a value specific to that population.

When dealing with populations of persons, true score theory adds another assumption about the errors, that their values are uncorrelated with the true values of the persons on any of the trials. With this assumption the theoretical validity of a measure is merely the square root of its reliability. This relationship shows how different the concepts of reliability and validity, on one hand, are from variance and bias, on the other. Given this definition, the traditional statement that 'no measure can be valid without also being reliable, but a reliable measure is not necessarily a valid one,' is not true. In contrast, a sample statistic may have an expected value over samples equal to the population parameter (unbiasedness), but have very high variance from a small sample size. Conversely, a sample statistic can have very low sampling variance (from an efficient sample design) but have an expected value very different from the population parameter (high bias).

Validity and reliability can be assessed only with multiple indicators. Psychometricians use a set of adjectives for 'validity' in the context of multiple measures of a single construct and/or multiple constructs. Bohrnstedt (1983) makes the distinction between theoretical validity, which is defined on a single indicator, and *empirical validity,* an estimation of theoretical validity that can be implemented only with another measure of the same construct. Sometimes *criterion validity* is used to denote that the other measure is assumed to be measured without any variable error. There are at least two types of empirical validity, which differ in the characteristics of the criterion chosen. *Predictive validity* is the correlation between one measure of a construct and another (presumably with better error features) taken at a later time. *Concurrent validity* is the correlation between a measure and some criterion measured at the same point in time. These error measurement procedures in psychometrics resemble the reinterview studies in surveys, used to measure response variance on the part of the sample. In these the same question is asked of the respondent at two points in time, and the response variance is measured as a function of the differences in the answers (Groves, 1989, pp. 23-23).

There are two additional terms, however, that need clarification. "*Random measurement error*, as used by Andrews (1984, p. 412), refers to 'deviations (from true or valid scores) on one measure that are statistically unrelated to deviations in any other measure being analyzed concurrently.' In the language of survey statistics this would refer to lack of correlation between two variables in their response deviations. *Correlated measurement error* means 'deviations from true scores on one measure that do relate to deviations in another measure being concurrently analyzed.' Thus, correlated measurement error means something very different from the *correlated response variance* used by survey statisticians. The latter refers to correlations among respondents contacted by the same interviewer (or other administrative units) in deviations obtained on one indicator. The correlated measurement errors could arise from the fact that two indicators share the effects of a common method of measurement. Such a viewpoint is central to the multitrait multimethod approach to estimating construct validity. This alternative measurement model retains all the basic concepts of error, but necessarily alters the computational forms of error estimates" (Groves, 1989, p. 26).

Empirical Estimation of Survey Measurement Error

This section describes some techniques for evaluating and controlling measurement error in surveys. The methods discussed are those used at NCES: (1) Reinterview Studies, (2) "Multiple Indicators" Studies, (3) Record Check Studies, and (4) Cognitive Studies.

1. Reinterview Studies

A reinterview—*replicated measurement on the same unit in interview surveys*—is a new interview which repeats (a subset of) the questions of the original interview. When implementing reinterview methodology, there are two underlying assumptions: 1) the reinterview is independent of the first interview, and 2) the original interview and the reinterview either use the same mode of data collection and are conducted under the same general conditions or the reinterview and a reconciliation provide "true" values. Reinterview studies requiring two sets of measurements on the sample or part of it have been implemented since the early days of sample surveys (Mahalanobis, 1946). There are two major purposes for conducting reinterview studies at NCES: (1) to estimate simple response variance or reliability and (2) to estimate response bias.

A reinterview used to measure either *simple response variance* or *reliability* must be an independent replication of the original interview. Independence is threatened, however, by conditioning, which occurs when respondents remember their first answer during the reinterview.

Theoretically, the measurement of response bias requires the existence of data from which the true value may be estimated; however, often these data do not exist. In practice, reinterview programs frequently estimate a measure of response bias by including a process known as *reconciliation*. This is when the respondent is asked to reconcile answers that differed between the original and the reinterview. Reconciliation can occur during or at the end of the reinterview or in a separate, third contact. NCES practice is to conduct reconciliation as part of the reinterview to avoid unnecessary respondent burden and cost. Reconciliation in NCES reinterview studies is typically done using computer-assisted telephone interview (CATI), which prompts the interviewer when the original and the reinterview responses are different. Typically, if no attempt to reconcile the original and the reinterview is made, then the results are interpreted as measures of simple response variance.

The different purposes for which reinterviews may be used necessitate different methodologies and thus dictate different reinterview designs. Forsman and Schreiner (1991) describe four *basic reinterview designs*. Two focus on evaluating interview performance (one of which was specifically developed to detect interviewer falsification), and two on estimating measurement error components of the interview data—one estimating simple response variance and reliability and the other estimating response bias.

Forsman and Schreiner (1991) explain that "each basic design is characterized by the following six factors:

- *The method of reinterview sample selection*. The reinterview sample can be a one-stage sample of respondents, households, or clusters of households (such a cluster may consist of, e.g., four neighboring households). The reinterview sample can also be a two-stage sample, where the original interviewers are primary sampling units, and respondents (or households or clusters) within interviewers are secondary sampling units (ssu). Such a two-stage sample permits a proper allocation of ssu's over interviewers.

- *The choice of reinterviewer*. The reinterviewers can be selected from the same pool of interviewers as the original interviewers. They may also be selected from among the most experienced interviewers in this pool. A third option is to select the reinterviewers from a group of supervisors.

- *The choice of respondent*. The respondent can be the same as in the original interview; he or she can be chosen according to the same procedure as in the original interview ("original respondent rule"); the respondent might be the most knowledgeable person in the household, or each person could respond for himself or herself ("self-response").

- *The design of the reinterview questionnaire*. The reinterview questionnaire may be exactly the same as the original questionnaire, or may contain a subset of the original questions. To achieve "true" values, the reinterview questionnaire may contain probing questions.

- *Whether or not to conduct reconciliation*. When the responses obtained during the reinterview differ from those obtained in the original interview the differences are evaluated through a process called reconciliation. During reconciliation the respondent is provided with the information received in both interviews and asked to determine what is the correct information.

- *The choice of mode*. The choice is between telephone and face to face interviews (Forsman and Schreiner, 1991, pp. 282-83).

If the purpose of the reinterview is to estimate response variance or reliability the questions are repeated exactly, the responses are not reconciled, and the mode is the same as in the original interview. When estimating bias, however, the purpose is to obtain the "true" response. Here, the reinterview design should include the most experienced interviewers and supervisors. Likewise, reinterviews designed to measure response bias should target the most knowledgeable respondent, not necessarily the original respondent. If estimating response bias, the questions can be modified to elicit more accurate responses, reconciliation is used, and the mode of data collection need not be the same as the original interview. With CATI, the purpose of the reinterview can be to estimate error components alone since a centralized telephone facility can be monitored to deter falsification and to provide feedback to the interviewers.

Chapter 2 of this report summarizes the various issues pertaining to NCES reinterview studies. Chapter 3 of this report describes selected NCES reinterview studies that estimate simple response variance and response bias, while chapter 4 of this report describes selected NCES reinterview studies that estimate reliability and validity from the psychometric perspective.

2. "Multiple Indicators" Studies

Groves (1989) describes *multiple indicators* studies as another approach that uses replicated measures to estimate measurement error, but it uses multiple measurements of the same characteristic in a single survey. In this approach measurement error associated with a particular method of data collection and/or a particular question can be assessed. "Measurement error" here is defined as a component of variance in the observed value of indicators, not corresponding to variability in the true values of the underlying measures. In the terminology of this report, it corresponds to variable errors of measurement only. "Method" has been used to mean the mode of data collection (personal, telephone, or self-administered), the format of the question (5 point scale, 10 point scale, open questions), the respondent rule (self-response, proxy response), or various other characteristics of the measurement.

Chapter 5 of this report summarizes selected NCES "multiple indicators" studies.

3. Record Check Studies

Record check studies are used to estimate *response bias*. As described in the section on reinterview studies, the measurement of response bias theoretically requires the existence of data from which the true value may be estimated. When these data do not exist, reinterview studies frequently use reconciliation. When these data do exist and are available, record check studies are possible. Such a study generally assumes that information contained in the records is without error, that is, the records contain the true values on the survey variables.

Groves (1989) describes three kinds of record check study designs: the reverse record check study, the forward record check study, and the full design record check study. The different designs are based in part on the relation of the survey sample to the external source of data providing the comparisons.

In the *reverse record check study*, which Groves also refers to as the retrospective design, the researcher goes back to the records which were the source of the sample to check the survey responses. That is, the survey sample is drawn from a record file considered to contain accurate data on a trait or characteristic under study, and the survey includes some questions on information already in the records. The survey data are compared with the record data to estimate measurement error.

The weakness of reverse record check studies is that they cannot by themselves measure errors of overreporting—falsely reporting an event. They can only measure what portion of the records sample correspond to events reported in the survey and whether the characteristics of the events are the same on the records as in the survey report.

In a *forward record check study*, the researcher obtains the survey data first and then moves to new sources of record data for the validity evaluation. Thus, in this design, the sample is drawn from a separate frame. Once the survey responses have been collected, the researcher searches for relevant records containing information on the respondents and makes comparisons. Some surveys may be designed to include questions asking about where records containing similar information on the sample person can be found.

Forward record check studies work well for measuring overreports in a survey, but they are not commonly used. They generally entail contacting several different record-keeping agencies and may require asking the respondents for permission to access their record files from the different agencies. They are also limited in their measurement of underreporting: "They learn about the failure to report events only when mention of those events appear on records corresponding to other events which *are* reported. Records are not searched for those respondents who fail to report any event" (Groves, 1989, pp. 301-302).

The *full design record check* study combines features of the reverse and forward record check designs. The survey sample comes from a frame covering all persons of the population (reverse record check design) and researchers seek records from all sources relevant to those persons (forward record check design). Thus, researchers measure survey errors associated both with underreporting and overreporting by comparing all records corresponding to the respondent. However, this design requires a data base that covers all persons in the target population and all events corresponding to those persons.

All validity evaluation designs share three limitations. As mentioned earlier, there is the assumption that the record systems do not contain errors of coverage, nonresponse, or missing data. Second, it is also assumed that the individual records are complete and accurate, without any measurement errors. The third limitation involves matching errors—difficulties matching respondent survey records with the administrative records—and these could affect the estimation of measurement errors. As Groves explains, "*If mismatches occur at random within the subsets,* the expected mean difference between interview responses and mismatched records will be equal to that of the expected mean difference between interview responses and correctly matched records. However, even under such restrictive assumptions, the variance in response errors will be overestimated with the possibility of mismatching and the regression of measured response error on the matched record value will have a smaller slope than that of correct response error on the correct record value" (Groves, 1989, p. 302).

Chapter 6 of this report summarizes selected NCES record-check studies.

4. Cognitive Studies

Forsyth and Lessler (1991) contend that "if we are to understand the sources of survey measurement error and find ways of reducing it, we must understand how errors arise during the question-answering process. This will allow us to develop better questions that will yield more accurate answers. The primary objective of cognitive laboratory research methods is not to merely study the response process, but through careful analysis to identify questioning strategies that will yield more accurate answers" (p. 394). As Nolin and Chandler (1996) explain, the methods of cognitive research can be used to increase understanding of the ways that respondents comprehend survey instructions and questions, recall requested information, and respond to the influence of word and question order.

Cognitive research draws on three different literatures: research in cognitive psychology on memory and judgment, research in social psychology on influences against accurate reporting, and evidence from survey methodology research regarding response errors in surveys. Literature in survey methodology concentrates on models of measurement of response errors, rather than on explaining their presence. For example, survey methodology has documented response errors and identified respondent groups and response tasks that are more prone to these errors. Theories of cognitive psychology have been applied to survey measurement to gain insight into how the respondent's attributes and actions may affect the quality of survey data. These theories focus on how people encode information in their memories and how they retrieve it later. Social psychological literature, on the other hand, emphasizes the influences on communication of answers to survey questions. For example, social psychological literature concentrates on understanding why some respondents appear to work more diligently at obtaining the correct answer, or how the interviewing environment can influence respondents toward one answer or another (Groves, 1989, p. 409).

Researchers generally agree on five stages of action relevant to survey measurement error

- Encoding of information—how the respondent obtains, processes, and stores information in memory
- Comprehension—how the respondent assigns meaning to the interviewer's question
- Retrieval—how the respondent searches for memories of events or knowledge relevant to the question
- Judgment of appropriate answer—how the respondent chooses from alternative responses to the question
- Communication—how the respondent answers through all the other personal characteristics and social norms that might be relevant (Groves, 1989)

Beyond acceptance of these five stages, cognitive research takes different paths.

Cognitive Research Methods

Forsyth and Lessler (1991) conducted a literature review of cognitive research methods used to study the survey question-answering process and discussed the topic with others who have conducted cognitive research. They concluded that no guidelines were available for choosing one cognitive research method over another. While a number of response models have been developed, there is yet little consensus on how the models are implemented. The Oksenberg and Cannell (1977) and Tourangeau (1984) models assumed a basic sequence that respondents followed when answering a question, but there is no consensus on the procedural details of these methods. Forsyth and Lessler "believe that this lack of consensus may be due, in part, to a lack of theoretical and empirical work that explores how methodological details can affect cognitive laboratory results" (Forsyth and Lessler, 1991, p. 395). Nonetheless, they offer a summary of four general sets of methods that have been implemented. (See table 1.)

Table 1. -- Cognitive laboratory research methods currently being used in the United States to study the question-answering process

General Type of Method	Specific Method
Expert evaluation	Interactional behavior coding Cognitive forms appraisal Expert analysis
Expanded interviews	Concurrent think-aloud interviews Follow-up probes Memory cue tasks Retrospective think-alouds and probe questions
Targeted methods	Paraphrasing Free-sort classification tasks Dimensional-sort classification tasks Vignette classifications Rating tasks Response latency Qualitative timing
Group methods	Focus groups Group interviews Group experiments

Methods were identified during a literature review as well as informal discussions with cognitive laboratory research measurement staff at the Bureau of the Census, National Center for Health Statistics (NCHS), Bureau of Labor Statistics (BLS), Westat, Inc., and Research Triangle Institute (RTI).

SOURCE: Derived from table 20.1, Forsyth and Lessler, (1991), "Cognitive Laboratory Methods: A Taxonomy," in Biemer, Groves, Lyberg, Mathiowetz and Sudman (eds.) *Measurement Errors in Surveys,* p. 397.

Expert evaluation methods—interactional behavior coding, expert analysis, and cognitive forms appraisal—involve no interaction with respondents. In interactional behavior coding, an observer codes "interactions between interviewers and respondents during the question-answering process" (Forsyth and Lessler, 1991, p. 396). In expert analysis, "a researcher (who may or may not have originally constructed or formulated the questions to be asked) reviews a questionnaire to gather an understanding of the response task and to note potential problems" (Forsyth and Lessler, 1991, p. 397). While the researcher may classify observations on, for example, the types of mistakes respondents might make, these observations are not subject to a formal coding scheme. In contrast, when questions are analyzed under cognitive forms appraisal, the analysis is based on a model and questions are assigned codes that "describe the response process and are directed at identifying problems" (Forsyth and Lessler, 1991, p. 397).

Expanded interview methods refer to interviews where the survey questions are accompanied by probes about how the respondents perceive the survey items and how they decide to answer them. These methods include concurrent think-aloud interviews, follow-up probe questions, memory cue tasks, and retrospective think-aloud and probe methods. In concurrent think-aloud interviews, respondents are instructed to voice their thoughts as they attempt to answer survey questions. “Think-aloud results have been used to identify difficulties in question comprehension, perceptions of the response task, memory recall strategies, difficulties in selecting a response, interpretations of question reference period, and reactions to sensitive questions (Forsyth and Lessler, 1991, p. 398).

Probe questions are used to focus respondents’ attention on particular aspects of the questions or on the whole question-answering process. If used with concurrent think-aloud techniques, they may direct a respondent to focus on, for example, what procedures the respondent is using to recall information. Follow-up probing may be used after analysis of the question-answering procedure has highlighted some focal issues researchers wish to explore. Retrospective think-alouds and probe questions, on the other hand, are used after respondents have completed an interview under “normal” conditions. Memory cue tasks are used “to assess recall errors due to a respondent’s failure to remember events during an interview” (Forsyth and Lessler, 1991, p. 399).

Targeted methods use survey items as stimulus material for other tasks. These methods include paraphrasing, free-sort classification tasks, dimensional-sort classification tasks, vignette classifications, rating tasks, response latency, and qualitative timing. In paraphrasing, respondents are asked to restate questions in their own words to determine whether they understood the questions. The three classification methods are used to determine how respondents conceptualize the topics covered by the questionnaire. In free-sort classification, respondents are asked to sort a set of cards that list survey items into groups using any criteria they choose. Dimensional-sort classifications have respondents sort items according to gradations of some characteristic. Finally, in vignette classification methods the respondents “are asked to read short descriptions of situations (“vignettes”) and select category labels that best describe the situations” (Forsyth and Lessler, 1991, p. 400).

Rating tasks are used to identify questions that respondents have difficulty answering. For example, respondents may be asked to rate the degree of confidence that they have in their answers or to rate the sensitivity of questions. Response latency research is another way at testing the difficulty of questions for respondents. It measures “the time elapsed between the presentation of a question and the indication of a response” (Forsyth and Lessler, 1991, p. 401) and is based on the assumption that questions which require respondents to dig into their memories will have longer response latencies. The final targeted method, qualitative timing, is similar to response latency, but instead of using special equipment to measure the exact time, an observer codes the interval between the question and the response into categories.

Group methods bring several people together to discuss topics of interest or to complete experimental versions of a questionnaire in a controlled setting. Group methods include focus groups, group interviews, and group experiments. One of the reasons group formats are important is the social factors that distinguish group tasks from other laboratory tasks. Focus groups are probably the best known format of group interviews. They may represent subgroups or cross-sections of a survey’s target population. Their task may be to complete a questionnaire and then discuss it, or they may be used to gather information on how people think about specific issues. In group experiments, respondents usually complete experimental versions of a questionnaire (Forsyth and Lessler, 1991).

In summary, all of these methods provide more information about the question-answering process than can be obtained through simply asking the survey questions and recording the answers. The methods differ according to their timing and the amount of control the researcher has over what is observed. The task timing may be either concurrent, immediately after the respondent answers the questions, delayed, or unrelated. Either the respondent decides what information will be observed, as in concurrent think-aloud interviews, or response data are independently processed by the researcher as in behavior coding. All cognitive laboratory methods are basically qualitative studies even though some of the methods do collect quantitative information.

NCES has used several cognitive research techniques, including interactional behavior coding, a form of expert evaluation; concurrent think-aloud interviews, a form of expanded interviews; and paraphrasing, a targeted method. Chapter 7 of this report summarizes selected NCES cognitive studies using these methods.

CHAPTER 2
Profile of NCES Reinterview Studies

NCES reinterview studies vary across surveys from small sample reinterview studies conducted as part of a survey field test to larger samples that range between 1 and 11 percent of the full scale study. These studies have been used for two major purposes

- Identifying specific questions that may be problematic for respondents and result in low reliability and validity during field test studies
- Quantifying the magnitude of the measurement error during full-scale studies

When the purpose of the reinterview is specifically to gain insight into the adequacy of questions, conducting the reinterview as part of the field test allows time to change the questions prior to the full scale study. Of course, it is not certain that questions changed as a result of a field test reinterview study are without problems. But, if there are questions dealing with concepts difficult to measure, reinterview studies conducted as part of the field test can give at least a limited indication of their adequacy.

Reinterview studies conducted as part of the full scale NCES studies have emphasized estimating and reporting the response variance and/or bias for selected items.

Many of the NCES reinterview studies examined in this report were conducted using a CATI procedure in a centralized setting (NHES, B&B, BPS, NPSAS:96). Since the CATI interviews are closely monitored, it is unlikely that a telephone interviewer could invent or falsify interviews. Therefore, this aspect of measurement error was not part of the focus of the NCES reinterview studies reviewed in this report.

Other NCES surveys, including Baccalaureate and Beyond (B&B), the Beginning Postsecondary Students Longitudinal Survey (BPS), High School and Beyond (HS&B), the National Household Education Survey (NHES), the National Postsecondary Student Aid Study (NPSAS), the National Study of Postsecondary Faculty (NSOPF), Recent College Graduates (RCG), the Schools and Staffing Survey (SASS), and the Teacher Follow-up Survey (TFS) conduct not only reinterview studies, but in some cases, a combination of methodologies to evaluate and control measurement error.

HS&B, NHES, RCG, SASS, and TFS conducted reinterview studies as part of the full scale study. NPSAS and NSOPF conducted reinterviews as part of the field test for the study. B&B and BPS conducted reinterviews as part of both full scale and field test studies.

Several NCES surveys have conducted reinterview studies for more than one round or cycle of the survey, specifically B&B, BPS, NHES, SASS, and TFS. Most of the studies do not include the same items on subsequent rounds of the reinterview, however. The BPS reinterview studies, for example, are designed "to build on previous analyses by targeting revised or new items, and items not previously evaluated" (Pratt, Burkheimer, Jr., Forsyth, Wine, Veith, Beaulieu and Knepper, 1994, pp. 65-66). Most of the NCES reinterview studies were developed to estimate response variance, but some, such as the Adult Education component of NHES:95, included a response bias study.

Characteristics of NCES Reinterview Studies

Table 2 summarizes some of the characteristics, such as sample size, sampling percentage of the original sample, the response rate, and the primary purpose of each of the major reinterview studies that NCES has conducted. The *reinterview sampling percentage* is the percentage of the completed interviews in the original sample that were selected for the reinterview study and the *response rate* is the proportion of completed reinterviews to the number of completed original interviews targeted for reinterview.

Table 2. -- Reinterview sample size, sampling percentage of original sample, response rate, and primary study purpose for studies (*studies in italics could not be included in this report)*

	Reinterview sample size	Sampling percentage[1]	Reinterview response rate[2]	Primary study purpose
NHES				
1991 Early Childhood Education	604	4%	88%	Response variance
1993 School Readiness	977	9%	90%	Response variance
1993 School Safety & Discipline	1,131	6%	88%	Response variance
1995 Adult Education	1,289	6%	86%	Response variance
1995 Adult Education Bias Study	230	1%	90%	Response bias
RCG				
1991	583	4%	88%[3]	Response variance & bias
SASS				
1987-88 Administrator Survey	1,309	≅10%	87%	Response variance
1990-91 Administrator Survey	1,048	≅10%	94%	Response variance
1993-94 Administrator Survey	*1,154*	*≅10%*	*82%*	*Response variance*
1987-88 School Survey	1,309	≅10%	87%	Response variance
1990-91 School Survey	1,034	≅10%	91%	Response variance
1993-94 School Survey	*900*	*≅10%*	*62%*	*Response variance*
1987-88 Teacher Survey	1,126	≅1%	75%	Response variance
1990-91 Teacher Survey	980	≅1%	83%	Response variance
1993-94 Teacher Survey	*1,261*	*≅1%*	*73%*	*Response variance*
1993-94 Library Survey	1,343	23%	72%	Response variance
1989 TFS	1,497	18%	81%	Response variance
1992 TFS Current (stayers & movers)	678	14%	93%	Response variance & bias
1992 TFS Former (leavers)	747	49%	92%	Response variance & bias

Table 2. -- Reinterview sample size, sampling percentage of original sample, response rate, and primary study purpose for studies (*studies in italics could not be included in this report*—Continued

	Reinterview sample size	Sampling percentage[1]	Reinterview response rate[2]	Primary study purpose
B&B				
1993-94 Field Test	200	13%	53%[3]	Response variance
BPS				
90/92: 1st Followup Field Test	125	11%	92%[3]	Reliability
90/92: 1st Followup	[4]	[4]	[3,4]	
90/94: 2nd Followup Field Test	113	11%	84%[3]	Reliability
NPSAS				
1992-93 Field Test	[5]	[5]	[5]	Reliability
1996 Field Test	252	7%	91%	Reliability
1996	*250*	*<1%*	*94%*	*Reliability*
NSOPF				
1993 Faculty	--	≅24%	--	Reliability

[1] The sampling percentage was calculated as the reinterview sample size divided by the number of completed interviews in the original survey.

[2] The reinterview response rate was calculated as the number of completed reinterviews divided by the reinterview sample size, where the reinterview sample is a subsample of the eligible original completed interviews.

[3] A reinterview sample was selected from which only a targeted number needed to be completed.

[4] We have incomplete information on the BPS 1st followup. The methodology report states that 191 sample members participated in the reliability reinterview; 9,011 initial sample members were fully (8,495) or partially (516) interviewed for the 90/92 followup.

[5] We have incomplete information on the NPSAS 1992-93 field test reinterview sample size. The methodology report states there were 7,417 eligible student records in the field test and that field test reinterviews were conducted with 237 students.

SOURCE: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. U.S. Bureau of the Census.

NCES reinterview studies typically attempt to reach a target number of reinterviews or to reinterview a certain percentage of the original sample. For example, the 1990-91 SASS reinterview study had the goal of reinterviewing 10 percent of the School and Administrator samples and 1 percent of the Teacher sample, resulting in a reinterview sample of just over 1,000 for each of its components. NCES may oversample to ensure that the target number or certain percentage of the original sample is reached. For instance, the RCG:91 reinterview study had a target number of 500 completed reinterviews. To make it easier to reach that number within the short time frame of the reinterview study, the sample size was made almost 17 percent larger (583).

Methodological Issues in NCES Reinterview Studies

Various methodological issues besides sample size and response rates may affect the reinterview results, including eligibility, time lag, the reinterview instrument itself, reinterview mode, and respondent burden. Each of these issues involve trade-offs.

Eligibility

Eligibility requirements are often stipulated for selection into the reinterview study. For example, to be eligible for the RCG:91 reinterview, the respondent must have been a bachelor's degree recipient, a graduate who had never refused to participate and who was interviewed for the main survey between August 15 and September 30. These eligibility requirements excluded respondents interviewed in the first 3 weeks of data collection (when interviewers were less familiar with the survey) and, by establishing a cutoff date, ensured that at least 2 weeks had elapsed between the original and the reinterview (Brick, Cahalan, Gray, Severynse and Stowe, 1994).

Response Rates

Reinterview response rates are very important in determining an accurate measure of response error. Reinterview nonresponse affects the quality of the estimates just as nonresponse at the time of the original interview affects the estimates developed for the survey parameters, as well as estimates of their standard errors. A large segment of nonresponse at the time of reinterview increases the possibility that the estimates developed for response variance and response bias are themselves biased. Since the sampled units have already been sensitized by the first interview, the assumption that nonresponse is random becomes harder to justify.

Time Lag

The lag between the original interview and the reinterview can vary from a few days to several months. Ideally, the more likely it is that some characteristics will change between the interviews, the more desirable a short time lag. However, conditioning—when respondents remember their first answer at the time the reinterview is conducted—threatens the independent replication assumption. O'Muircheartaigh (1986) suggests regression analysis to test independence between the two interviews. Experience suggests wording questions to refer to specific time periods.

It is difficult to determine the actual time lag between the original and the reinterview for most of the NCES surveys. It is usually stated as a range such as "the reinterviews were conducted in October and November, about 4 to 6 weeks after the original interview" (Brick, Cahalan et al., 1994, p. 3-3). Early Childhood Education (ECE) reinterviews for NHES were designed for 14 days after the completion of the original ECE interview, but they were actually completed between 14 and 20 days after the original interview. Reinterviews in NPSAS were conducted between one and three months after the original interview.

Reinterview Instrument

The reinterview instrument is a subset of the original questionnaire, but the question wording is almost always identical between the original and the reinterview. In some cases, however, adjustments were made to the question wording (e.g., compare the NSOPF 1992-93 field test questionnaire and reinterview questionnaire). This is most often the case if the reinterview is conducted as part of the field test and not as part of the full scale study, or if the purpose of the reinterview study is to estimate response bias and includes a reconciliation.

Mode

Most surveys and their corresponding reinterviews are conducted using the same mode (i.e., telephone/CATI). However, the Schools and Staffing Survey (SASS) is conducted by mail with telephone/CATI followup. Here, conducting all the reinterviews by telephone violates survey error model assumptions that require the reinterview to be an independent replication of the original interview in order to estimate response variance accurately. Therefore, SASS included research in its 1991 reinterviews to determine the impact mode change might have on data quality. Most of the mail respondents were reinterviewed by mail and the telephone follow-up cases were reinterviewed by telephone. Reinterviews conducted by mail showed lower response variance than the telephone reinterviews; however, this was not a randomized experiment. For details on this research, see the appendix.

Bushery, Royce, and Kasprzyk (1992) suggest four possible reasons why the reinterviews completed by mail show lower response variances than the reinterviews completed by telephone. First, only respondents who answered the original survey by mail were eligible for the mail reinterview. These respondents were more likely to be more cooperative and answer the questions more carefully in both interviews. Second, respondents interviewed by mail may take more time than those interviewed by telephone to look up the answers to questions from records or may go through a more careful, lengthy thought process to provide the needed facts. Respondents interviewed by telephone may not feel free to take the time to look up records while the interviewer is waiting on the phone. Third, mail respondents may leave more difficult or uncertain questions blank. Telephone interviewers may manage to obtain answers to a difficult question, but the answers may be unreliable. Fourth, mail respondents may photocopy the original questionnaire after completing it and refer to their original answers when completing the mail reinterview. However, the authors feel this last explanation would have only accounted for a small part of the mail-mail versus telephone-telephone differences; they consider some combination of the first three explanations is the most plausible.

Respondent Burden

The possible effects of respondent burden have long been an important issue when a questionnaire is designed. Past experience with mail questionnaires has led researchers to believe that the longer the questionnaire, the higher the nonresponse rate—assuming all other factors are equal (Sudman and Bradburn, 1982, pp. 226-27). However, Frey (1989, p. 258) suggests that respondent burden may not be as significant a problem as originally thought, at least for telephone questionnaires. He cites a Bureau of Social Science Research report that found the factors of time demand, recall, and repeat administration had no effect on response patterns to telephone questionnaires. While the exact role of respondent burden is still unclear, consideration of respondent burden is still important when designing questionnaires for reinterview purposes, when the overall sample size is limited by the number of initial respondents. Moreover, the respondents often do not perceive the real purpose of the reinterview; they ask, "Why do I have to answer the same questions again?" Thus, it is all the more important that the questionnaire be as brief as possible.

Lessons Learned from NCES Reinterview Studies

Questionnaire construction and question formulation lie at the heart of survey design. NCES reinterview studies illustrate the variety of measurement errors embedded in the art and science of survey design. Many of the measurement errors discussed here can be eliminated through researcher "due diligence" and creativity. Others, such as the constraints imposed by the fallibility of the human memory and the presence of diverse cultural filters, are not so easily addressed. The reinterview studies presented in chapters 3 and 4 illustrate the high response variance/low reliability and high response bias/low validity that occur in even the most thoughtfully designed surveys as a result of lack of focus and specificity of questions, ambiguity or vagueness of language, sensitivity of the issues raised, and faulty assumptions of respondent knowledge. Chapter 3 addresses the statistical approach to measurement error through the analysis of response variance and response bias; chapter 4 explores the psychometric approach to this subject through analysis of reliability and validity.

While a variety of survey pretesting approaches through the use of cognitive laboratories, behavior coding, and interviewer debriefing will certainly alert survey researchers to "difficulties" with particular questions and enable them to reformulate or eliminate them for use in the final survey, often such troublesome questions remain and come to light only in reinterview studies. Similarly, response bias may only be detected through analyses of reinterview data. In studying response bias, researchers should consider the use of "intensive reinterviews" under which interviewers trained in special techniques seek to identify particularly sensitive questions.

To optimally employ the reinterview in assessing response variance, the NCES experience reviewed in these chapters clearly shows that the researcher should strive to **reproduce as completely as possible the conditions of the original survey**. Researchers should employ the identical data collection mode in their reinterviews that they employed in their original surveys. If the original survey was mailed to respondents and self-administered, then the reinterview should follow this same procedure. If the survey was conducted by telephone, then the reinterview should be conducted by telephone, as well. (In telephone reinterviews, NCES data indicate that best results occur when the interviewer is initially unaware of the responses from the original interview, next compares initial and reinterview responses, and then seeks to reconcile any differences through either a third call reconciliation, or through a CATI-mediated reconciliation process at the time of the telephone reinterview).

In conducting reinterviews, researchers should also be careful to draw sufficiently large samples so as to ensure the precision of estimates. If researchers are unable to include the original interview in its entirety in the reinterview (which is the optimal approach), then in selecting questions for inclusion in a reinterview, they should be attentive to the context of questions within the original questionnaire—removing certain questions from their original context may well transform the respondent's answer.

The NCES reinterview studies indicate that response bias may be reduced if respondents are informed of the overall intent of the survey in which they are asked to participate. The studies further indicate that the more factual and direct the survey question, the more recent or current the data requested, and the more circumscribed the choices provided, the lower the response variability between interviews. Response variance was reduced if questions regarding use of time on particular tasks were tied to a specific time frame; clear definition of any time periods addressed in surveys was also seen as improving response reliability. Definition of the terms employed in a survey was also valuable in reducing response variance.

Questions asking respondents their views/opinions on issues or calling for more complex responses than a simple statement of facts, resulted in higher response variance. Questions exploring respondents' satisfaction with particular services or programs are also subject to significant variance/low reliability across studies. Including two distinct questions within one larger question appears to confuse respondents, leading to high response variance. Similarly, questions employing abstract concepts requiring respondents to engage in a more elaborate cognitive process elicited greater response variance, as did questions which assumed respondent knowledge about the identity or workings of various administrative structures.

Interviewer training focusing on areas of potential difficulty for respondents or highlighting items of particular importance to researchers, supplemented by monitoring and supervision of interviewers during the survey process, itself, appears to reduce response variance and increase the reliability of survey results. To aid researchers in assessing survey findings, the respondents could be asked screening questions concerning their degree of confidence in their responses to a particular question or set of questions.

Detailed Findings from NCES Reinterview Studies

Chapters 3 and 4 will describe in more detail the reinterview studies conducted by NCES that were reviewed for this report. Reinterview studies that estimated response variance and response bias will be given first, followed by reinterview studies that estimated reliability. Thus, chapter 3, Reinterview Studies: Simple Response Variance and Response Bias, will look at reinterview studies conducted on components of the National Household Education Survey (NHES) in 1991, 1993, and 1995; Recent College Graduates in 1991; the Schools and

Staffing Survey (SASS) in 1987-88, 1990-91, and 1993-94 (Library Survey only); the Teacher Followup Survey in 1989 and 1992; and the Baccalaureate and Beyond Longitudinal Study (B&B) 1993-94 field test. Chapter 4, Reinterview Studies: Reliability and Validity, looks at reinterview studies conducted on the Beginning Postsecondary Students Longitudinal Study (BPS) 1992 and 1994 field tests and 1992 full scale study, the National Postsecondary Student Aid Study (NPSAS) 1992-93 and 1996 field tests, and the National Study of Postsecondary Faculty (NSOPF) 1993 field test.

CHAPTER 3
Reinterview Studies: Simple Response Variance and Response Bias

Several reinterview studies conducted by NCES examined simple response variance, and response bias, as part of an examination of measurement error. This is what was done at NCES by the National Household Education Survey (NHES), Recent College Graduates (RCG), and Schools and Staffing Survey (SASS) reinterview studies. These reinterview studies were conducted during the full scale study. In addition, the Baccalaureate and Beyond Longitudinal Study (B&B) 1993-94 Field Test conducted a reinterview study from which estimates of simple response variance were examined. Before looking at the results obtained by these studies—and not all of the reinterview studies for even these four surveys could be included—it may be helpful to discuss in more detail the theoretical development of the estimators of simple response variance and response bias.

Since it is difficult to directly estimate measurement error in a survey setting, models have been proposed to represent the most important structures of the error process. In essence, the models assume that the correct answer to a question may not actually be reported due to any number of sources of error. Thus, a measurement error model attempts to reflect the general nature of the errors, taking into account the data collection process. For example, a model might assume that in identical, independent replications of the data collection, the value reported would, on average, be the same as the correct value.

A measurement error model is useful only if it includes the major components of error. In addition, the model assumptions should be true. For example, if the model assumes that errors are independent, but they are actually highly correlated, then the estimates of the model parameter may be misleading. The measurement error models examined below are essentially those originally proposed by Hansen, Hurwitz, and Pritzker (1964).

Simple Response Variance Model

Measurement error in a survey response can be viewed as a random variable arising from the sampling of a hypothetical error population. Thus, the response to a particular survey item is the result of two "stages" of random sampling: the sampling of an individual unit from a population of individuals and the sampling of errors within individuals from an infinite hypothetical populations of trials. We will consider a simple model which assumes that the correct value differs from the observed value by an unobserved additive error term. For unit *i* (*i* = 1, 2,..., n) and time trial *t* (*t* = 1,2,...), the assumed model is

$$y_{ti} = \mu_i + \varepsilon_{ti} \qquad (3.1)$$

where y_{ti} is the observed value at trial t for the ith respondent, μ_i is the unobserved correct value for the ith respondent, and ε_{ti} is the unobserved error at trial t for the ith respondent. To complete the specification of the model, we further assume

$$\begin{aligned}
&E(\varepsilon_{ti}|i) = 0 \\
&Var\left(\varepsilon_{ti}|i\right) = \sigma_i^2 \\
&E(\sigma_i^2) = \sigma^2 \\
&Cov\left(\varepsilon_{ti}\varepsilon'_{t'i'}\right) = 0 \;\; for \;\; i \neq i' \text{ or } t \neq t'
\end{aligned} \qquad (3.2)$$

The model implies that there are no systematic biases in the estimates (the mean of the errors is zero, and the variances are not equal) and the errors are not correlated. The latter means that the errors in an observation for a trial do not affect other observations in the same trial and the errors across trials for the same observation are uncorrelated.

Under the measurement error model specified by (3.1) and (3.2), the ordinary measure of the precision of the estimate differs from the usual expression. For example, in a simple random sample, the variance of a mean, can be calculated over all possible trials and for all samples as

$$\begin{aligned}
Var(\overline{y_t}) &= Var(\overline{\mu}) + \frac{\sigma^2}{n} \\
&= SV \; + \; SRV
\end{aligned} \qquad (3.3)$$

where

$$\overline{y_t} = \frac{1}{n}\sum y_{ti} \, , \overline{\mu} = \frac{1}{n}\sum \mu_i \, , \text{ and n is the sample size.}$$

The first term on the right hand side of (3.3) is the **sampling variance (*SV*)** of the estimate. The SV is the ordinary variance of the estimate if there is no measurement error. The second term on the right hand side of (3.3), often called the **simple response variance (*SRV*)** of the estimate, is the variability of the responses to the item averaged over conceptual repetitions of the survey under the same conditions.

Sometimes expression (3.3) gives the erroneous impression that the usual methods of estimating the variance of an estimate such as the mean must be modified to account for the additional term. However, Hansen, Hurwitz, and Pritzker (1964) showed that the ordinary estimate of the variance at trial t, written as

$$\frac{s_t^2}{n} = \frac{1}{n(n-1)}\sum (y_{ti} - \overline{y_t})^2 \qquad (3.4)$$

is an unbiased estimate of $Var(\overline{y_t})$, where the expectation is taken over all possible trials and all samples.

Thus, if the assumptions of this measurement error model hold, the estimates from the survey will be unbiased and the estimated variance will include both the *SV* and the *SRV*. Despite this, it is still valuable to estimate the relative contribution of the *SRV* to the random error because the *SRV* can be reduced by different data collection methods (e.g., ways of phrasing the questions). If the *SRV* is a large fraction of the total error, then methods to reduce it can significantly reduce the total error in the estimates.

The model has potential weaknesses when there is a correlation between the responses in the original interview and the reinterview. A correlation may exist because the respondent recalls answers to the original question or is somehow influenced by the original survey. Another reason this model might not be appropriate is the correlation between the responses of the sampled units that were conducted by the same interviewer.

Other potential weaknesses in the model arise from invalid assumptions about the means of the error term. For example, the error term may not have zero mean over replications of the survey. These types of failures are likely to be of greatest concern for categorical data.

Despite its inherent limitations, model (3.1) and (3.2) can provide a useful approximation of the contribution of measurement error. To produce these estimates in a reinterview setting, the parameters of the model must be estimated from the original and the reinterview data. The trials described in model (3.1) and (3.2) are defined so that t=1 is the original interview and t=2 is the reinterview.

Typically, there are two statistics used by NCES reinterview studies to examine aspects of the reliability of reporting:

- Gross Difference Rate (*GDR*)
- Index of Inconsistency (*IOI*)

Gross Difference Rate. Under the assumptions of model (3.1) and (3.2), the response bias is defined to be zero and is not estimated. The SRV can be estimated by the **gross difference rate (*GDR*)**, where *GDR* is

$$GDR = \frac{1}{n}\sum\left(y_{1i} - y_{2i}\right)^2 \qquad (3.5)$$

Thus, the gross difference rate is the average squared difference between the original interview and the reinterview responses.

Under model (3.1) and (3.2), the *GDR* can be rewritten as

$$\begin{aligned} GDR &= \frac{1}{n}\sum\left(\mu_i + \varepsilon_{1i} - \mu_i - \varepsilon_{2i}\right)^2 \\ &= \frac{1}{n}\sum\left(\varepsilon_{1i} - \varepsilon_{2i}\right)^2 \end{aligned}$$

$$= \frac{1}{n} \sum (\varepsilon_{1i}^2 + \varepsilon_{2i}^2 - 2\varepsilon_{1i}\varepsilon_{2i}) \qquad (3.6)$$

Now, taking the expectation of (3.6) over all possible trials gives

$$E(GDR) = 2\sigma^2 \qquad (3.7)$$

Thus, when a reinterview is conducted for a sample of individuals, the *GDR* is an unbiased estimate of $2\sigma^2$. In other words, the *GDR* divided by 2 is an unbiased estimate of *SRV.* These results are based on simple random samples. To hold for more complex designs the estimators must be revised to include the sample weights.

Index of Inconsistency. A natural estimator of the proportion of the random error that is associated with measurement error is given by the **index of inconsistency (*IOI*)**

$$IOI = \frac{SRV}{SRV + SV} \cong \frac{GDR}{2s^2}, \qquad (3.8)$$

where s^2 can be estimated by the average of the ordinary variance estimates as defined in equation (3.4) for the original and reinterview. Other estimators of the denominator of *IOI* are possible. The *IOI* obtains values between 0.0 and 1.0, inclusive. Estimates of *IOI*, on the other hand, can go over 1.0. The estimate can also go over 1.0 even when the assumptions are met, though it rarely does.

Special Case of Categorical Variables. For characteristics that have exactly two possible outcomes, the **gross difference** is equal to the percentage of cases reported differently in the original interview and the reinterview. The ***GDR*** is the ratio of the gross difference divided by the estimated total number of cases.

With **dichotomous variables**, the estimators are often presented in a very simple table showing the original and reinterview estimates (or counts if the design is simple random sampling). Table 3 shows the general format for reporting outcomes by the original interviews and reinterviews for dichotomous variables which take the value 1 for cases with a characteristic and the value 0 for cases without the characteristic.

Table 3. -- Interview by reinterview table

		Original Interview		
		Number of cases with characteristics	**Number of cases without characteristics**	**Total**
Reinterview	**Number of cases with characteristics**	a	b	a + b
	Number of cases without characteristics	c	d	c + d
Total		a + c	b + d	n = a + b + c + d

From tables formatted in this fashion, the percent *GDR* takes on a very simple form:

$$GDR\% = 100 \times \frac{b+c}{n} \qquad (3.9)$$

Thus, the *GDR%* is the percentage of cases that were reported differently in the original and reinterview surveys. It is equal to the percentage of cases reported as having a characteristic in the original interview but not having it in the reinterview, plus the percentage of cases reported as not having the characteristic in the original interview but having it in the reinterview. That is, the *GDR* is the ratio of the estimated number of cases misclassified in the original interview divided by the estimated total number of reinterview surveys.

Similarly, from table 3, the percent ***IOI*** also takes on a very simple form:

$$IOI\% = 100 \times \frac{b+c}{2np(1-p)} \qquad (3.10)$$

where p is $\frac{a+c}{n}$

Equations (3.9) and (3.10) easily be seen to be a special case of equations (3.5) and (3.8) respectively when the only two valid responses are zero and one.

For categorical variables with more than two response values, the expressions for the *GDR* and *IOI* still can be written in forms that are simpler than expressions (3.5) and (3.8). For example, the *GDR* is the sum of the off-diagonal elements of the original interview by reinterview table divided by the total for the table, expressed as a percentage. The *IOI* can be written as an average of the indices for the 2 x 2 sub-tables, often called the **L-fold index of inconsistency**. The U.S. Bureau of Census (1985) defines these terms more explicitly.

Response Bias Model

A different model can be formulated if the original response has a systematic error or bias that does not occur in the reinterview. The consequences of assuming that the second trial or the reinterview has less error than the original survey response are considered below.

This new model retains the simple additive error structure of (3.1), but the assumptions on the error terms are different, since $\varepsilon_{2i} = 0$. The following results follow immediately from the assumptions about the error terms

$$
\begin{aligned}
&E(\varepsilon_{ti}|i) = \beta_i \neq 0 \quad \textit{for } t = 1,\\
&Var(\varepsilon_{ti}|i) = \sigma_i^2 \quad \textit{for } t = 1\\
&Var(\varepsilon_{ti}|i) = 0 \quad \textit{for } t = 2\\
&Cov(\varepsilon_{ti}, \varepsilon_{ti'}) = 0 \quad \textit{for } \; i \neq i'
\end{aligned}
\qquad (3.11)
$$

Note that in this model, the error term for the first trial no longer averages to zero. The estimate based on the original interview could be subject to a **response bias,** where the bias is defined as

$$\beta_i = \frac{1}{N}\sum\left(y_{1i} - \mu_i\right). \qquad (3.12)$$

The response bias for the second trial is zero by assumption. The model specified by (3.1) and (3.11) will be called model (3.11)

In order to meet the conditions of model (3.11), the result from the reinterview should be free of measurement error. While this is not completely possible under the constraints of a reinterview, several different procedures have been proposed in the literature to obtain more accurate responses in the reinterview than were obtained in the original interview. These include using more experienced interviewers or supervisors, using improved data collection methods, using additional probing questions, and asking the respondent to reconcile the differences in responses.

Reconciliation is a means of improving responses. Since it is very unlikely that the reconciled responses are actually error free, they can be used to identify the expected direction of bias, and the relative amount of bias, but cannot provide precise estimates of the size of the bias. Furthermore, the reconciliation process does not detect consistent errors made in both the original and the reinterview.

If the reconciled interviews are free of measurement error, the *GDR* (computed as the difference between the original and the reconciled responses) no longer provides an unbiased estimate for the *SRV*. Using expression (3.6), it can be shown that the *GDR* is an overestimate of the *SRV* (Hansen, Hurwitz and Pritzker, 1964). Therefore, the *GDR* estimated using reconciled reinterview responses is an upper bound on the *SRV*.

Net Difference Rate. Of course, the main reason for doing the reconciliation is to provide at least a rough guide to the size of the response bias. An unbiased estimate of the response bias under model (3.11) is given by the **net difference rate (*NDR*)**, which can be written as

$$NDR = \frac{1}{n}\sum\left(y_{1i} - y_{2i}\right). \tag{3.13}$$

For the **binary case**, the net difference is the count of cases with a characteristic as reported in the original interview and the count of cases in the reinterview. That is, *(a + c) - (a + b) = c - b*, using the terms in table 4. The *NDR* is the ratio of the net difference to the estimated total number of interviews and the *NDR*, expressed as a percentage, is

$$NDR\% = 100 \times \frac{c-b}{n}. \tag{3.14}$$

In the *NDR* calculation equal number of errors in opposite directions offset each other and the remaining non-offsetting part of the total error is counted. While in the *GDR* calculation there is no opportunity for one error to offset the other resulting in every error being counted.

For items with **numeric data**, the net difference rate is the average difference between the original and reinterview. Note, for expediency sake, items which are measured in constant linear units (e.g., number of hours) and are symmetric about the diagonal are often treated in the same manner as items with multiple categories.

While expression (3.13) is valid for quantitative and dichotomous variables, it is less justified when the responses are categorical, unordered data. For example, for a variable such as race, which takes on the value 1, 2, or 3 corresponding to white, black, and Native American. This expression for the net bias actually weights the responses. In this case, the difference between response categories 1 and 3 would result in a larger contribution to the net bias than the difference between 1 and 2. Since these are unordered responses, this approach is questionable. Because of this, the net difference rate for a few categorical, unordered response variables can be computed differently. The net difference rates can be computed without weighting the responses. In other words, the difference between white and black would count the same as the difference between white and Native American. However, for these types of unordered measures the net difference rate is more of a general indicator of offsetting error than a direct measure.

While the *NDR* computed based on the reconciled responses can be used to estimate the expected direction and magnitude of response bias, it does not have the same properties when computed using the unreconciled responses. The net difference rate computed from the unreconciled reinterview data can be used to examine whether the two interviews result in similar estimates. If the two interviews are independent, then the expected value of the net difference rate should be equal to 0. A high *NDR* suggests that the reinterview may not have replicated the original survey very well. This could result in the gross difference rate being an overestimate of the *SRV*.

Special Case for Dichotomous Variables. In discussing model (3.2), we mentioned some of the problems of using this model with categorical variables. We expand on that discussion below with particular attention to dichotomous variables that take on the value of one if the sampled unit has the characteristic and zero otherwise.

With a dichotomous variable, the conditions on the moments of the model (3.2) can be written in terms of the probabilities of misclassifying the sampled unit (falsely classifying the unit as having or not having the characteristic). Biemer and Forsman (1992) show that both the response bias and the *SRV* are functions of these probabilities of misclassification and the proportion of the units that have the characteristic. They show that the response bias is zero only under special conditions. For example, the response bias is not equal to zero when the misclassification (false positives and false negatives) are equal, except for characteristics held by exactly 50 percent of the population.

These results have implications for the interpretation of the response bias and the *SRV* for dichotomous variables. The assumption of zero response bias in (3.2) does not mean that the probability of misclassification is the same in both directions. Rather, it means the number of sampled units erroneously classified as having the characteristic will, on replications of the survey, equal the number of units erroneously classified as not having the characteristic.

The *SRV* is still estimated unbiasedly by half of the *GDR*, but it does not directly measure the probabilities of misclassification. Thus, the *IOI* is an estimator of the impact of misclassification errors on the estimates rather than a direct measure of the misclassification probabilities. The appendix in the U.S. Bureau of the Census report (1985) describes these issues in more detail and gives some tables to demonstrate these points.

GDR, *NDR*, and *IOI* Results in this Report

Gross and net difference rates and indices of inconsistency are examined for items studied in the NCES reinterviews. As described above, the uses of these statistics can be summarized in table 4.

Table 4. -- Uses of reinterview statistics, by type of reinterview response

	Type of reinterview responses	
Statistic	**Unreconciled**	**Reconciled**
Gross difference rate (*GDR*)	Measure of random error (simple response variance)	Model diagnostic
Net difference rate (*NDR*)	Model diagnostic	Measure of systematic error (response bias)
Index of inconsistency (*IOI*)	Ratio of simple response variance to total random error	

SOURCE: Derived from exhibit 3-2, Brick, Cahalan et al., (1994), *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates,* p. 3-14.

To aid the reader in the interpretation of these statistics, we have categorized the *IOI* results from NCES reinterview studies consistently throughout the chapter as follows:

- An *IOI* of less than 20 is **low** relative response variance;
- An *IOI* between 20 and 50 is **moderate** relative response variance;
- An *IOI* above 50 is **high** relative response variance.

The *GDR* is more difficult to interpret than the index of inconsistency. Large *GDRs* indicate serious response variance in the data. Unfortunately, a small *GDR* is no guarantee of good consistency. In a low-frequency category, even a small *GDR* can represent high response variance relative to total variance. Thus, when available, we will provide the proportion in category along with the *GDR*. Mean *GDRs* for categories of items have been calculated and can be used to compare the general reliability of responses among similar categories across surveys.

The *NDR* computed from the original and reconciled reinterview with a value close to zero indicates that the response bias is not very large. When available, *t*-test results are provided for items with *NDR* values significantly different from zero (i.e., in which the *NDR* estimates the bias to be statistically significant).

The remainder of this chapter will describe the methodology, design, and results of the following reinterview studies

- **National Household Education Survey (NHES)**
  - NHES:91 Early Childhood Education
  - NHES:93 School Readiness
  - NHES:93 School Safety and Discipline
  - NHES:95 Adult Education
  - NHES:95 Adult Education Bias Study
- **Recent College Graduates (RCG)** 1991
- **Schools and Staffing Survey (SASS)**
  - 1987-88 Administrator Survey
  - 1990-91 Administrator Survey
  - 1987-88 School Survey
  - 1990-91 School Survey
  - 1987-88 Teacher Survey
  - 1990-91 Teacher Survey
  - 1993-94 Library Survey
  - 1989 TFS Current (stayers & movers)
  - 1989 TFS Former (leavers)
  - 1992 TFS Current (stayers & movers)
  - 1992 TFS Former (leavers)
- **Baccalaureate and Beyond (B&B)** 1993-94 Field Test

For each reinterview study, tables and figures display unweighted summary statistics. Where the information is available, tables show the reinterview sample size, the number of completed reinterviews, and the response rate, as well as how the reinterview sample size compares with the number of completed interviews in the original survey (sampling percentage). The average rank assigned to subject area measurements in the tables uses the three rank categories of low, moderate, and high discussed above to describe the simple average of the *IOI* of all items included in that subject area. The numerical average unreconciled *GDR* for subject areas is shown in the figures. Finally, when the items with high *IOI* are identified, the item question wording follows along with the *IOI* value, and the standard error when available. In cases where a comparison of results from a previous year's reinterview study is presented, both *IOI* and *GDR* values are given along with the estimated percent and standard error when available. NHES:95 calculated just the *GDR*, but not the *IOI*. In this study, *GDRs* and associated estimated percents and standard errors are presented for items of concern.

National Household Education Survey (NHES)

The National Household Education Survey (NHES) is a data collection system designed to address a wide range of education-related issues. It provides descriptive data on the educational activities of the U.S. population and offers policymakers, researchers, and educators a variety of statistics on the condition of education in the United States. It collects its data through telephone interviews, using random digit dialing (RDD) and computer-assisted telephone interviewing (CATI) procedures. The sample is drawn from the noninstitutionalized civilian population in households having a telephone in the 50 states and the District of Columbia. In each NHES, between 45,000 and 65,000 households are typically screened to identify persons eligible for one of the topical components. Generally, each collection covers two topical components, and interviews are conducted with between 10,000 and 15,000 respondents for each component.

NHES full scale surveys were conducted in 1991, 1993, 1995, and 1996. The 1991 survey components were early childhood education and adult education. The 1993 survey components were school readiness and school safety and discipline. The 1995 NHES components repeated those of 1991, addressing early childhood program participation and adult education, although both components had been substantially redesigned to incorporate new issues and new measurement approaches. The 1996 NHES components were parent/family involvement in education and civic involvement. By repeating components, NHES can monitor educational activities over time (Nolin and Chandler, 1996).

Reinterview Studies

NCES conducted four comprehensive reinterview studies for the 1991, 1993, and 1995 full scale NHES surveys. The reinterview study for NHES:91 was administered only on the early childhood component. In NHES:93 both components underwent reinterviews, while only the adult education component was reinterviewed for NHES:95. Table 5, below, provides a summary of the NHES surveys and reinterview studies. [Note: Reinterviews were also done as a part of NHES:96. Results were not available in time to include in the report.]

Table 5. -- 1991, 1993, and 1995 NHES components and reinterview studies

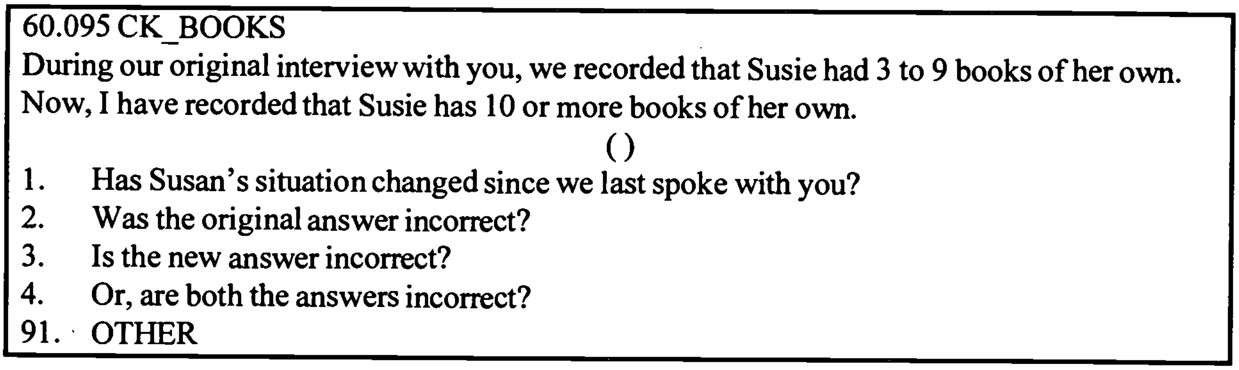
Studies	Component I	Component II	Reinterview
NHES:91	Early Childhood Education (ECE)	Adult Education (AE)	ECE
NHES:93	School Readiness (SR)	School Safety & Discipline (SS&D)	SR, SS&D
NHES:95	Early Childhood Program Participation (ECPP)	Adult Education (AE)	AE

The primary objectives for the NHES reinterview studies were

- To identify unreliable survey items
- To quantify the magnitude of the response variance for groups of items collected from the same respondent at two different times
- To provide feedback to improve the design of questionnaire items for future surveys

Interviewers reconciled some of the original and the reinterview responses for NHES 1991, 1993, and 1995 components using a CATI reconciliation screen. A typical example is depicted in figure 1.

Figure 1. -- CATI reconciliation screen



60.095 CK_BOOKS
During our original interview with you, we recorded that Susie had 3 to 9 books of her own.
Now, I have recorded that Susie has 10 or more books of her own.

( )

1. Has Susan's situation changed since we last spoke with you?
2. Was the original answer incorrect?
3. Is the new answer incorrect?
4. Or, are both the answers incorrect?

91. OTHER

SOURCE: Derived from exhibit A, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-11.

The NHES:91 and NHES:93 reconciliation process focused on determining the reason for discrepancies between the original interview and the reinterview. Results from the reconciliation were grouped into four categories

- The situation changed between the original interview and the reinterview
- The original data item was recorded or reported incorrectly
- The reinterview data item was recorded or reported incorrectly
- Both data items were recorded or reported incorrectly

To examine response bias more closely, NHES:95 included an *intensive* reinterview for the Adult Education component. This is described separately, as the NHES:95 Adult Education Bias Study.

All NHES reinterview studies have used gross difference rate (GDR) and net difference rate (NDR), and all except NHES:95 used index of inconsistency (IOI), as measures of response variability and response bias for critical items in the surveys. The reconciliation process and computation of NDR is more specifically used to measure response bias.

The NHES reinterview studies for each year are described separately. A short overview is followed by a discussion of the studies' methodology and design and a summary of the results of the response variance measurement calculations.

NHES:91 Early Childhood Education Reinterview

The Early Childhood Education (ECE) reinterview study targeted a sample of respondents who had previously completed the ECE interview. The questions asked were identical to those in the original survey, so the responses were used to measure response variance and provided feedback to improve the questionnaire design.

Methodology and Design

The NHES:91 reinterview was originally designed to be conducted 14 days after the completion of the original ECE interview (that is, the "extended" interview, not the screener); however, it took longer than 14 days to complete some reinterviews because all interviews in a household had to be completed before sampling for the reinterview could take place, and some reinterviews were done in less time so that all could be completed before data collection ended. The majority of reinterviews, about 73 percent, were conducted between 14 and 20 days after the original interview. About 10 percent were completed less than 14 days after the original interview.

The ECE component represented two groups of children: preprimary school children, not yet enrolled in first grade, and primary school children, enrolled in first grade or beyond. The sampling within these two groups was proportional to their representation in the full sample, a useful means of avoiding differential sampling of the children within the groups. The reinterview sample size was 604, or 4 percent of the full sample; the response rate was 88 percent (see table 6). The items selected for reinterview were substantively important and not highly time dependent. They concentrated on care and education arrangements for the preprimary child, and on enrollment characteristics for primary school children.

Table 6. -- NHES:91 ECE reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
ECE	604	534	88%	4%	Response variance

SOURCE: Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report*, pp. D-8 and D-9.

The NHES:91 reinterview study was conducted using the same CATI methods as the original interview. These methods provided an opportunity to control access to earlier responses for independent reinterviews. The entire reinterview was conducted first: interviewers read identical items to the same parent/guardian who responded to the original interview. This person had been identified as the person who knew the most about the child's care and education for the original ECE interview. After all the items for the reinterview were asked, the original and reinterview responses were compared electronically. The CATI system then produced a series of edit-check screens to resolve differences between the initial and the reinterview responses. Until this time, the interviewer was unaware of the responses from the original interview. For any items which had different responses, the interviewer asked whether the original or reinterview response was correct. This procedure accomplished the reinterview study objective to provide information on the reasons for the differences without underestimating response errors (Brick, Collins, Celebuski, Nolin, Squadere, Ha, Wernimont, West, Chandler, Hausken and Owings, 1991, pp. D-9 and D-10).

Summary of Results

The NHES:91 reinterview results suggested that the ECE interview measured some variables with relative success, but it also revealed items that needed to be handled carefully when tabulating findings and for which alternative methods of collection should be considered (Brick et al., 1991, p. D-25).

Both the preprimary and primary school children reinterviews included questions on current enrollment (whether the child was attending school and, if so, what grade) and home environment (reading and television habits). The two items worth noting were

- P19/E36 How often do you or other family members read stories to (child)? [never/several times a year/several times a month/at least three times a week/every day]
- P22/E40 How many hours each day does (child) watch television or videotapes? [do not have a TV/less than one hour per day/hours per day watching TV or VIDEOS]

Brick et al. felt the IOI of 42.0 for the television question might be due to the "general ambiguity in the item, the crude measurement scale (whole hours) relative to the internal variability in the item, and differing circumstances" (1991, p. D-20). The reading question also raised concern (GDR 23.3, IOI 33.5). Brick et al. concluded that this item "also had specific categories and...had to be classified into...precoded categories. This type of classification may be difficult for the respondent to do accurately, which would result in the...measurement errors" (1991, p. D-20).

Table 7 shows the average rank of subject area measurements for preprimary and primary school children items. All of the seven enrollment items had low IOIs. For the four home environment items, there was one low and three moderate IOIs. Note that no items had high IOI.

Table 7. -- NHES:91 average rank of subject area measurements for preprimary and primary school children items

Subject areas	IOI	Items with high IOI (above 50%)
Current enrollment	Low	--
Home environment	Moderate	--

SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report*, p. D-17.

The preprimary reinterview subject areas included relative care, nonrelative care, daycare, prekindergarten/nursery school, and kindergarten. Specific items asked about where the child was cared for, and by whom. If the child went to a daycare center, items asked if the center had instruction, if it was a Head Start program, and how often the parent had contact with the care provider. The prekindergarten/nursery school questions were similar to the daycare questions. Kindergarten questions asked whether the child went to one or two kindergartens, if it was full- or part-day, etc. The items that concerned researchers were

- P3G Is (child's) program at this daycare center a Head Start Program? [NA or missing/yes/no] (GDR 12.9, IOI 41.3)
- P4G Is the program at the (first/next) (nursery school/prekindergarten) a Head Start Program? [NA or missing/yes/no] (GDR 13.5, IOI 43.9)

These measurement errors combined with an examination of the children's characteristics indicated response problems. It was decided that children may have been incorrectly identified as Head Start children, possibly because parents were unsure what constituted a Head Start program, possibly because children were enrolled in programs that also enrolled Head Start children. A related question (P4TYPE: Is the (first/next) program a nursery school, prekindergarten, or Head Start program?) had a moderate IOI.

Two other related items in the preprimary series were noted in the report

- P3J How often do you talk with (child's) primary care provider or teacher about how (he/she) is doing at this daycare center? [NA or missing/every time or most times child goes/frequently/ occasionally/rarely or never/no experience or newly enrolled]
- P4J How often do you talk with (child's) primary care provider or teacher about how (he/she) is doing at (this) (nursery school/ prekindergarten/Head Start program)? [NA or missing/every time or most times child goes/frequently/occasionally/rarely or never]

P3J had a GDR of 18.2 and an IOI of 25.4. P4J had a GDR of 18.4 and an IOI of 27.9. It was felt that parents might have been unsure what constituted "talking" with the provider. Some might have included general conversation made when picking up their child, others might have restricted the meaning to formal discussions.

The average rank of the IOIs for the subject areas are shown in table 8. More than half (13) of the preprimary items had moderate IOIs, and 3 had a high IOI.

Table 8. -- NHES:91 average rank of subject area measurements for preprimary school children items

Subject areas	IOI	Items with high IOI (above 50%)
Relative care	Low	--
Nonrelative care	Moderate	--
Daycare	Moderate	P3B
Prekindergarten/nursery schools	Moderate	P4F
Kindergarten	Moderate	P15

SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-17.

The three items with high IOIs are

- P3B How many different daycare centers does (child) currently go to? [# of daycare centers currently attend] (IOI 51.5)
- P4F Does the (nursery school/prekindergarten/Head Start program) (child) goes to have an educational program? [NA or missing/yes/no] (IOI 50.9)
- P15 Is this (child's) first or second year of kindergarten? [first/second] (IOI 101.0)

The primary reinterview subject areas included grade levels, retention, current school year, and prior arrangements for child care/early education. Grade levels items asked about the grade the child had attended the previous year, whether the child had ever attended kindergarten, and the child's age when starting kindergarten. Retention items asked if the child had attended kindergarten for 1 or 2 years, changed schools, or repeated grades. Current school year items asked whether the child attended private or public school, how many days the child had homework, how often the parent talked to the child about school, and whether the parent contacted the teacher. Prior arrangements repeated some of the questions asked of preprimary children: whether the child had ever received care from a relative or nonrelative and whether the child had gone to a daycare center, nursery, prekindergarten, or Head Start program.

Three items with moderate IOIs drew researchers attention

- E3 In what month and year did (child) start kindergarten? [year child started kindergarten]
- E18 How often do you [or (child's) (other parent)] talk with (him/her) about (his/her) experience in school? [not at all/rarely/occasionally/ regularly]
- E33 Did (any of) the (daycare centers) (or) (early childhood programs) (child) went to have an educational program? [yes/no]

E3 had an IOI of 48.0 and a GDR of 27.4. Brick et al. (1991) conjectured that parents had to mentally construct their answer and that could have contributed to much of the problem. E18 had an IOI of 28.7 and a GDR of 2.7 and E33 had an IOI of 31.5 and a GDR of 5.3. Such low GDRs indicate that there was no substantial problem.

The average rank of the GDRs and IOIs for the subject areas are shown in table 9. Most of the primary items had low IOIs, but six had moderate IOIs, and one had a high score (50.2).

Table 9. -- NHES:91 average rank of subject area measurements for primary school children items

Subject areas	IOI	Items with high IOI (above 50%)
Grade levels	Moderate	E2
Retention	Low	--
Current school year	Low	--
Prior arrangements	Moderate	--

SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-18.

The one item with a high IOI was

- E2 (Before starting first grade) did (child) ever attend kindergarten? [yes/no] (IOI 50.2)

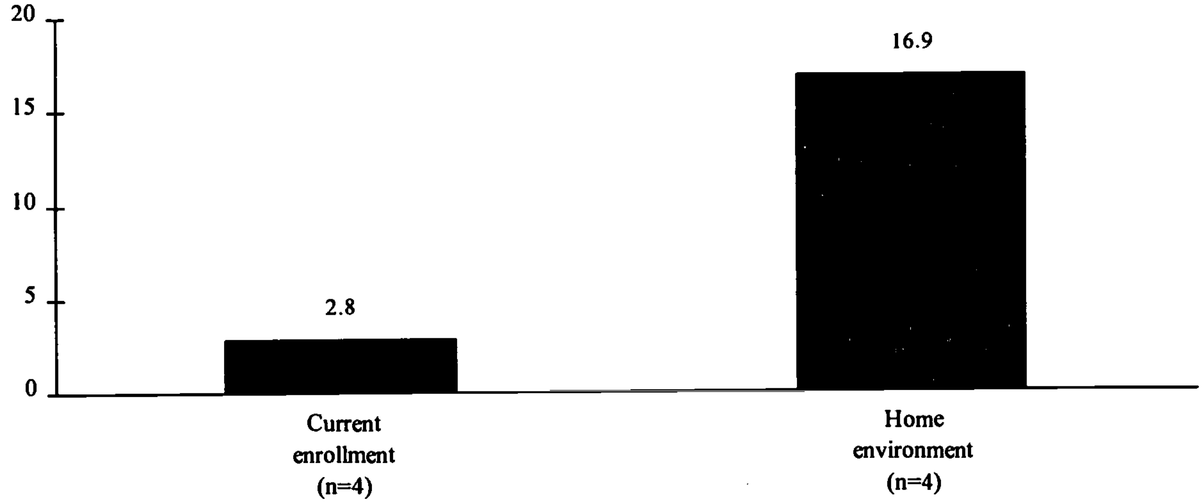
Overall, the results showed that about three-fourths of the reinterview items for NHES:91 had GDRs that were less than 10 percent. Only six items had GDRs greater than 15 percent. Ninety-one percent of the items had low to moderate IOIs while only 9 percent had high response variability.

Brick et al. felt that only two items in the entire reinterview study needed any substantial revision

- How many hours spent watching TV? (home environment)
  The lack of a time frame made this item ambiguous.
- Is the daycare center a Head Start program? (daycare centers)
  Children may have been incorrectly identified as Head Start children.

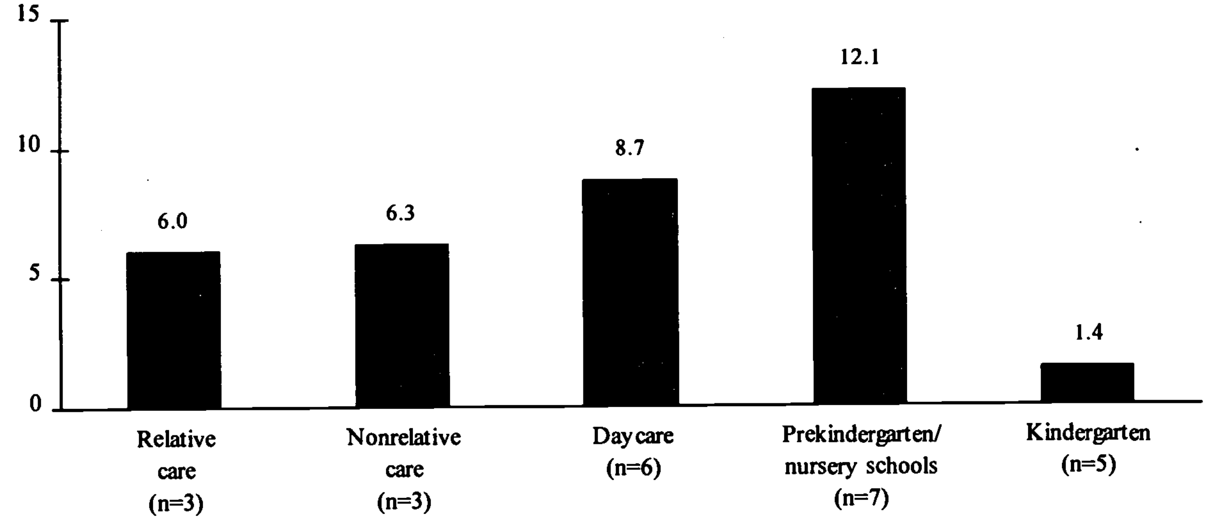
Figures 2 to 4, below, illustrate NHES:91 Early Childhood reinterview mean unreconciled GDRs by subject areas. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 2. -- Mean unreconciled GDR, NHES:91 Early Childhood Education (preprimary and primary reinterviews)



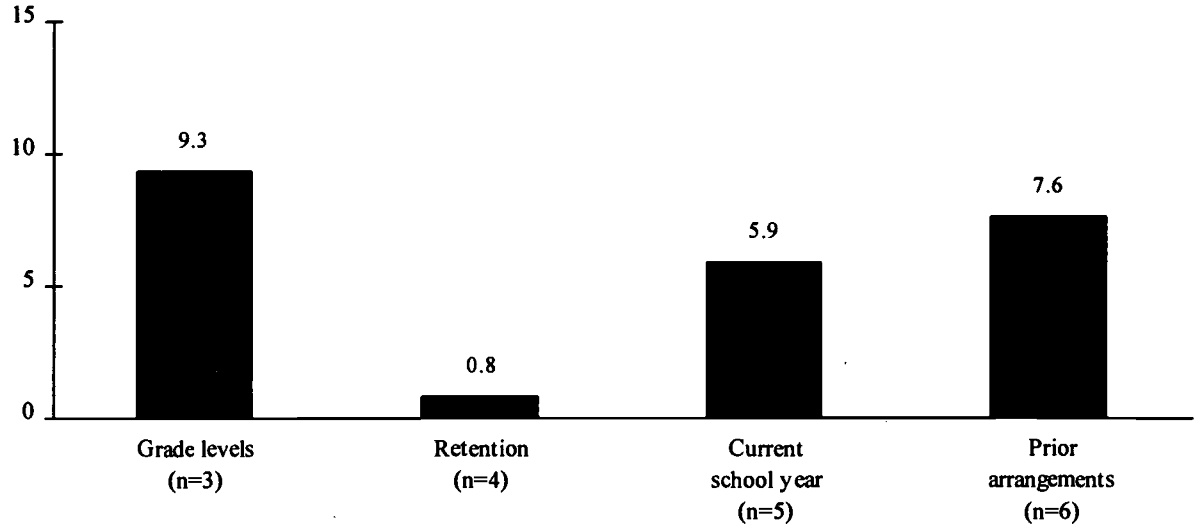
SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-17.

Figure 3. -- Mean unreconciled GDR, NHES:91 Early Childhood Education (preprimary reinterview)



SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-17.

Figure 4. -- Mean unreconciled GDR, NHES:91 Early Childhood Education (primary reinterview)



SOURCE: Derived from table 3, Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-18.

The NHES:91 looked at response bias, but with minimal results. The results for the NHES:91 study showed that 87 percent of the items had an NDR of less than 5 percent. Only four items, which were restricted to subgroups of the set of children with smaller sample sizes, had an NDR greater than 10 percent (P3G, P4G, and P3J were already seen).

- P3G Is (child's) program at this daycare center a Head Start Program? [NA or missing/yes/no] (GDR 12.9, IOI 41.3, NDR -12.9))
- P4G Is the program at the (first/next) (nursery school/prekindergarten) a Head Start Program? [NA or missing/yes/no] (GDR 13.5, IOI 43.9, NDR -13.5)
- P3J How often do you talk with (child's) primary care provider or teacher about how (he/she) is doing at this daycare center? [NA or missing/every time or most times child goes/frequently/ occasionally/rarely or never/no experience or newly enrolled] (GDR 18.2, IOI 25.4, NDR 18.2)
- P4TYPE Is the (first/next) program a...nursery school, prekindergarten or Head Start program? (GDR 23.3, IOI 40.3, NDR 16.7)

Unfortunately, the study results do not include tests to indicate how many NDRs are significantly different from zero.

NHES:93 School Readiness and School Safety and Discipline Reinterviews

The two topical components of NHES:93 were the School Readiness (SR) interview of parents of children, ages 3-10 and enrolled in second grade or below, and the School Safety and Discipline (SS&D) interview of parents of students enrolled in grades 3-12 (youths enrolled in grades 6 through 12 for whom the parent interview was already completed were subsampled and also interviewed). A random sample of completed interviews was selected for the reinterview study. The subset of the original SR and SS&D questionnaire items chosen were selected because they were substantively important, not highly time dependent, and not examined in the NHES:91 reinterview.

The reinterview sample sizes were substantially increased from the 604 of the NHES:91 reinterview study to obtain more reliable estimates of the response variance for key questions (table 10). The results were used to identify questions that posed difficulty for respondents.

Table 10. -- NHES:93 School Readiness (SR) and School Safety & Discipline (SS&D) reinterview studies

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
All	2,108	1,879	89%	7%	Response variance
SR	977	882	90%	9%	Response variance
SS&D					
parents of 3-5 graders	256	227	89%	10%	Response variance
parents of 6-12 graders	315	277	88%	3%	Response variance
6-12 graders	560	493	88%	8%	Response variance

SOURCE: Derived from table 2, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), p. 7.

Methodology and Design

The NHES:93 reinterviews began the first week of March, 1993. The reinterviews were designed to be conducted at least 2 weeks after the original interview; however, since all scheduled interviews in a household needed to be completed in order for any of them to be eligible for reinterview, the time lag was usually longer. Forty percent were conducted between 2 and 4 weeks after the original interview, and 51 percent more than 4 weeks after the original. The lag was reduced near the end of data collection in April to complete all reinterviews. No substitutions were allowed for the original respondents. As in NHES:91, the reinterview study was a CATI survey administered from a central location. Items for the reinterview were selected from specific subject areas. The NHES:93 reinterview used GDR, IOI, and NDR to estimate measurement error. In addition, some reconciliation was done between original interview responses and reinterview responses to provide a measure of response bias.

Summary of Results

The School Readiness (SR) subject areas included developmental questions, general topics, reading and meals, and television. Developmental items asked questions about things that children master at different ages; for example, identify certain colors by name, button clothes, have temper tantrums, etc. General topic items asked a variety of questions about early childhood programs, the child's adjustment to kindergarten or primary school, teacher feedback on child's school performance and behavior, whether the child had received any special help in school, and about the child's health. The reading and meals subject items include questions on how often parents read with children and family meal practices. The television subject area asked about the child's television and video viewing habits.

Most SR reinterview items had low to moderate IOIs, showing that the items were reported consistently. Items which referred to events in the past week, specifically reading and meal practices, had the highest response variance. Since the reinterview referred to events that occurred in a different week than the original interview, this result was not unexpected: responses could be different and still be correct. The time factor may also have contributed to the higher than average response variance for items concerning teachers' comments about the child during the school year.

The SR reinterview included some items about Head Start participation that had been revised since the NHES:91 interviews. The revised items asked more directly about Head Start participation and had relatively low response errors. Reinterview studies contributed to the improvement of these items from NHES:91.

- R32 Is (child) now attending or enrolled in Head Start? [yes/no] (GDR 4.9, s.e. 1.8; IOI 31.3, s.e. 11.0, estimated percent 9)
- R33 [Prior to starting (kindergarten/first grade), did/has] (child) ever (attend/attended) Head Start? [yes/no] (GDR 3.6, s.e. 0.6; IOI 19.7, s.e. 3.2, estimated percent 10)

The average rank of the IOIs for the SR subject areas are shown in table 11.

Table 11. -- NHES:93 average rank of subject area measurements for the School Readiness (SR) reinterview

Subject areas	IOI	Items with high IOI (above 50%)
Developmental	Moderate	R31
General topics	Moderate	R40; R51c-e; R52a, e, g-j, l; R79e
Reading and meals	High	R97, R98, R115
Television	High	R92c-f

SOURCE: Derived from table 13, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), pp. 36-37.

The items with high IOIs are listed below by subject area.

Developmental

- R31 Does (he/she) bend over to look very closely at pictures or drawings? [yes/no] (IOI 56.0, s.e. 10.9)

General topics

- R40 Have any of the (Head Start programs) (or) (nursery schools, prekindergartens, preschools, or day care centers) (child) has gone to had an educational program? [yes/no] (IOI 52.6, s.e. 9.5)
- R51c Did (he/she) pretend to be sick to stay home from school? [more than once a week/once a week or less/not at all] (IOI 53.3, s.e. 5.8)
- R51d Did (he/she) say good things about school? [more than once a week/once a week or less/not at all] (IOI 75.4, s.e. 16.9)
- R51e Did (child) say (he/she) liked (his/her) teacher? [more than once a week/once a week or less/not at all] (IOI 50.8, s.e. 4.3)
- R52a (Child) has been doing really well in school? [yes/no] (IOI 58.7, s.e. 11.2)
- R52e (Child) has often seemed sad or unhappy in class? [yes/no] (IOI 61.3, s.e. 10.6)
- R52g (Child) has been having trouble taking turns, sharing, or cooperating with other children? [yes/no] (IOI 52.3, s.e. 9.8)
- R52h (Child) gets along with other children or works well in a group? [yes/no] (IOI 74.2, s.e. 7.0)
- R52i (Child) is very enthusiastic and interested in a lot of different things? [yes/no] (IOI 53.2, s.e. 6.1)
- R52j (Child) lacks confidence in learning new things or taking part in new activities? [yes/no] (IOI 60.2, s.e. 8.8)
- R52l (Child) is often sleepy or tired in class? [yes/no] (IOI 61.8, s.e. 17.9)
- R79e Has (child) received any special help in school this year for children who are having trouble with English as a second language? [yes/no] (IOI 56.6, s.e. 20.1)

Reading and meals

- R97 How many times in the past week have you or has someone in your family read to (child)? [one or two times/three or more times] (IOI 51.3, s.e. 6.5)
- R98 Was that every day in the past week? [yes/no] (IOI 63.5, s.e. 12.1)
- R115 During the last week, on how many days did the whole family sit down to eat dinner together? [0-7 days] (IOI 64.6, s.e. 3.6)

Television

R92 On average; about how many hours of television or video tapes does (child) watch at home each weekday, that is, Monday through Friday?

- R92c How about between 3 pm and dinner time? [0-1- hrs.] (IOI 69.9, s.e. 2.8)
- R92d How about after dinner time? [0-10 hrs.] (IOI 66.7, s.e. 2.8)
- R92e How many hours does (child) watch television or video tapes at home on Saturday? [0-16 hrs] (IOI 64.5, s.e. 2.7)
- R92f How many hours does (child) watch television or video tapes at home on Sunday? [0-14 hrs] (IOI 62.8, s.e. 2.5)

The NHES:93 reinterview was also designed to test the reliability of composite variables. The two SR composites studied in Brick, Rizzo, and Wernimont (1997, p. 41) both appeared more reliable than the individual items. For example, the derived variable "Percentage of hours spent watching TV" had a lower IOI than all of the individual items. A developmental score from items asked of parents of preschool children also showed more reliable results. Composites were most useful in handling difficult concepts by subsetting related items.

The NHES:93 School Safety and Discipline (SS&D) component was designed to gather information about the school environment, safety at school, school discipline policy, and alcohol/other drug use and education. The parents/guardians of students enrolled full-time in grades 6 through 12 were asked about school characteristics, school environment, school safety, school discipline policy, tobacco, alcohol, and other drugs used and drug education, child characteristics, family characteristics, community characteristics, and parent and household information. Youth in grades 6 through 12 were asked a subset of these questions plus eight additional items concerning school safety and school discipline policy. They were also asked about privacy in responding. The parents of students in grades 3 through 5 were administered a subset of the parent items, those relevant to the school experiences of students in the lower grades.

The SS&D component generally indicated small measurement errors, although they were somewhat larger than those for SR, while reliability was nearly equal for all three types of respondents. Interestingly, the reliability of items for the 6th through 12th graders was similar to the reliability rates of their parents, suggesting that youth can respond effectively to telephone surveys like NHES. The general environment items relating to alcohol and drug education had estimates of response variance that were high for all three types of respondents; they are listed separately in the tables below. Brick, Rizzo, and Wernimont recommend further work on these items before they are used in future studies.

Items included in the NHES:93 reinterview can be grouped into categories of related items. Average IOIs were calculated for these categories and are shown in tables 12 to 14. For SS&D, these subject areas are classified in terms of parents of 3rd through 5th graders, parents of 6th through 12th graders, and the 6th through 12th graders themselves. Note that the items asked only of parents are designated by the prefix "P", items asked only of youths by "Y", and items asked of both parents and youth "PY". The questions following the tables are items with an IOI greater than 50.

SS&D, parents of 3rd through 5th graders

Table 12. -- NHES:93 average rank of subject area measurements for the School Safety & Discipline (SS&D) reinterview of parents of 3rd through 5th graders

Subject areas	IOI	Items with high IOI (above 50%)*
General environment	Moderate	P83c
Drug & alcohol education	Moderate	PY68b, PY68c, PY68d
Incidents	High	PY28, PY29, PY37, PY40, PY42, PY45

*Items asked only of parents are designated by the prefix "P"; items asked only of youths by "Y", and items asked of both parents and youth "PY".

SOURCE: Derived from table 6, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), p. 22.

General environment

- P83c Would you say that you are very satisfied, somewhat satisfied, somewhat dissatisfied, or very dissatisfied with the academic standards of the school? [very satisfied/somewhat satisfied/ somewhat dissatisfied/very dissatisfied] (IOI 79.5, s.e. 28.1)

Drug and alcohol education

PY68 There are many different ways that alcohol or other drug education can be presented to students. Did (child/you) receive alcohol or other drug education in school this year...

- PY68b A special course about alcohol or other drugs? [yes/no] (IOI 56.2, s.e. 10.5)
- PY68c At assemblies or demonstrations outside of classes? [yes/no] (IOI 50.4, s.e. 10.3)
- PY68d In other school activities or clubs? [yes/no] (IOI 61.9, s.e. 10.0)

Incidents

- PY28 Did it [stealing] happen to (child/you) this school year? [yes/no] (IOI 62.0, s.e. 18.8)
- PY29 (Have you heard/Do you know) of money or other things being taken directly from students or teachers by force or threat of force at school or on the way to or from school this school year? [yes/no] (IOI 81.1, s.e. 26.0)
- PY37 Did it [bullying] happen to (child/you) this school year? [yes/no] (IOI 56.6, s.e. 19.4)
- PY40 Did (child/you) see an incident like this [physical attack] happen to someone else [yes/no] ? (IOI 116.4, s.e. 159.2)

- PY42 Did it [physical attack] happen to (child/you) this school year? [yes/no] (IOI 76.6, s.e. 89.2)
- PY45 Have any of those incidents that happened at (child's) school this year interfered with (his/her) opportunity to learn? [a lot/somewhat/a little/not at all] (IOI 55.7, s.e. 15.1)

Parents of 6th through 12th graders

Table 13. -- NHES:93 average rank of subject area measurements for the SS&D reinterview of parents of 6th through 12th graders

Subject areas	IOI	Items with high IOI (above 50%)
General environment	Moderate	PY21a, PY59, P83d
Drug & alcohol education	High	PY68b-c
Drug & alcohol use	Moderate	PY63
Incidents	Moderate	PY30, PY35, PY40

SOURCE: Derived from table 7, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), pp. 23-24.

General environment

- PY21a When you think about (child's/your) experiences at (his/her/your) school since the beginning of this school year, would you strongly agree, agree, disagree, or strongly disagree with the statement "(child is/I am) challenged at school." [strongly agree/agree/ disagree/strongly disagree] (IOI 56.7, s.e. 10.7)
- P59 Now I have some questions about the school's discipline policy. Does it cover alcohol and other drug possession, use, and distribution? [yes/no] (IOI 80.6, s.e. 23.2)
- P83d Would you say were are very satisfied with the order and discipline at the school? [very satisfied/somewhat satisfied/somewhat dissatisfied/very dissatisfied] (IOI 53.1, s.e. 7.9)

Drug and alcohol education

PY68 There are many different ways that alcohol or other drug education can be presented to students. Did (child/you) receive alcohol or other drug education in school this year...

- PY68b A special course about alcohol or other drugs? [yes/no] (IOI 61.3, s.e. 9.1)
- PY68c At assemblies or demonstrations outside of classes? [yes/no] (IOI 75.0, s.e. 9.7)

Drug and alcohol use

- PY63 (Have you heard of/Have you seen) any students (having been) drunk or showing the effects of alcohol when they were at (child's/your) school this year? [yes/no] (IOI 57.6, s.e. 10.1)

Incidents

- PY30 Did (child/you) see an incident like this [force or threat of force] happen to someone else? [yes/no] (IOI 74.8, s.e. 32.0)
- PY35 (Was (child/you) see an incident like this [bullying] happen to someone else? [yes/no] (IOI 119.8, s.e. 48.1)
- PY40 Did (child/you) see an incident like this [physical attack] happen to someone else? [yes/no] (IOI 56.0, s.e. 15.6)

6th through 12th graders

Table 14. -- NHES:93 average rank of subject area measurements for the SS&D reinterview of 6th through 12th graders

Subject areas	IOI	Items with high IOI (above 50%)
General environment	Moderate	PY21e, Y44a
Drug & alcohol education	High	PY68a - PY68d
Drug & alcohol use	Moderate	--
Incidents	Moderate	PY29, PY30, PY35, PY37, PY39, PY40, Y52f, Y52h

SOURCE: Derived from table 8, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), pp. 26-27.

General environment

- PY21e The principal and assistant principal maintain good discipline at (child's/my) school. [strongly agree/agree/disagree/ strongly disagree] (IOI 63.8, s.e. 10.5)
- Y44a [Did you do any of the following things because you were worried that someone might hurt or bother you?] Take a special route to get to school? [yes/no] (IOI 59.0, s.e. 15.2)

Drug and alcohol education

PY68 There are many different ways that alcohol or other drug education can be presented to students. Did (child/you) receive alcohol or other drug education in school this year...

- PY68b A special course about alcohol or other drugs? [yes/no] (IOI 62.4, s.e. 6.1)
- PY68c At assemblies or demonstrations outside of classes? [yes/no] (IOI 51.9, s.e. 4.0)
- PY68d In other school activities or clubs? [yes/no] (IOI 53.6, s.e. 4.7)

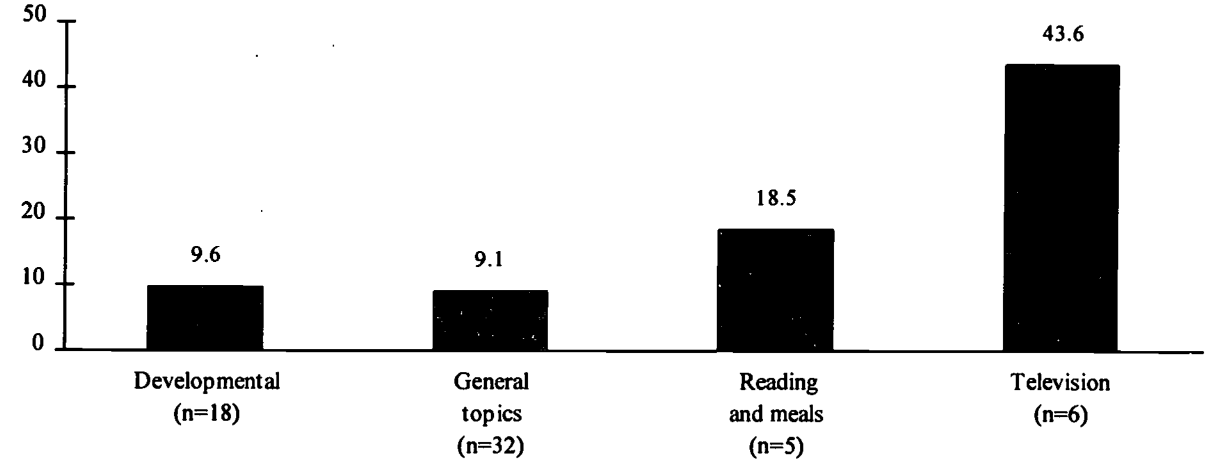
Incidents

- PY29 (Have you heard/Do you know) of money or other things being taken directly from students or teachers by force or threat of force at school or on the way to or from school this school year? [yes/no] (IOI 67.0, s.e. 9.5)

- PY30 Did (child/you) see an incident like this happen to someone else? [yes/no] (IOI 99.9, s.e. 24.2)
- PY35 (Was (child/you) see an incident like this [bullying] happen to someone else? [yes/no] (IOI 68.7, s.e. 10.2)
- PY37 Did it happen to (child/you) this school year? [yes/no] (IOI 58.4, s.e. 7.2)
- PY40 Did (child/you) see an incident like this [physical attack] happen to someone else? [yes/no] (IOI 50.3, s.e. 11.1)
- Y52f Did you bring mace? [yes/no] (IOI 62.0, s.e. 35.2)
- Y52h Did you bring a stick, club, or bat? [yes/no] (IOI 58.1, s.e. 44.9)

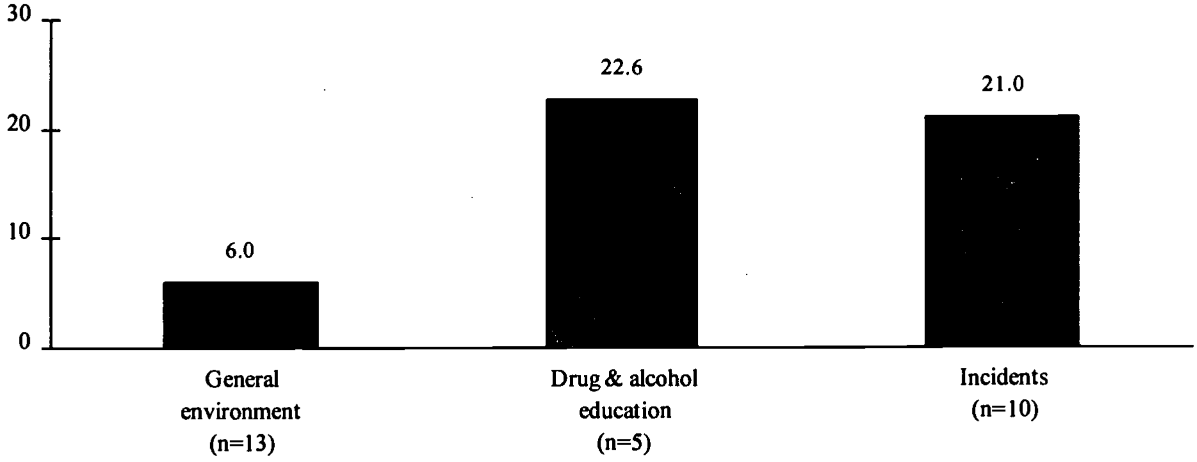
Figures 5 to 8 illustrate mean unreconciled GDRs for the subject areas in NHES:93 School Readiness and School Safety & Discipline reinterviews.

Figure 5. -- Mean unreconciled GDR, NHES:93 School Readiness



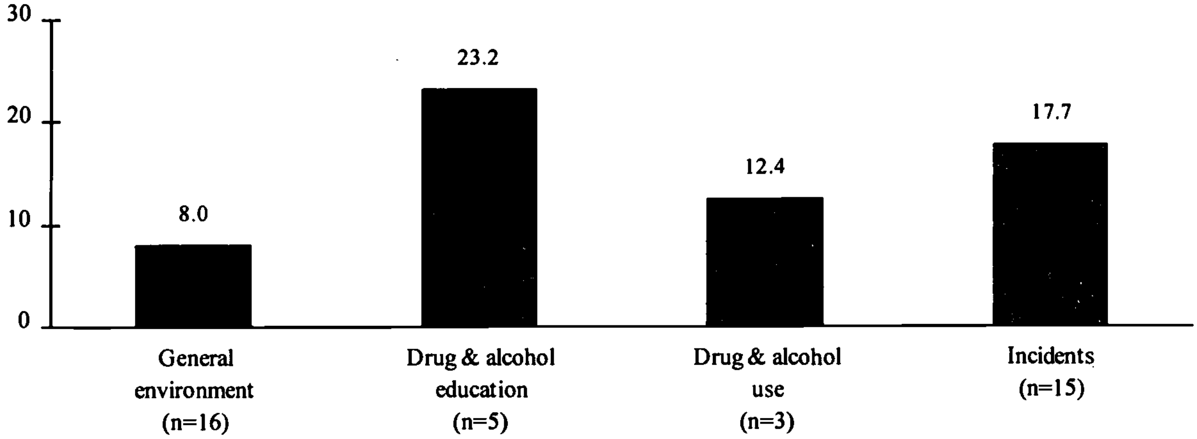
SOURCE: Derived from table 13, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), pp. 36-37.

Figure 6. -- Mean unreconciled GDR, NHES:93 School Safety and Discipline (parents of 3rd-5th graders)

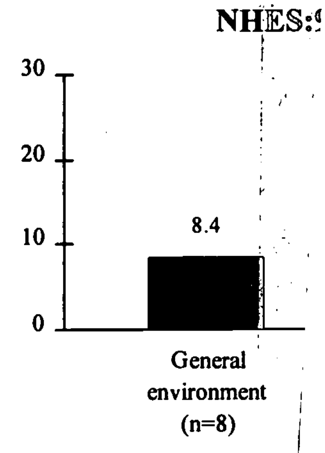


SOURCE: Derived from table 6, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), p. 22.

Figure 7. -- Mean unreconciled GDR, NHES:93 School Safety and Discipline (parents of 6th-12th graders)



SOURCE: Derived from table 7, Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), pp. 23-24.



SOURCE: Derived from table 8, B ool
Readiness Components (NCES 97-339), pp. 26-27.

Brick, Rizzo, and Wernimont's conclusions on their use of net difference rates are as follows.

> *[For SR], the net difference rates computed from the original interview and unreconciled reinterview responses...can be used to assess the assumption that the two interviews were conducted under the same general conditions. Of the 55 items in the reinterview, the t-statistics for 15 items are greater than 2.0. This is nearly one-quarter of the items and is greater than the 5 percent that would be expected by chance alone. There does not appear to be a pattern to the estimated biases; six are positive and nine are negative. These results are not very supportive of the assumption that conditions for the original and reinterview were the same; rather, they raise concerns about how valid the gross difference rates are as measure of response variance* (1997, p. 38).
>
> *[For SSD], the median net difference rates are generally close to zero, indicating that the response bias, as measured by the net difference rate, is not very large. The number of items with t-statistics greater than 2.0 is larger than expected by chance; 17 of the 105 items have t-statistics greater than 2.0, while only about 5 would be expected to be this large by chance. There is no pattern in these statistics to suggest that a certain type of respondent or grouping of items is more subject to response bias. These results show that either the items were not subject to large response bias or that the reconciliation process is not capturing the "true" values for the respondents* (1997, p. 28).

NHES:95 Adult Education Reinterview

As mentioned earlier, the components included in NHES:95 were Early Childhood Program Participation and Adult Education, essentially the same components as in NHES:91. The NHES:95 reinterview study only examined and estimated measurement errors as components of nonsampling error in the Adult Education (AE) survey.

Methodology and Design

A subset of items from the original interview was selected and the original and reinterview responses were then compared to estimate the consistency of reporting. The items selected provided key statistics or were used for critical estimates, substantively important, not highly time dependent, and those not examined in previous NHES reinterviews.

Both the original interviews and the reinterviews were administered using a CATI system. In general, the sampling for reinterview was not completed until 2 weeks after all original interviews in the household were complete; however, this lag was relaxed toward the end of the data collection period to sample all eligible AE interviews. Interviews were sampled at different rates for participants and nonparticipants (i.e., people who did not participate in adult education activities), with a total of 1,289 cases selected for reinterview. Characteristics of the NHES:95 reinterview study are shown in table 15.

Table 15. -- NHES:95 Adult Education (AE) reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
All	1,289	1,109	86%	6%	Response variance
AE participants	917	882	96%	10%	Response variance
AE nonparticipants	372	227	61%	4%	Response variance

SOURCE: Derived from tables 2 and 3, Brick, Wernimont and Montes, (1996), *The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component* (Working Paper 96-14), pp. 8-9.

The NHES:95 reinterview study calculated GDR and NDR to estimate measurement error.

Summary of Results

Items were selected in three subject areas: 1) adult education participation (including items on how much time and how much of the respondent's own money went toward the total sum of books, courses, transportation, child care, etc. for English as a second language classes, basic skills and GED (General Educational Development) preparation, and courses taken as part of a program leading to a credential or degree) 2) education background (including items on years of schooling, degrees obtained, English speaking ability, labor force status, and job benefits), and 3) barriers to participation (i.e., obstacles that may have prevented respondents from taking part in adult education activities).

The GDRs were satisfactory for the adult education participation and the education background items, indicating that responses to those questions were consistent. For the adult education participation category, all items, including an overall participation composite variable and a question about using computer-based or interactive video instruction without an instructor being present had GDRs below 15 percent. The overall participation composite variable had a GDR higher than most of the individual items (GDR 13.4, s.e. 3.7; estimated percent 44.9, s.e. 142.3). The items that measured participation in work-related and personal development courses contributed to this inconsistency. However, the composite variable was not included when the mean unreconciled GDR was calculated (see figure 9). Calculations for reconciled items in Adult Education revealed even smaller response variability, down to 4.0 from 5.5. For the 23 education background items, only 3 items had a GDR over 10 percent, and none were over 17.

Brick, Wernimont, and Montes (1996) identified only three background items with GDRs large enough to possibly merit further consideration. They suggested NCES might consider improving the question wording when those items were next used. They pointed out, for example, that the unemployment question (I25) is a compound question: respondents might not be sure whether the descriptor (for as long as a month) applied to either or both parts of the question (p. 21).

- A10V Did you earn a vocational or technical diploma after leaving high school? [yes/no/refused/don'tknow] (GDR 17.0, s.e. 10.9; estimated percent 13.3, s.e. 14.1)
- I14 Does your occupation have legal or professional requirements for continuing training or education? [yes/no/refused/don'tknow] (GDR 14.7, s.e. 6.0; estimated percent 21.5, s.e. 25.2)
- I25 At any time during the past 12 months, have you been unemployed and looking for work for as long as a month? [yes/no/refused/don't know] (GDR 15.6, s.e. 19.5; estimated percent 36.8, s.e. 1.5)

The GDRs for barrier to participation items were much higher than for the other subject areas, indicating that responses were not consistent. Only four (out of 15) barrier items had GDRs less than 10 percent, and the highest GDR approached 50 percent. This inconsistency may have been related to factors like recoding the questions, additional eligibility criteria, and small sample sizes. Most of the estimates for barrier items were not significantly greater than zero, as the standard errors were high. Nonetheless, barrier items had some response problems and did not appear to be reliable. When considering both barrier to participation items and other characteristics of participating in AE activities, one-third of the 30 items had GDRs greater than 20 percent, another third had GDRs between 10 and 20 percent, and the remaining third had GDRs of less than 10 percent. Brick, Wernimont, and Montes concluded that

> *Given the response problems associated with questions on barriers, it may be useful to reconsider the method of addressing obstacles. For example, given the difficulty associated with discriminating between major and minor obstacles,*

> *this difference might be dropped. Other types of questions about behaviors of adults might also be investigated rather than the current questions. For example, rather than asking directly about barriers to participation, questions could be asked about whether adults took any steps to try to either take courses or obtain information about courses and, if so, whether they were discouraged for specific reasons* (1996, p. 26).

The use of NDRs as a measure of response bias will be discussed in the next section, NHES:95 Adult Education Bias Study, but in NHES:95 Brick, Wernimont, and Montes (1996) used NDRs to test the assumption that independent replication makes GDR a valid measure of response variance. When the NDRs were calculated for comparison of the original and reinterview values, none of the items in the adult participation area were significantly different from zero, supporting the assumption of independence and the validity of GDR as a measure of response variability.

Brick, Wernimont, and Montes (1996, p. 26) recommended increasing the sample size of the reinterview to improve the precision of estimates (lower standard errors) and suggested that specific groups in the reinterview not be oversampled, as the nonparticipants were in this study.

Table 16 summarizes the items identified for further consideration in the Adult Education (AE) reinterview.

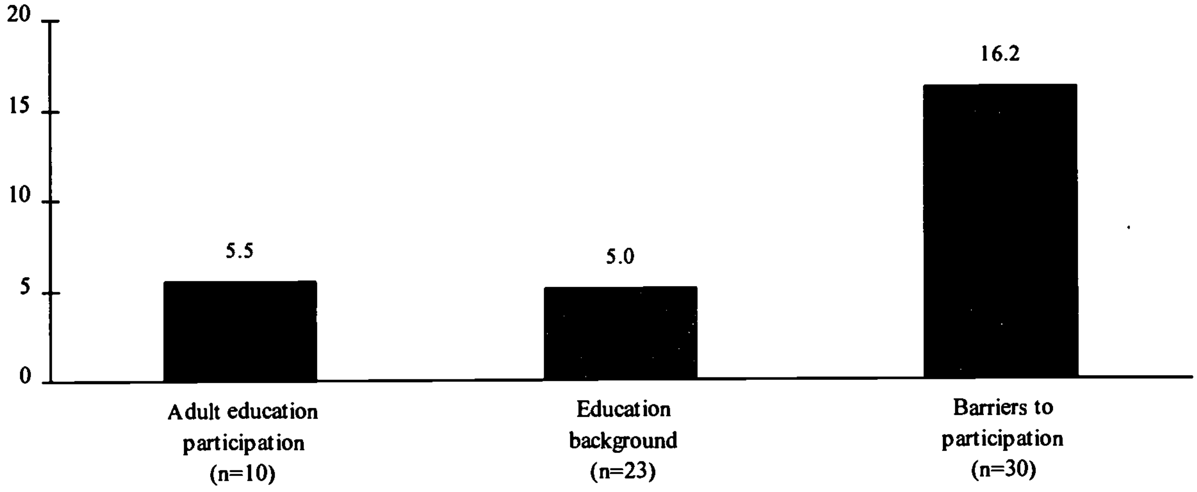
Table 16. -- NHES:95 items identified for further consideration in the Adult Education (AE) reinterview

Subject areas	Items identified for further consideration
Adult Education Participation	--
Education Background	A10V, I14, I25
Barriers to Participation questions	--

SOURCE: Derived from tables 5, 7, and 8, Brick, Wernimont and Montes, (1996), *The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component* (Working Paper 96-14), pp. 17, 20, 22.

Figure 9 illustrates mean unreconciled gross difference rates (GDRs) for each subject area.

Figure 9. -- Mean unreconciled GDR, NHES:95 Adult Education

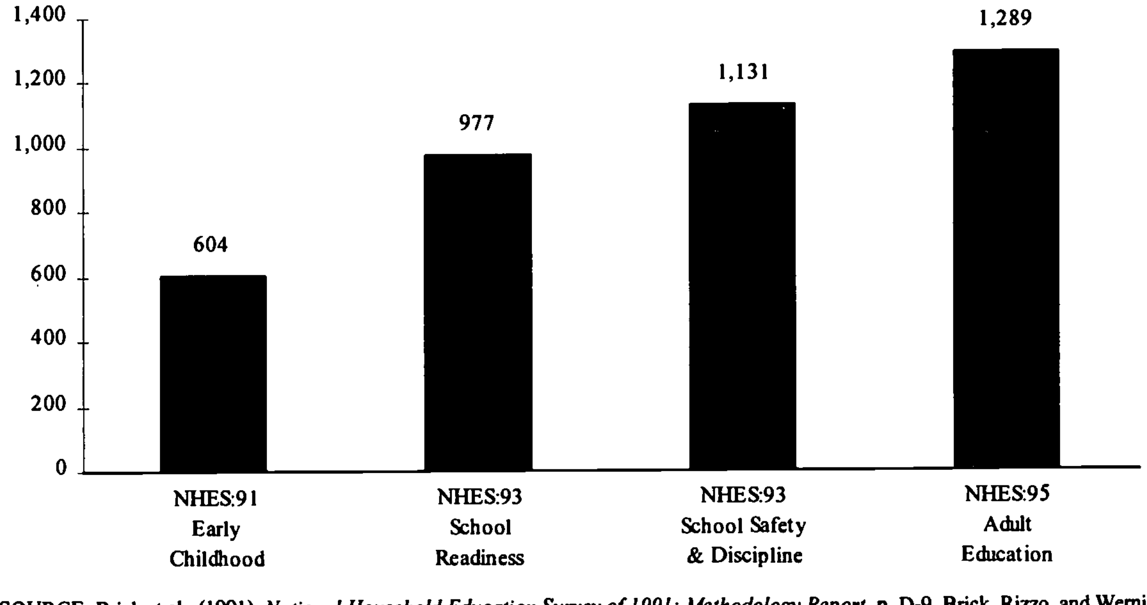


SOURCE: Derived from tables 5, 7, and 8, Brick, Wernimont and Montes, (1996), *The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component* (Working Paper 96-14), pp. 17, 20, 22.

Review of NHES Reinterviews

The sample size for the studies increased over time (see figure 10) although response rates and sampling percentages for the NHES:93 study were higher than for NHES:91 and 1995.

Figure 10. -- Sample size of NHES reinterviews



SOURCE: Brick et al., (1991), *National Household Education Survey of 1991: Methodology Report,* p. D-9. Brick, Rizzo, and Wernimont, (1997), *Reinterview Results for the School Safety & Discipline and School Readiness Components* (NCES 97-339), table 2, p. 7. Brick, Wernimont and Montes, (1996), *The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component* (Working Paper 96-14), table 3, p. 9.

The major emphasis of the NHES 1991, 1993, and 1995 reinterview studies was to measure response variability. Overall, the results were positive. For NHES:91, nearly 85 percent of the items in the reinterview study had low to moderate response variability. Brick et al. (1991) recommended some items related to a specific subject be modified, and the modified questions had lower response variability in the NHES:93 study. The overall results of the NHES:93 study showed that while the School Safety & Discipline (SS&D) items had higher inconsistency than the School Readiness (SR) program items, which were all reliable, no severe problems were detected. Within the NHES:95 reinterview study, only a few items in one subject area had high response variability. The reinterview responses were consistent for most items; only minor modifications were suggested.

There were also limited attempts to measure response bias. A more complete bias study, described below, was undertaken in 1995.

NHES:95 Adult Education Bias Study

Reconciliation was used in the NHES:91 and NHES:93 reinterview studies to examine response bias. However, the estimates of response bias based on the reconciled responses were not very different from the response based on the unreconciled reinterview data: "regardless of how these differences between the reinterview and reconciled reinterview response are interpreted, the reconciliation produced little additional information" (Brick, Kim, Nolin and Collins, 1996, p. 4). NHES:95 used an alternate method to estimate response bias: an *intensive* reinterview separate from the already described reinterview study to study response variability.

Methodology and Design

The basic objective of the Adult Education (AE) Bias Study was "to obtain more detailed and accurate information by understanding the respondent's perspective and the reasons for his or her answers" (Brick et al., 1996, p. 8). Specifically, four Bias Study goals were identified

- To examine the potential bias due to either underreporting or overreporting participation in adult education activities
- To examine bias in the estimates of participation rates in work-related and personal development courses (these courses were identified in the NHES:95 field test as the types of AE that were most susceptible to underreporting)
- To assess the validity of the responses to the barriers to participation items contained in the survey
- To explore reasons for switching participation status between the NHES:95 screener and the AE extended interview (Brick et al., 1996, pp. 5-7)

The method used was an intensive reinterview, "more a directed conversation between the respondent and the interviewer rather than a formally scripted interview" (Brick et al., 1996, p.

8). It involved many techniques discussed in the section on cognitive research in chapter one, such as asking open-ended questions and using probes to encourage respondents to elaborate.

The protocol developed contained information on the respondent's participation status, education level, and responses to the original survey. There were also some open-ended questions and suggested techniques for eliciting detailed answers. The protocol was revised after interviewer training. The initial questions and suggested probes were reduced in number and reworded into a more conversational tone. The examples of educational activities were simplified and reorganized by type (hobby/special interest, personal development, work-related). The protocol was also revised to include a timeline with major dates that the interviewer could use for reference (Brick et al., 1996, pp. 9-10).

Interviewers were carefully selected. The regular pool of NHES interviewers have been trained to read questions verbatim and to avoid behavior that might influence respondents. The five interviewers selected for this study had previous experience in conducting semi-structured interviews and had experience in a wide variety of educational studies. They received training in additional intensive techniques, such as active listening, giving neutral feedback to encourage in-depth responses, probing for details, and the creative use of silences. They were also given background information on the AE component and an overview of the purpose of the Bias Study (Brick et al., 1996, pp. 10-11).

The Bias Study budget only allowed for a very limited sample size. The sample was randomly selected from both participants and nonparticipants who completed the AE interview and were not included in any other special studies (i.e., NHES:95 Splice Sample Interview, NHES:95 Expanded Screener Interview, and the reinterview study already discussed). In order to be eligible for the study, the following conditions had to be met

- The original interview was never a refusal or a language problem
- The original interview was conducted in English (as opposed to Spanish)
- Only one interview in a household was eligible
- All other extended interview in the household had been completed

The eligible cases were first classified into the following non-exclusive, six groups for sampling

- Adults who completed the original interviews as participants in adult education
- Adults who were sampled as nonparticipants but completed the interviews as participants in adult education
- Adults who were sampled as participants but completed the interviews as nonparticipants in adult education
- Adults who were sampled as low-education nonparticipants and answered the questions about obstacles to taking courses
- Adults who were sampled as low-education nonparticipants and did not answer the questions about obstacles to taking courses
- Adults who were sampled as high-education nonparticipants or participants

The sample was drawn by randomly selecting adults from the first group, then sampling all eligible adults in the second group, provided they were not already sampled, etc. A sample of 230 adults was selected from all 6 groups; 206 respondents completed intensive reinterviews (see table 17) (Brick et al., 1996, pp. 11-13).

Table 17. -- NHES:95 Adult Education (AE) reinterview bias study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
AE	230	206	90%	1%	Response bias

SOURCE: Brick et al., (1996), *Estimation of Response Bias in the NHES:95 Adult Education Survey* (Working Paper 96-13), p. 13.

Data collection for the Bias study began 10 weeks after the beginning of data collection for the full survey and extended over 6 weeks. Because all cases that met the four criteria for eligibility were included in the pool for sampling, some respondents were recontacted shortly after the original interview and some after a longer time: the time interval between the original interview and the intensive reinterview was between 6 and 18 weeks.

Individual interview protocols—protocols containing a sampled individual's name, telephone number, level of education, participation status, whether interest or barriers questions had been answered in the original survey, and whether participation status had switched during the interview—were prepared by project staff and reviewed by interviewing staff prior to placing telephone calls (Brick et al., 1996, p. 14).

Summary of Results

Participation rates. The bias study showed substantial underreporting by adults who participated in AE in 1995. Brick et al. felt that the estimates of participation were less than they would have been if adults had completely understood the intent of the survey. One hypothesis for the underreporting was that "some respondents may have created a response paradigm that restricted their answers to more formal courses when the questions about the work-related and personal development courses were asked" (1996, p. 35). If this is so, intervention is needed to modify this behavior. Brick et al. suggested that cognitive research on different types of intervention would be useful.

Barriers to participation. The bias study confirmed the reinterview findings that responses to barriers questions were not very consistent. It also uncovered respondent confusion between lack of interest and barriers to participating in AE activities. The authors felt strongly that the approach to asking about barriers needed to be completely reconsidered, and recommended additional cognitive interviews.

Switching participation status. The bias study did not substantially add to the understanding of this topic. As recognized before the study, some proxy respondents may not know the correct answers, and many do not appear to understand the full range of activities that should be included when answering screening questions. This study did not lead to suggestions for ways to improve the situation.

Brick et al. concluded that the intensive reinterview methodology used for the NHES:95 bias study appeared "to have good potential as a method for detecting biases especially if more traditional methods like record check studies are not feasible" (1996, p. 38). However, they suggested that the sample design should be a valid probability sample large enough so that reliable estimates of biases and their standard errors could be calculated. Intensive interviews were not as successful as the standard NHES reinterview for estimating consistency of reporting. The bias study was also not particularly successful in eliciting reasons from respondents. Cognitive methods might be better suited to that purpose. Finally, because intensive reinterviews are more costly that regular reinterviews, they suggested that this method be used "primarily when there is an indication of reporting errors that might result in biases and the estimates subject to the biases are important to the survey objectives" (1996, p. 39).

Recent College Graduates (RCG:91) Reinterview Study

The 1991 Survey of Recent College Graduates (RCG:91) provided data on the occupational and educational outcomes of bachelor's degree and master's degree recipients one year after graduation. Telephone interviews were conducted between July 1991 and December 1991 using computer-assisted telephone interviewing (CATI). Only graduates who met the following eligibility requirements were included in the survey

- Received a bachelor's or master's degree from the sampled institution
- Received the degree between July 1, 1989 and June 30, 1990
- Lived in the United States at the time of the survey

In RCG:91, 400 higher education institutions and 18,000 graduates were sampled. The weighted institution response rate was 95 percent, while the weighted graduate response rate was 83 percent (Brick, Cahalan et al., 1994, p. 1-1).

Both a reinterview study and a record check study were conducted for RCG:91. The record check study will be discussed in chapter 6.

The reinterview study was conducted to estimate the impact of measurement errors on estimates from the survey. Its goals were

- To identify unreliable survey items
- To quantify the magnitude of the measurement error
- To provide feedback on the design of questionnaire items for future surveys

Selected items were reconciled at the end of each reinterview. The net difference rate (NDR), based on the reconciled data, was used as a measure of the direction and magnitude of the potential response bias. The reinterview design attempted to maximize the ability to estimate the random component of measurement error using the gross difference rate (GDR) and the index of inconsistency (IOI), based on the unreconciled data.

Methodology and Design

The reinterview study was a one-stage sample of the original interview. Its goal was to complete 500 reinterviews from a target sample of 583 original interview respondents. To be included in the reinterview study, respondents had to be

- Bachelor's degree recipients
- Graduates who had never refused to participate
- Graduates interviewed in the original survey between August 15 and September 30

The eligibility requirements excluded respondents interviewed in the first 3 weeks of data collection (when interviewers were less familiar with the survey) and, by establishing a cutoff date, ensured that at least 2 weeks elapsed between the original and the reinterview.

The reinterviews were conducted in October and November, about 4 to 6 weeks after the original interview. The supervisors selected their better interviewers for the reinterview study (Brick, Cahalan et al., 1994, p. 3-2); however, the same interviewer was not allowed to conduct the original and reinterview with a respondent to protect the assumption that the two interviews were independent. The reinterview study used CATI as its mode of data collection.

Table 18. -- RCG:91 reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
RCG:91	583	512	88%*	4%	Response bias & variance

* A reinterview sample was selected from which only a targeted number (500) needed to be completed.

SOURCE: Brick, Cahalan et al., (1994), *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates* (NCES 95-640), p. 3-2.

The question wording for the reinterview was the same as the original interview, but only a subset of the items were included. Question selection was based on three general considerations: reliability, variety, and context. NCES wished to examine the reliability of key questions for reporting and comparing over time. By using a variety of questions, NCES could gather information on which types of questions were most subject to inconsistency. Context was a consideration because some questions were connected with other questions and were difficult to replicate in isolation. Specifically, questions were selected for inclusion in the reinterview because they represented sensitive items, date recall problems, and complex questions (Brick, Cahalan et al., 1994, pp. 3-2, 3-3, and 3-11).

The items chosen for reconciliation had typically been used for analysis in other RCG reports. Reconciliation occurred after the completion of the whole reinterview, not on a question-by-question basis, and interviewers were not aware of the original responses during the reinterview. Although the data were reconciled, the data from the original interview were retained on the final database. Respondents were also asked what they thought was the reason for the discrepancies. The most common reason cited was recall problems (36%), but 20 percent of the respondents reported the question was unclear or the response category did not fit their situation, and 11 percent claimed the interviewer recorded the wrong response. The reasons for discrepancies reported during reconciliation provided insight for revising the survey items (Brick, Cahalan et al., 1994, pp. 3-3 and C-7).

Two measurement error models were estimated from the reinterview data. The first model (the simple response variance model) assumed the errors were all from random sources. This model was then expanded to allow for systematic errors or biases. Both models assumed that the interviewers were not a source of systematic error in the data collection process, but the first assumed that the measurement errors were the same across sampled graduates and from one trial to the next. Thus, if the reinterview were uncorrelated with the original interview, then the number of original and reinterview errors should be roughly equal.

There are some challenges to this assumption, however. Specifically, the correlation in responses between the original and the reinterview may be a result of respondent recall (see discussion on time lag in chapter 1), or could be attributable to the interviewer. The error term may not have zero mean over replications of the survey and the error variances may be heterogeneous. These latter two conditions were of more concern for categorical level data.

The second model, the response bias model, assumed the response bias for the second trial—the reinterview—was zero. In order to meet those conditions, the reinterview had to be free of measurement error. Selected questions were reconciled in an attempt to obtain more accurate responses. Because the assumption that the reinterview had no measurement error was almost impossible to support, reconciled responses were used only to identify the expected direction of bias, and the relative amount of bias, but not to provide precise estimates of the size of bias. The reconciliation process was also limited because it did not detect consistent errors made in both the original and the reinterview (Brick, Cahalan et al., 1994, pp. 3-3 to 3-6).

The RCG:91 IOI reinterview results were interpreted based on the Bureau of the Census guidelines presented in *Evaluating Censuses of Population and Housing* (U.S. Bureau of the Census, 1985). The one exception was that RCG used a cut-point of greater than 45 for labeling IOI as high whereas the Bureau of the Census uses 50 as the cut-point. Thus, the RCG cut-points were less than 20 for low, 20 to 45 for moderate, and greater than 45 for high.

Summary of Results

The unreconciled GDR and IOI estimates, their standard errors, and the reconciled GDR and NDR estimates with their standard errors, are given for each item selected for reinterview in Brick, Cahalan et al. (1994), as well as the mean unreconciled GDRs for the categories of items. Their general conclusions are presented below.

The areas considered key topics in the RCG:91 reinterview were employment experience, additional education, and teacher status. Employment items asked about assistantships, any other work for pay, whether the respondent was looking for or available for work, and the main reason the respondent was not working. Additional education asked about any formal training the respondent had during or after completing the 1989-90 degree. Teacher status examined teacher eligibility and certification. Of the 16 reinterview items in these categories, only two had an IOI greater than fifty percent. One item related to employment experience, while the other was a question about teacher certification and employment.

- 24 Were you looking for work during the week of April 22, 1991? [yes/no] (IOI 58.8, s.e. 11.0)
- 62 Prior to completing the requirements for your 1989-90 degree, were you at any time employed as a school teacher at any grade level, from prekindergarten through grade 12? Please exclude student or practice teaching and work as a teacher's aide. [yes/no] (IOI 63.0, s.e. 7.3)

Question 24 asking graduates if they were looking for work was only asked of the subset of the sample of graduates who were unemployed. The reduced sample size may have contributed to a larger GDR and IOI. There was also a potential for recall problems since the question referred to a specific period of time. No explanation was offered for the possibly high random measurement error for question 62 asking if the graduates were employed as teachers before completing their degrees. However, questions asking for retrospective information have been linked to higher variability (Brick, Cahalan et al., 1994, p. 3-17).

The response variance levels for the RCG:91 subject areas, and the questionnaire item number for questions with high variability, are shown in table 19.

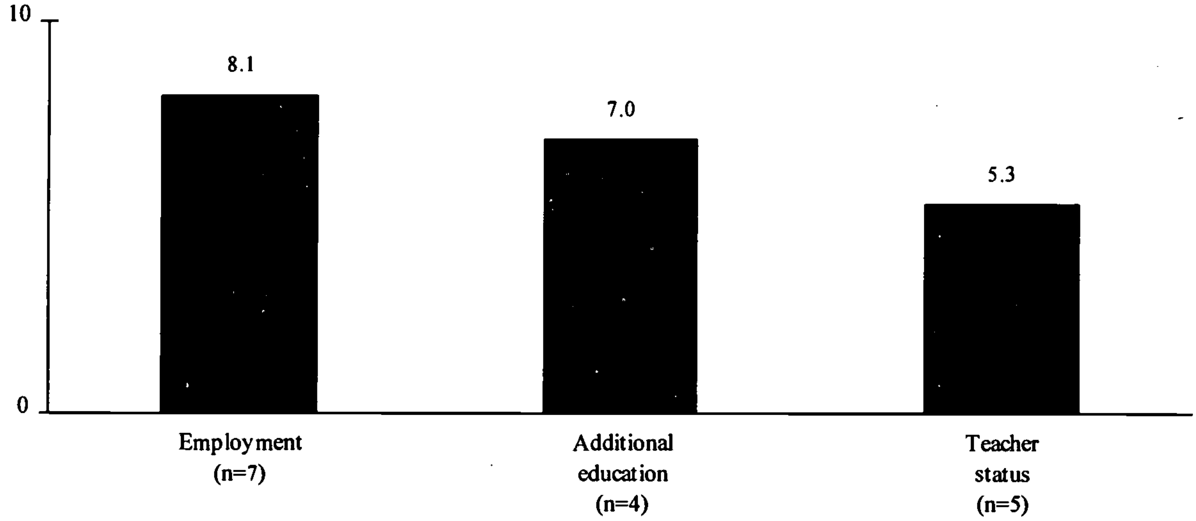
Table 19. -- RCG:91 average rank of reinterview subject area measurements

Subject areas	IOI	Items with high IOI (above 50%)
Employment	Low	24
Additional education	Low	--
Teacher status	Low	62

SOURCE: Derived from table 3-1, Brick, Cahalan et al., (1994), *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates* (NCES 95-640), p. 3-12.

Figure 11 shows mean unreconciled GDRs for some RCG question categories. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 11. -- Mean unreconciled GDR, RCG:91



SOURCE: Derived from table 3-1, Brick, Cahalan et al., (1994), *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates* (NCES 95-640), p. 3-12.

The GDRs and IOIs for items selected because of anticipated high error (such as sensitive, date, and ambiguous/complex items) were somewhat larger than for the key items, but could still be considered moderate. Thirty-two percent had indices below 20 (low), 50 percent had indices between 20 and 45 (moderate), and 17 percent had indices of more than 45 (high). Taking the reconciled values as the correct responses, Brick, Cahalan et al. estimated the response bias for the key items. They found that the NDRs were small for almost all of the key items; only two items had an NDR of greater than 5 percent, indicating that either the response bias was small for the items that were reconciled or the reconciled reinterview did not result in significantly reducing any of the bias that may have been associated with the items. They concluded that the reinterview "did not provide evidence of significant response bias for nearly all of the key items in the RCG:91" (1994, p. 3-14).

The RCG:91 reinterview findings supported the conclusion of other reinterview studies (HS&B, SASS) that items asking for factual and status information have lower response variability than questions asking for opinions, or more complex responses. Other studies have also found that items asking for recent or current information have lower variability than those that are retrospective or ask for future expectations. Generally, the more open-ended the response choices, the more specific a date requested, and the higher the number of response categories, the greater the variability. For example, questions in which the graduate was asked to provide the exact month and year had the highest gross difference rates (Brick, Cahalan et al., 1994, p. 3-16).

The overall conclusions of Brick, Cahalan et al. were that even though measurement errors were an important source of error in RCG:91, the estimates from the survey were not greatly distorted by these errors. The relatively small GDRs indicated responses were consistent; however, the IOIs being generally moderate implied that improvements in questionnaire wording and construction might help to reduce measurement errors in future surveys.

Schools and Staffing Survey (SASS) Reinterview Studies

The Schools and Staffing Survey (SASS) is a periodic, integrated system of surveys designed to collect data on characteristics of public and private school teachers, administrators, and their workplaces. The first two rounds of SASS included the School Survey, the School Administrator Survey, the Teacher Demand and Shortage Survey (TDS), the Teacher Survey, and, one year later each time, the Teacher Followup Survey. In 1993, SASS added three new components: the Library, Librarian, and the Student Records surveys. Survey data are collected by mail, with telephone followups to nonrespondents.

For each round of SASS, the sample selection proceeds in stages. First, a sample of schools is selected. The sample is designed to provide separate data for public and private schools, with detail by state for the public sector and by association group for the private sector. The same sample is used for the School Administrator Survey and the Library surveys. For the sample of private schools, the questions for the TDS are included in the questionnaire for the School Survey. The second selection stage requires each local education agency (LEA) administering one or more of the sample schools in the public sector to become part of the TDS sample. In the third selection stage, a list of teachers is obtained from each sample school and a sample selected for inclusion in the Teacher Survey.

Finally, a subsample of the teachers who participated in the Teacher Survey and continued teaching in the same or another school is selected and contacted during the following school year for the Teacher Followup Survey. These teachers receive the questionnaire for current teachers. Additionally, all teachers who responded in the Teacher Survey and are no longer teaching in an elementary or secondary school are contacted in the Teacher Followup Survey. These participants receive the questionnaire for former teachers.

SASS was conducted in school years 1987-88, 1990-91, and 1993-94. The Teacher Followup Survey was conducted one academic year after each SASS, in 1988-89, 1991-92, and 1994-95.

Reinterview Studies

SASS includes reinterview studies as part of its survey design, in addition to other types of measurement error studies. For example, SASS conducts the cognitive research discussed in chapter 6. [Note: The 1993 Teacher List Validation Study, while essentially a reinterview study, is included in the chapter 7, Cognitive Studies, because of its close connection with the SASS Teacher Listing Form Study.]

SASS reinterview studies in 1987-88 and 1990-91 consisted of a subset of questions administered to a subset of each sample—school administrators (School Administrator Survey), schools (School Survey), and teachers (Teacher Survey). Reinterviews were not conducted with LEAs (Teacher Demand and Shortage Survey) because multiple persons completed the forms and tracking all respondents would have been difficult. We present a summary of the combined public and private reinterview results for each component. A discussion of the combined public and private results of the Library Survey reinterview study from the 1993-94 SASS follows. Finally, we will turn to the Teacher Followup reinterview results from 1988-89 and 1991-92. Our discussion, once again, combines the public and private results.

Table 20 lists the SASS reinterview studies described in this section.

Table 20. -- Schools and Staffing Survey (SASS) components and selected reinterview studies

SASS component	Reinterview
Administrator Survey	1987-88, 1990-91
School Survey	1987-88, 1990-91
Teacher Survey	1987-88, 1990-91
Library Survey	1993-94
TFS Current (stayers and movers)	1988-89, 1991-92
TFS Former (leavers)	1988-89, 1991-92

The SASS reinterview studies were primarily designed to estimate simple response variance; that is, to measure the consistency in responses by comparing original survey and reinterview responses and then computing the gross difference rate (GDR) and the index of inconsistency (IOI). The L-fold index is also frequently computed to measure the response variance for questions with more than two response categories. High response variance means the respondents are very inconsistent. The 1991-92 TFS reinterview also included extensive reconciliation in an attempt to find out why respondents' answers differed between the original interview and the reinterview. Another emphasis of SASS reinterview studies is simply to identify questions which may need revision in an effort to improve future SASS cycles.

We have already discussed the important assumption that reinterviews be independent in order to estimate response variance accurately (see chapter 1). The customary SASS reinterviews fail to replicate the original interview in two respects (Bushery, Royce and Kasprzyk, 1992, p. 458).

- All SASS reinterviews contained fewer questions than their original counterparts.
- The original SASS surveys used self-administered mail-return questionnaires (with telephone followup of nonrespondents). Prior to the mode effects study conducted in 1991, all the reinterviews were conducted by telephone.

For measuring response variance, the TFS reinterview and extensive reconciliation departed even more from the ideal model. First, the original TFS responses were transcribed onto the reinterview questionnaires. Secondly, the original interviewers were field representatives, while the reinterviewers were supervisory field representatives. Third, the reinterview and extensive reconciliation was administered exclusively by phone, whereas 56 percent of the original cases were self-administered.

School Administrator Reinterviews (SASS 1987-88 and 1990-91)

The School Administrator Survey questions public school principals and private school heads about their demographic characteristics, training, experience, professional background, and their perceptions of school climate and conditions.

Methodology and Design

The School Administrator reinterview studies sampled about 10 percent of all school administrators, who included principals, headmasters, or headmistresses. U.S. Bureau of the Census field representatives conducted the reinterviews by telephone. In 1987-88, the response rate was 87 percent. In 1990-91, it was 93.5 percent (see table 21).

Table 21. -- SASS School Administrator Survey reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
1987-88	1,309	1,139	87%	≅10%	Response variance
1990-91	1,048	980	94%	≅10%	Response variance

SOURCE: Derived from table 1, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 459; Jabine, (1994), *Quality Profile for SASS* (NCES 94-340*)*, p. 3.4.

The 1987-88 School Administrator reinterview asked both attitudinal and factual questions, mirroring the original survey. Because the 1987-88 reinterview results showed high levels of inconsistency for the attitudinal questions, NCES decided to concentrate the 1990-91 reinterview on factual questions "with the aim of improving future cycles of the SASS" (Bushery, Royce and Kasprzyk, 1992, p. 459).

Summary of 1987-88 Results

The 1987-88 school administrator reinterview included 11 factual items. Of the nine for which IOI could be reliably estimated, only one had a low IOI; four had a moderate IOI and four were in the high range (Jabine, 1994, p. 3.15).

The reinterview included 22 attitudinal items: when their IOIs were estimated, none were low, three were moderate, and 19 were high. One set of items asked principals for their views of the relative importance of 13 problem areas occurring in schools. Three of these items had estimated IOIs in the moderate range, while the rest were in the high range. A set of nine items asked principals for their evaluation of the relative influence of teachers, principals, and governing bodies on policies establishing curriculum, hiring new teachers and discipline. All items showed indices in the high range. As already mentioned, while attitudinal items were not entirely eliminated from subsequent SASS reinterviews, they were included to a much more limited extent.

The two questions reinterviewed in both 1987-88 and 1990-91 asked whether administrators had earned a bachelor's degree and a master's degree. The 1987-88 results for reports of bachelor's degrees (GDR 20.3, IOI 98.5) and master's degrees (GDR 9.9, IOI 49.4) prompted NCES to substantially revise the question format in 1991 from one "mark all that apply" question to two "yes or no" questions (Bushery, Royce and Kasprzyk, 1992, p. 460; Jabine, 1994, p. 3.14).

1987-88:

- Which of the following college degrees have you earned? (Mark all that apply.)
  _ Associate degree or vocational certificate
  _ Bachelor's degree
  _ 2nd Bachelor's degree
  _ Master's degree
  _ Professional diploma or education specialist (at least 1 year beyond Master's level)
  _ Doctorate (e.g., Ph.D., Ed.D.)
  _ First professional degree (e.g., M.D., L.L.B., J.D., D.D.S.)
  _ No degree or diploma

1990-91:

- Do you have a Bachelor's degree? [yes/no]
- Do you have a Master's degree? [yes/no]

Summary of 1990-91 Results

The 1990-91 school administrator reinterview included 26 factual items. Of these items, IOI could not be reliably estimated for one, five had a low IOI, 10 had a moderate IOI, and 10 were in the high range. There was no difference in response variance between public and private administrators.

The questions chosen for reinterview were from four subject areas: questions about the degrees the respondents had earned and their major fields of study; questions about training respondents had had before becoming administrators; questions about other school positions the respondent held, job experience, and future plans as an administrator; and questions about the administrators' annual salary. The average IOIs (i.e., L-fold index) for the 1990-91 School Administrator subject areas are shown in table 22.

The two degree questions were already discussed in the section on 1987-88 results. The results for the revised questions were, for "Do you have a Bachelor's degree?," a GDR of 1.3

(there were too few cases to estimate IOI) and, for "Do you have a Master's degree?," a GDR of 1.7 and an IOI of 11.3. The GDRs and IOI showed statistically significant differences from the 1987-88 results.

Table 22. -- 1990-91 SASS School Administrator Survey average rank of reinterview subject area measurements

Subject areas	IOI (L-fold index)	Items with high IOI (above 50%)
Degree information	Moderate	1e
Training	High	4, 6-1, 6-2, 6-0
Other positions held/experience— Future plans	High	3-1, 3-5, 3-6, 3-0, 7a
Annual salary	Low	--

SOURCE: Derived from tables B, E-G, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 5, 7-9.

The training questions and the questions about other positions held/experience-future plans had moderate to high response variance. For 80 percent of all respondents (82 percent of public respondents and 76 percent of private respondents), the annual salary reported in the original interview and that reported in the reinterview were within 5 percent. The annual salary question showed low response variance (Royce, 1994, p. 9).

The items with high L-fold index are listed below by subject area.

Degree Information

- 1e What was your second major or minor field of study? [84 fields within 8 categories] (L-fold index 55.3)

Training

- 4 Prior to becoming an administrator did you participate in any school training or development program for ASPIRING school administrators? [yes/ no] (L-fold index 61.2)

6 Aside from college course work for a degree, have you had any of these types of training for your current position? (Mark all that apply.)

- 6-1 _ Inservice training in evaluation and supervision (L-fold index 63.1)
- 6-2 _ Training in management techniques (L-fold index 63.9)
- 6-0 _ None of the above (L-fold index 68.2)

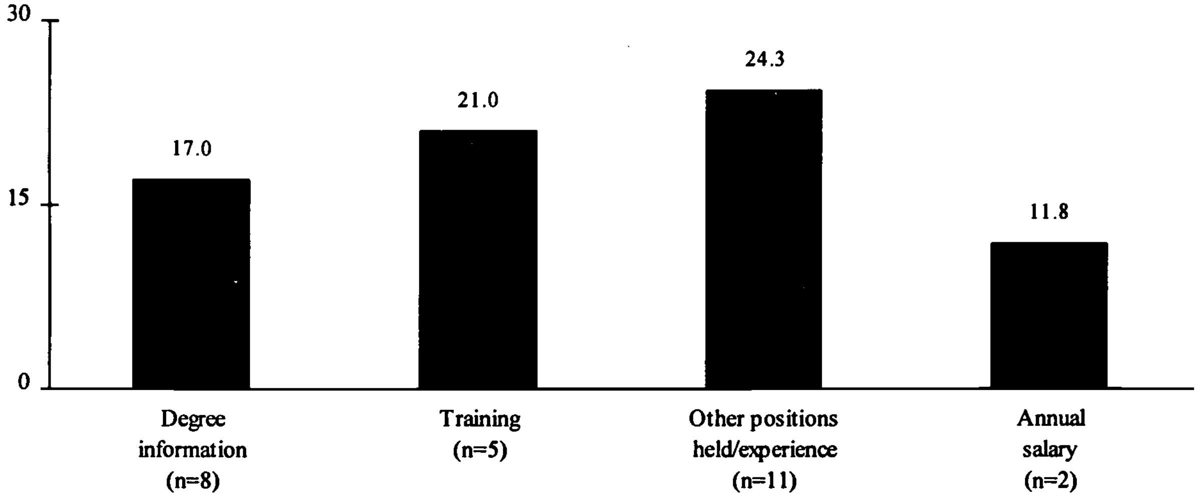
Other Positions Held/Experience - Future Plans

3 What other school positions, if any, did you hold before you became a principal?

• 3-1	Department head or curriculum coordinator (L-fold index 53.5)
• 3-5	Sponsor for student clubs, debate teams (L-fold index 75.9)
• 3-6	Other - Specify (L-fold index 96.6, GDR 65.4)
• 3-0	None (L-fold index 88.1)
• 7a	How long do you plan to remain a principal? (L-fold index 64.7)

Figure 12 illustrates mean unreconciled GDR for the different subject categories within the 1990-91 reinterview. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 12. -- Mean unreconciled GDR, 1990-91 SASS School Administrator Survey



SOURCE: Derived from tables B, E-G, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 5, 7-9.

School Reinterviews (SASS 1987-88 and 1990-91)

This is a survey of public and private schools about school programs and services offered, policies, and conditions; student characteristics; staffing patterns, student-teacher ratios, and teacher turnover. The private school questionnaire includes questions on aggregate demand for teachers (new and continuing); measures of teacher shortage; school policies on teacher salaries, compensation, retirement, and hiring.

Methodology and Design

The School reinterview studies sampled about 10 percent of all school principals. Bureau of the Census field representatives conducted the reinterviews. In 1987-88, the response rate was 87 percent. In 1990-91, it was 91 percent (See table 23).

Table 23. -- SASS School Survey reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
1987-88	1,309	1,139	87%	≅10%	Response variance
1990-91	1,034*	941	91%	≅10%	Response variance

*This number represents those eligible to participate in the reinterview and does not include nonrespondents in the full-scale survey.

SOURCE: Derived from table 1, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 459; Jabine, (1994), *Quality Profile for SASS* (NCES 94-340), p. 2.13.

In 1987-88, NCES used telephone reinterviews, the usual mode for reinterview studies. For the 1990-91 School Survey reinterview study, however, NCES revised reinterview procedures to conduct a mode study. NCES used a mail reinterview for mail respondents (as far as possible) and a telephone reinterview for telephone followup cases. NCES also requested the same respondent complete the reinterview questions as answered the original survey. All of the reinterview questions had lower response variance when the respondent was interviewed originally by mail and reinterviewed by mail, compared to when the respondent was interviewed originally by mail or telephone and reinterviewed by telephone (Royce, 1992).

The results of this analysis were discussed in chapter 2. Details of the study can be found in the appendix. Because of the outcome of the mode study, SASS reinterviews are "designed so that the data collection method [is] the same as that used in the original interview" (Gruber, Rohr and Fondelier, 1996, p. 311).

Summary of 1987-88 Results

Of the fourteen questions appearing in the reinterview survey, eight showed a high index of inconsistency, while six were moderate. There was no difference in response variance between public and private schools.

Questions on bilingual education, English as a second language, and extended day care all had GDRs of 16.2, 16.1, and 9.3, respectively. The question on community had a GDR of 34.7.

Summary of 1990-91 Results

Forty-four questions appeared in the reinterview survey. Of the 36 for which an L-fold index could be reliably estimated, 7 (19 percent) showed a high index of inconsistency, 12 (33 percent) were moderate, and 17 (47 percent) were low. The questions fell in four subject categories: questions about the student population and the teacher population at the school; questions regarding the kind of school, the community in which it was located, and the number of days in the school year; questions about programs that the school offered and the grade levels of instruction at the school; and, questions regarding teaching vacancies in the school for the year, evaluation programs for teachers, and programs to help beginning teachers.

The question on the community in which a school is located again had moderate response variability (GDR 30.4, L-fold index 37.6). NCES decided to drop the question since this information is available from geographic files.

As stated above, all of the reinterview questions had lower response variance when the respondent was interviewed originally by mail and reinterviewed by mail (MM), compared to when the respondent was interviewed originally by mail or telephone and reinterviewed by telephone (MT/TT). The MM and MT/TT scores are listed in Royce (1994).

The average rank of the IOIs for the 1990-91 school reinterview questions are shown by subject area in table 24.

Table 24. -- 1990-91 SASS School Survey average rank of reinterview subject area measurements

Subject areas	IOI (L-fold index)	Items with high IOI (above 50%)
Student population/teacher population	Moderate	8, 9a-b
Type of school/community	Moderate	--
Grades and classes	Moderate	6g, 7-1
Teaching vacancies/teacher programs	High*	10a-b

* The L-fold index could only be estimated reliably for 4 out of the 11 items.

SOURCE: Derived from tables J, M, P, and Q, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 12, 14, 17, and 19.

The items with high IOI are listed below by subject area.

Student population/teacher population

- 8 How many K-12 teachers are NEW to this school this year? (L-fold index 51.5)
- 9a How many K-12 teachers LEFT this school between October 1 of last school year and October 1 of this school year? (L-fold index 53.3)
- 9b Of those K-12 teachers who LEFT this school, how many are no longer teaching in an elementary or secondary school? (L-fold index 54.9)

Grades and classes

- 6g Are diagnostic and prescriptive services--services provided by trained professionals to diagnose learning problems of students and to plan and provide therapeutic or educational programs based upon such services available (to students in this school either during or outside of regular school hours, and regardless of funding source)? [yes (how many)/no] (L-fold index 59.7)

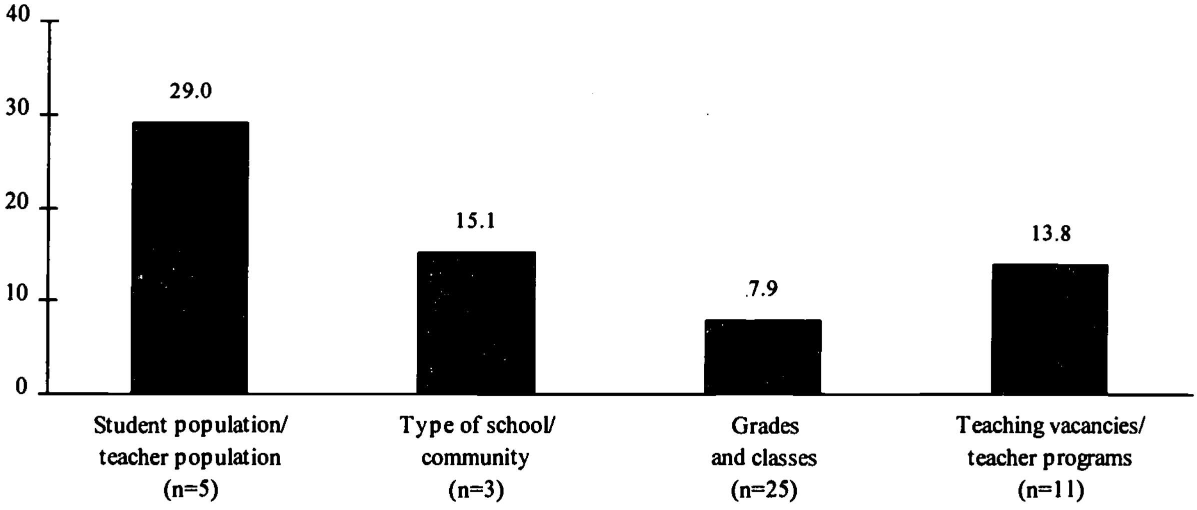
- 7-1 For what grade levels does this school offer instruction?--ungraded [Mark each grade *offered* (whether or not any pupils are enrolled): ungraded/ prekindergarten/kindergarten/1st-12th/postsecondary] (L-fold index 57.9)

Teaching vacancies/teacher programs

- 10a Were there teaching vacancies in this school for this school year; i.e., teaching positions for which teachers were recruited and interviewed? [yes/no] (L-fold index 55.1)
- 10b Did this school have any teaching vacancies this school year that could not be filled with a teacher qualified in the course or grade level to be taught? [yes/no] (L-fold index 52.6)

Figure 13 illustrates the mean unreconciled GDR for the different subject categories within the 1990-91 School Survey reinterview.

Figure 13. -- Mean unreconciled GDR, 1990-91 SASS School Survey



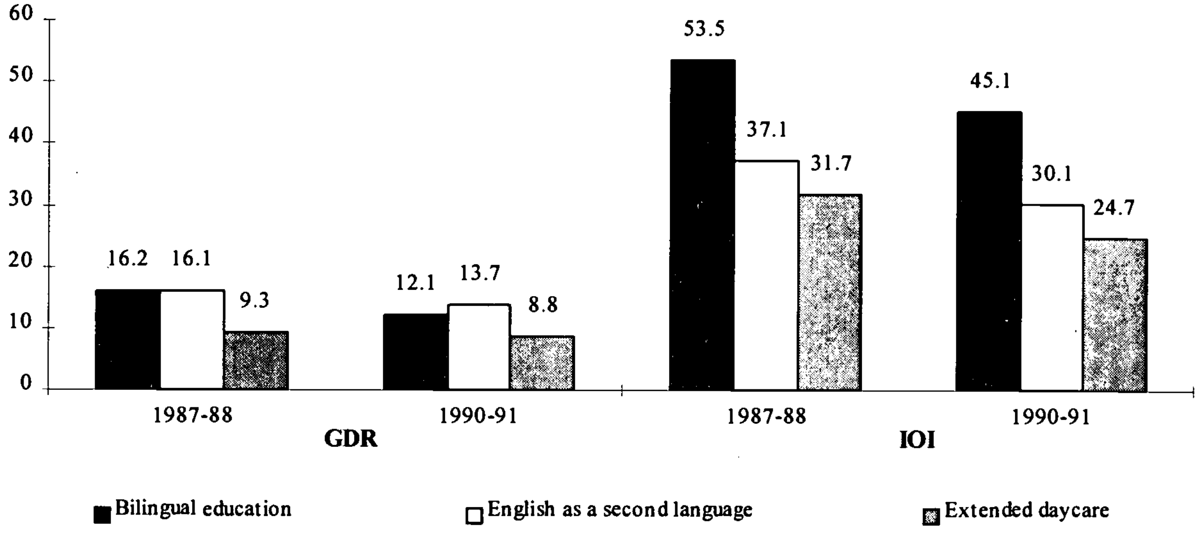
SOURCE: Derived from tables J, M, P, and Q, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 12, 14, 17, and 19.

Comparison of Results for the School Reinterviews: 1987-88 versus 1990-91

Of the 15 factual reinterview items common to both SASS School Survey reinterview studies, 11 received significant revisions in 1990-91. Looking at all of the factual items (most of them not the same in both rounds) from the School Survey reinterview for which IOI could be estimated in each response variance level by year, in 1987-88 there were zero lows, six moderates, and eight highs. In 1990-91, there were 17 lows, 12 moderates, and 7 highs. While it is difficult to make definitive statements due to the different methodologies in the two reinterviews, the authors of the study suggest that "efforts by NCES and Bureau of the Census to improve the questions and instructions for [1990-91] may have had some success" (Jabine, 1994, p. 2.14).

Figure 14 shows GDR and IOI results for three of the reinterview questions included in both 1987-88 and 1990-91. Although these questions were virtually unchanged between the two cycles, there were moderate statistically significant reductions from 1987-88 and 1990-91 in the GDRs for all three items and in the IOIs for two of them. Changes in reinterview methodology may have contributed to some of the reduced response variance. The overall response variance (L-fold index) for the community where the school is located also improved slightly, but remained in the moderate range.

Figure 14. -- Comparison of GDR and IOI results for the 1987-88 and 1990-91 SASS School Survey



SOURCE: Derived from table 5, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 461.

Teacher Reinterviews (SASS 1987-88 and 1990-91)

This is a survey of public and private school teachers about demographic characteristics; teacher preparation and qualifications, including education, training, teaching experience, and certification; career history and plans; teaching assignments; working conditions; and perceptions of school environment and the teaching profession.

Methodology and Design

The Teacher Survey reinterview studies sampled about one percent of all school teachers. Bureau of the Census field representatives conducted telephone reinterviews. In 1987-88, they achieved a 75 percent response rate, in 1990-91, an 83 percent response rate. See table 25.

Table 25. -- SASS Teacher Survey reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
1987-88	1,126	845	75%	≅1%	Response variance
1990-91	980	811	83%	≅1%	Response variance

SOURCE: Derived from table 1, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 459; Jabine, (1994), *Quality Profile for SASS* (NCES 94-340), p. 5.10.

As in the Administrator Survey, the 1987-88 Teacher reinterview asked both attitudinal and factual questions, mirroring the original survey. Because the 1987-88 reinterview results showed high levels of inconsistency for the attitudinal questions, NCES decided to concentrate the 1990-91 reinterview on factual questions "with the aim of improving future cycles" (Bushery, Royce and Kasprzyk, 1992, p. 459).

Summary of 1987-88 Results

The 1987-88 teacher reinterview included 62 items: 20 factual and 42 attitudinal items. Among the 20 factual items, eight did not meet the minimum requirements to compute a reliable IOI estimate. Five of the remaining 12 factual items and 39 out of 42 attitudinal items showed an IOI in the high range. Attitudinal questions covered teachers' views about problems in their schools, their influences on school and classroom policies and practices, and the extent to which school administrators and other teachers had been helpful to them.

The items asking about degrees had response variances high enough to cause some concern. The IOIs for four of the categories which were repeated in the 1990-91 reinterview were "Associate degree or vocational certificate" (IOI 36.9), "Bachelor's degree" (IOI 79.5), "Master's degree" (IOI 8.9), "Education specialist or professional diploma (at least one year beyond Master's level) (IOI 69.8).

Summary of 1990-91 Results

Fifty-six questions in six subject areas were included in the 1990-91 Teacher Survey reinterview study: questions about teaching vacancies in the school for the year, evaluation programs for teachers, and programs to help beginning teachers; questions about the teachers' teaching assignment and the grade levels of the students in the respondents' classes; questions about the respondents' teaching experience; questions about the respondents' degrees; questions about any training the respondents may have taken, and what type of teaching certificate they had; questions about respondents main activities the year before they began teaching, the respondents' main activity during the last school year, and what the respondents' main activity would be the next school year; and questions about the respondents' teaching salary and other salaries they may have received during the school year.

Of the 56 questions on the reinterview survey, 53 were factual and 3 were attitudinal items. Looking at IOI rates among the three opinion items, one was in the high range and two were in the moderate range. For the 53 factual questions, if we set aside Question 15 which asked about grade levels for the responding teacher's current classes and whose 16 possible response categories were treated as 16 separate items for estimating IOI, there remain 40 items to consider. Eight were in the low range, 14 in the moderate range, and 10 in the high range (eight did not meet the minimum requirements to compute a reliable IOI estimate). All 16 possible response categories for Question 15 had indices in the low range (Jabine, 1994, p. 5.29).

The average rank of IOIs for the 1990-91 Teacher subject areas are shown in table 26, and the items with a high IOI are listed below by subject area.

Table 26. -- 1990-91 SASS Teacher Survey average rank of reinterview subject area measurements

Subject areas	IOI	Items with high IOI (above 50%)
Assignment and activity (this year)	Moderate	10a-2
Teaching experience	Moderate	16
Degree information	Moderate	9a, 9b-1, 9b-2
Training/teaching certificate	High	11, 12, 13b, 14a
Main activity (past and future)	Moderate	--
Salaries	Moderate	19a(2), 19b(3)

SOURCE: Derived from tables R, S, V, W, X, and Y, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 21, 23, 25-28.

Assignment and activity (this year)

- 10a-2 Is your teaching assignment equally divided between two fields? [yes/ no] (L-fold index 75.9)

Teaching experience

- 16 How long do you plan to remain in teaching? [Mark only one: as long as I am able/until I am eligible for retirement/will probably continue unless something better comes along/definitely plan to leave teaching as soon as I can/ undecided at this time] (L-fold index 66.6)

Degree information

- 9a Do you have any other type of degrees? [yes/no] (L-fold index 51.9)
- 9b-1 Associate degree (L-fold index 54.2)
- 9b-2 Education specialist or professional diploma (at least one year beyond Master's level) (L-fold index 62.7)

Training/teaching certificate

- 11 Have you ever taken any college level courses in teaching methods or education? [yes/no] (L-fold index 68.8)
- 12 Have you ever taken any college level courses in the subject area which is your MAIN teaching assignment? [yes/no] (L-fold index 73.8)
- 13b What type of certification do you hold in this field? [Mark only one: advanced professional certificate/regular or standard state certification/ probationary certification/temporary, provisional, or emergency certification] (L-fold index 52.6)
- 14a During your first year of teaching, did you participate in a formal teacher induction program? [yes/no] (L-fold index 56.2)

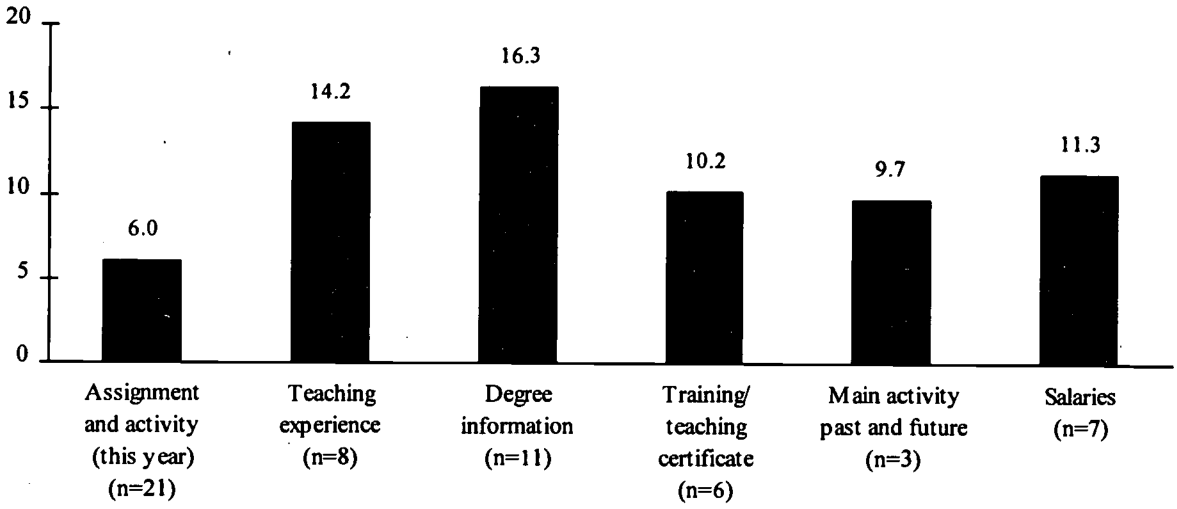
Main activity (past and future)—no items with high IOI.

Salaries

- 19a(2) During the summer of 1990, did you have any earnings from working in a nonteaching job in this or any other school? [yes ($)/no] (L-fold index 80.0)
- 19b(3) Have you earned income from any other sources this year? [yes ($)/no] (L-fold index 56.8)

Figure 15 illustrates mean unreconciled GDR for the different subject categories within the 1990-91 Teacher Survey reinterview.

Figure 15. -- Mean unreconciled GDR, 1990-91 SASS Teacher Survey



SOURCE: Derived from tables R, S, V, W, X, and Y, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), pp. 21, 23, 25-28.

Comparison of Results for the Teacher Reinterviews: 1987-88 versus 1990-91

In 1987-88, about two-thirds of the items included were attitudinal questions and nearly all of them (39 out of 42) had indices in the high range. These items covered teachers' views about topics like problems in their schools, their influence on school and classroom policies and practices, and the extent to which school administrators and other teachers had been helpful to them. In 1990-91, only three attitudinal items were covered in the teacher reinterview, one of which had an index in the high range.

Table 27. -- Comparison of 1987-88 and 1990-91 SASS Teacher Survey summary of IOI[1] results

		IOI			
Round and type of item	**Number of items**	**Low**	**Moderate**	**High**	**NA[2]**
1987-88:					
Factual	20	3	4	5	8
Attitudinal	42	--	3	39	--
1990-91:					
Factual	53	21	14	10	8
Factual, excluding item 15	37	8	14	10	5
Attitudinal	3	--	2	1	--

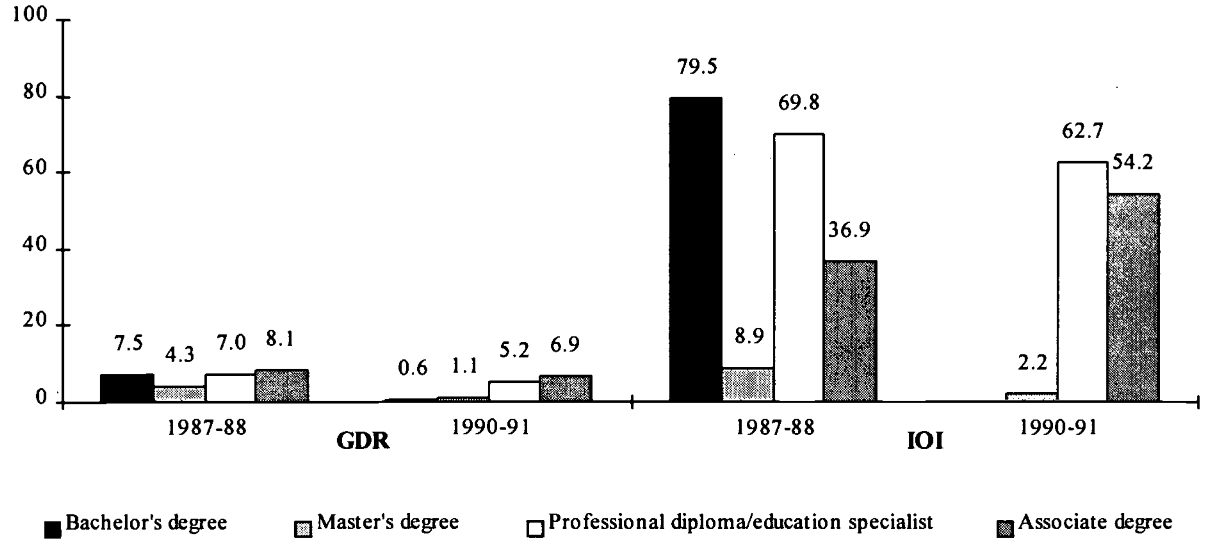
[1] Each item either had closed, multiple-response categories or was converted to the equivalent by assigning class intervals to open-end responses. For items with more than 2 response categories, the L-fold index of consistency was estimated.

[2] Did not meet the minimum requirements to compute a reliable IOI estimate.

SOURCE: Derived from table 5.9, Jabine, (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS)* (NCES 94-340), p.5.29.

Figure 16 shows the reinterview results for the questions on teachers' degrees earned for 1987-88 and 1990-91. The question format was substantially revised between the two rounds. The 1987-88 question provided a list of possible degrees and asked the respondent to "mark all that apply." The 1990-91 question asked "Do you have a bachelor's degree?" If yes, the next question asked "Do you have a master's degree?" The remaining degrees (associates, doctor's etc.) used a "mark all that apply" approach. The results suggest that the direct question format produces more reliable data for this variable.

Figure 16. -- Comparison of GDR and IOI results for "degrees earned", 1987-88 and 1990-91 SASS Teacher Survey



Note: IOI was not calculated for Bachelor's degree in 1990-91.

SOURCE: Derived from table 3, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 460.

Both reinterview rounds also included questions on teaching assignment, years in teaching, and plans to remain in teaching (an attitude type question). None of these questions exhibited significantly improved response variance.

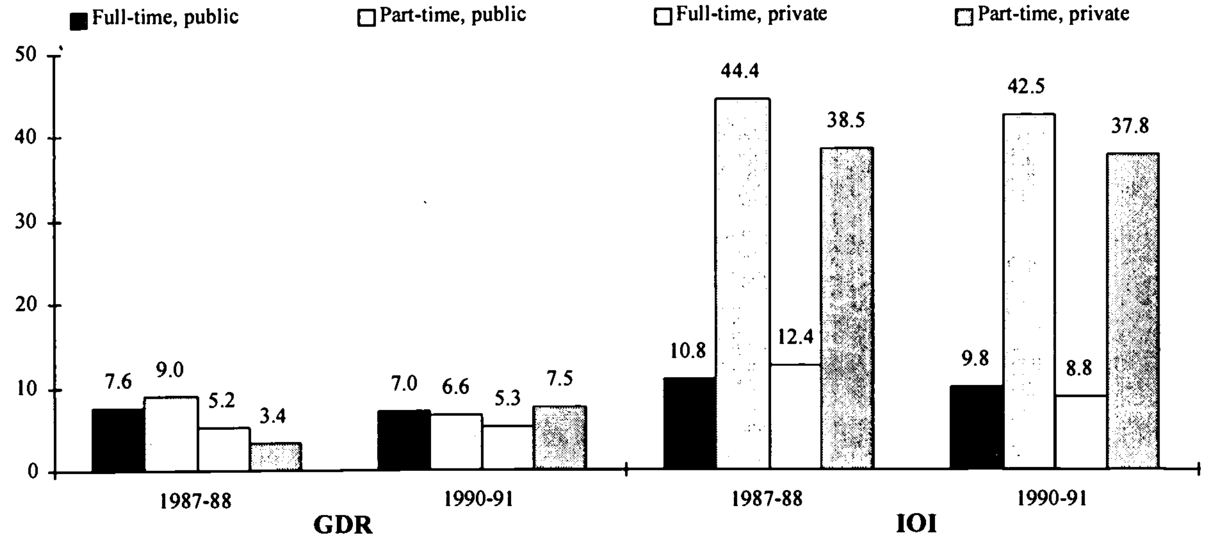
The teaching assignment questions reinterviewed in 1988 and 1991 were similar but not strictly comparable, making it difficult to compare the two, but response variance on the number of full-time teachers showed no significant change between 1987-88 and 1990-91.

The 1987-88 "years teaching" questions asked, "...how many years have you worked as a full-time teacher in public and/or private schools?" (repeated for part-time) and provided a cross-tabulation for the respondent to complete:

	Years full-time	Years part-time
Public		
Private		

In 1991, NCES changed the format to ask four separate questions: How many years have you worked as a full-time teacher in private schools? How many years have you worked as a part-time teacher in private schools? etc. NCES grouped the responses into the categories of interest: less than three years, three to nine years, 10 to 20 years, and more than 20 years. Unfortunately, there was no improvement. See figure 17.

Figure 17. -- Comparison of GDR and IOI results for "years teaching", 1987-88 and 1990-91 SASS Teacher Survey



SOURCE: Derived from table 4, Bushery, Royce and Kasprzyk, (1992), "The Schools and Staffing Survey: How Reinterview Measures Data Quality," in *American Statistical Association 1992 Proceedings of the Section of Survey Research Methods*, p. 460.

The response variance for one question reinterviewed in both cycles proved worse in 1990-91 than in 1987-88. Since the question, "How long do you plan to remain in teaching?" was not changed, NCES speculated that the teachers' attitudes in 1991 were "less stable than in 1988" (Bushery, Royce and Kasprzyk, 1992, p. 461).

Summary Comparison of the SASS 1987-88 and 1990-91 Reinterviews

In the 1987-88 SASS, NCES could estimate the IOI reliability for 35 of the 45 (78 percent) factual questions reinterviewed. For the 1990-91 SASS, NCES reliably estimated the IOI for 109 of the 126 (87 percent) factual questions reinterviewed (Bushery, Royce and Kasprzyk, 1992, p. 459). There was no difference in response variance between public and private administrators, schools, or teachers (Royce, 1994, p. 2).

Thirty-nine percent of the 1990-91 SASS reinterview questions showed low response variance. This was significantly better than the 11 percent of reinterview questions for SASS 1987-88 with low response variance (see table 28). Moreover, there was a 23 percentage point difference between 1990-91 and 1987-88 SASS items with a high response variance (26 percent versus 49 percent) (Royce, 1994, p. 1).

It is important to note that the comparisons across 1987-88 and 1990-91 are not strictly comparable. Different sets of questions were used for the two interviews. Among the 15 factual questions common to both years, 11 showed significant revisions in 1991. Four of these items displayed reduced response variance, which indicates "question improvement efforts have paid off, at least partially" (Bushery, Royce and Kasprzyk, 1992, p. 459).

Table 28. -- Summary of 1987-88 and 1990-91 SASS reinterview response variance results*

	Low		Moderate		High	
	Number	Percent	Number	Percent	Number	Percent
All three components:						
1988	4	11%	14	40%	17	49%
1991	43	39%	38	35%	28	26%
Administrator Survey:						
1988	1	11%	4	44%	4	44%
1991	5	20%	10	40%	10	40%
School Survey:						
1988	0	0%	6	43%	8	57%
1991	17	47%	12	33%	7	19%
Teacher Survey:						
1988	3	25%	4	33%	5	42%
1991	21	44%	16	33%	11	23%

*Questions for which index could be reliably estimated.

SOURCE: Derived from table A, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper 94-03), p. 1.

Library Survey Reinterview Study (SASS 1993-94)

School library media centers have become a topic of increasing interest and concern to a number of education policymakers and researchers over the past decade. The 1990-91 SASS obtained selected basic information on the availability, staffing, and role of school library media centers throughout the nation. The 1993-94 SASS followed this initial effort with an expanded survey component on school library media centers, including data collection on the background, training, and attitude of school librarians and the expenditures, materials, equipment, and services of school library media centers throughout the nation (Ingersoll, Han and Williams, 1994, pp. 1 and 4).

Methodology and Design

The Library Survey reinterview sample consisted of 1,780 library media centers pre-selected from library media centers in the final 1993-94 Schools and Staffing Survey: 833 public schools and 947 private schools. However, there were only 1,343 eligible schools among those (see table 29).

In the original interview, questionnaires were first mailed to the school's library media center. If the original questionnaire was not returned by the due date, a second mailout was sent. If the original questionnaire was still not returned, then the interview was attempted by computer-assisted telephone interviewing (CATI). The reinterview was conducted using the same mode as the original interview; that is, if the original interview was completed by mail, the reinterview was completed by mail; if the original interview was completed by CATI, the reinterview was completed by CATI. For both modes of reinterview, operators attempted to reinterview the same respondent who filled out the original questionnaire. The overall response rate was 72 percent.

Table 29. -- SASS Library Survey reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
1993-94	1,343	959	72%	23%	Response variance

SOURCE: Feindt, (1996), *Reinterview Report: Response Variance in the 1993 Library Survey,* p. 4. Monaco et al., (forthcoming), *An Analysis of Response Rates in the 1993-94 Schools and Staffing Survey.*

The reinterview instrument contained a subset of questions selected from the original LS-1A, Public School Library Media Center, and LS-1B, Private School Library Media Center, questionnaires. Two measures of response variance were calculated for each survey item: L-fold GDR and L-fold index. The library reinterview also estimated net difference rates (NDRs) as an indication of how well the reinterview met the model assumptions.

Summary of Results

The reinterview instrument contained 32 items, but several of the questions had multiple response categories which were each analyzed separately. Thus, the five questions in one subject area accounted for 28 items. Subject areas included in the Library Survey were the number of staff and their qualifications; the library media center's acquisitions, holdings, and expenditures; the school and the library media center's computer technology; library media center facilities; and check-out privileges.

Staffing. The first three questions asked the number of state-certified media specialists working in the school's library center (Question 1a), the number who are not certified as library specialists (Question 1b) and the number of paid employees working in the school's library media center (Question 1c). These items were each analyzed as seven separate questions

- none
- full-time
- at least 3/4 time but less than full time
- at least 1/2 time but less than 3/4 time
- at least 1/4 but less than 1/2 time
- less than 1/4 time
- total

Respondents seemed to have some difficulty using these subcategories, and it was suggested that the next survey use more concrete definitions, such as giving the number of hours which define "working at least 3/4 time but less than full-time," and explaining exactly what is meant by "professional staff member." In addition, the "none" category did not appear to be used correctly. It displayed moderate L-fold GDR for all three items (22.8, 26.9, and 21.2) and moderate to high L-fold indices (43.9, 59.3, and 45.5). "These questions need a complete overhaul in format" (Feindt, 1996, p. 2).

Collection and expenditures. Respondents were asked to report acquisitions, holdings, and expenditures for the library media center during the 1992-93 school year in six categories: books, serial subscriptions, video materials, other audio-visual materials, microcomputer software, and CD-ROMs (Question 5). The final question in this section asked for the total expenditure for computer hardware for the school's library media center (Question 7a). The average L-fold GDR was 28.1 and the average L-fold index was 37.3 (moderate). Subscriptions—acquired showed the highest rates, a GDR L-fold of 44.4 and an L-fold index of 49.2.

Among the other problems noted with questions in this section were evidence of "heaping"—the tendency of respondents to "heap" responses at rounded estimates instead of providing exact responses—at the $1,000 and $2,000 intervals and low agreement rates between the reinterview responses and the original responses. It appeared that respondents had difficulty giving consistent and exact numbers. Three suggestions made to improve responses in the next survey were to ask respondents to round to the nearest "X" dollar or to provide respondents with a range of values or to keep the original question wording but to adjust the original answer categories to reflect the expected heaping. There were also suggestions about wording in some questions; for example, that the term "personal computer" or "PC" should be used instead of microcomputer (Feindt, 1996, pp. 24-32).

Technology. This section examined whether the school had any microcomputers (Question 11a), the number of microcomputers under the supervision of the library media staff (Question 11b), and whether the library media center had various equipment or services (Question 12c). The average L-fold GDR was 6.4, while the average L-fold index was 23.2 (moderate). Question 11a, whether the school had any computers, showed the highest L-fold index (48.4). It was felt that even though response variance in this section was fairly low, it might be worthwhile to provide definitions for some of the terms used; for example, what does "automated circulation system" mean? (Feindt, 1996, pp. 34-5).

Facilities. This section examined library media center facilities in three items which asked: how the library media center is organized (Question 17), seating capacity of the library center (Question 18), and types of spaces available (Question 19). The average L-fold GDR was 20.6, while the average L-fold index was 52.1 (high).

The last item, the types of seating capacity, was divided into 10 subquestions. For the L-fold index, three were in the moderate range, while the rest were high. All 10 subquestions had NDRs statistically different from zero, suggesting the response error model assumptions may not have held for these questions. One reason suggested for the high response variance was the questionnaire format: a "mark all that apply" format is usually not the best format to use. Sudman and Bradburn suggest caution in interpreting this kind of question:

> *While the presence of a check mark indicates a positive instance, the omission of it may indicate that in fact the adjective does not apply, or that respondents did not notice [the response option] because they were hurrying over the list, or that they were not sure whether it would apply* (Sudman and Bradburn, 1982, p. 168, quoted in Feindt, 1996, p. 38).

It was recommended not to use this type of question in the future.

Check-out privileges. The last section of the reinterview examined the maximum number of various types of materials students may check out at a time (Question 27) and which groups of people are allowed to check out materials (Question 28). The average L-fold GDR was 18.0, while the average L-fold index was 32.3 (moderate).

The maximum number of materials that a student may check out was analyzed as six separate questions. The mostly moderate L-fold indices may have been due to confusing response categories. The second item, persons allowed to check out materials, was divided into four response categories: prekindergarten students, kindergarten students, parents, and other members of the community. Prekindergarten showed an L-fold GDR of 18.4 and an L-fold index of 33.1. It was thought that there may have been some confusion between "No" and "No prekindergarten (or kindergarten) at this school."

The average IOI ranks for the library reinterview subject areas are shown in table 30.

Table 30. -- 1993-94 SASS Library Survey average rank of reinterview subject area measurements

Subject areas	L-fold IOI	Items with high IOI (above 50%)
Staffing	Moderate	1b (2 response categories)
Collection and expenditures	Moderate	--
Technology	Moderate	--
Facilities	High	19 (7 response categories)
Scheduling and transactions	Moderate	--

SOURCE: Derived from tables B-M, Feindt, (1996), *Reinterview Report: Response Variance in the 1993 Library Survey,* pp. 13-47.

The items with high IOIs are listed below by subject area.

Staffing

1b How many professional staff members working in this school's library media center are NOT certified as library media specialists?

- None (L-fold index 59.3)
- Less than 1/4 time (L-fold index 50.7)

Other response categories: Full-time, at least ¾ time but less than full-time, At least ½ time but a less than ¾ time, At least ¼ time but less than ½ time

Facilities

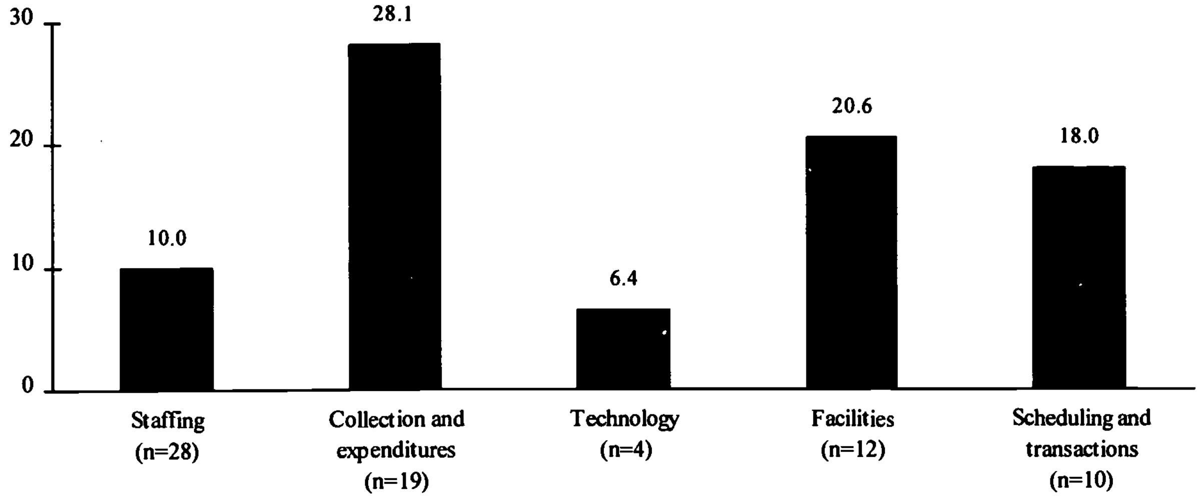
19 Which of these types of spaces are available in the library media center?

- Individual reading, viewing, and listening (L-fold index 58.1)
- Small group (5 or less) activity areas? (viewing or listening) (L-fold index 68.2)
- Large group (more than 5) activity area (L-fold index 67.1)
- Production areas for classroom teachers (L-fold index 52.4)
- Production areas for students (L-fold index 60.5)
- Storage (equipment, etc.) (L-fold index 54.9)
- None of the above (L-fold index 63.1)

Other response categories: conference rooms/computer access or lab/workroom for library media staff

Figure 18 shows the mean unreconciled GDR for the different question categories within the 1993 library reinterview.

Figure 18. -- Mean unreconciled GDR, 1993-94 SASS Library Survey



SOURCE: Derived from tables B-M, Feindt, (1996), *Reinterview Report: Response Variance in the 1993 Library Survey*, pp. 13-47.

Teacher Followup Survey (TFS) Reinterview Studies

The Teacher Followup Survey (TFS) is an important component of the Schools and Staffing Survey (SASS). It is treated separately because it is implemented one year after SASS. The survey identifies and collects national-level data from public and private school teachers who have remained in the same school as the previous year (stayers), as well as those who have changed schools (movers), and those who have left the teaching profession (leavers). These data are used to provide information about teacher attrition and retention in the public and private schools and to project teacher demand.

The questionnaire for continuing teachers asks primary occupational status (full-time, part-time); primary teaching assignment, by field; teaching certificate; level of students taught; school community type; reasons for leaving previous school; possible areas of dissatisfaction; new degrees earned or pursued; expected duration in teaching; level of satisfaction; marital status; number of children; academic year base salary; and combined family income. The questionnaire for former teachers repeats many of these categories, asking primary occupational status (full-time); type of business; primary activity; time planning to spend in current job; new earned degrees, by type and field; plans for returning to teaching; reasons for leaving teaching; possible areas of dissatisfaction; salary; marital status; number of children; and combined family income.

TFS was first conducted in the 1988-89 school year with a sample from the 1987-88 SASS. This report examines the TFS reinterview studies for 1988-89 and 1991-92. The reinterview study for 1988-89 was similar to other SASS reinterviews designed to measure response variance estimates. However, for the 1991-92 TFS, the reinterview study also attempted to uncover why respondents' answers differed between the original TFS and the reinterview by employing an extensive structured reconciliation. Below we provide the study design and results for the reinterviews conducted and a comparison of results for the two reinterview studies.

TFS 1988-89 Reinterview Study

Methodology and Design

The purpose of the 1988-89 TFS reinterview was to improve the survey questions and to measure response error. Two reinterview samples were selected: 750 current teachers (stayers and movers), and 750 former teachers (leavers). The 1988-89 TFS reinterview had an overall response rate of 81 percent. Data were collected by Bureau of the Census field representatives over the telephone.

Table 31. -- 1988-89 TFS reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
Total	1,497[1]	1,220[2]	81%	18%	Response variance
Current	--	687	--	--	Response variance
Former	--	450	--	--	Response variance

[1] There were three missing cases.

[2] The 1,220 completed reinterviews include 83 respondents who switched leaver/stayer status between the original interview and the reinterview.

Source: Derived from table C, Royce, (1990), *1989 Teacher Followup Survey (TFS) Reinterview Response Variance Report*, p. 5. Faupel, Bobbitt and Friedrichs, (1992), *1988-89 Teacher Followup Survey Data File User's Manual* (NCES 92-058).

Summary of Results

The current teacher reinterview contained 32 items (19 factual and 13 opinion), mostly addressing teachers' opinions, attitudes, and expectations. The former teacher reinterview contained 24 items (2 factual and 22 opinion). Among current teachers, seven out of the nine factual items for which an index of inconsistency could be computed were in the moderate or low range. The two factual items with high IOI related to teacher certification in the field of their primary and secondary teaching assignments. Eight out of nine attitudinal items were in the high range. Among former teachers, 13 out of the 20 items for which an index of inconsistency could be computed were in the high range (see table 32).

Table 32. -- 1988-89 TFS summary IOI[1] results

		IOI			
Teacher status and type of item	**Number of items**	**Low**	**Moderate**	**High**	**NA[2]**
Current (movers and stayers):					
Factual	19	3	4	2	10
Attitudinal	13	--	1	8	4
Former (leavers):					
Factual	2	1	1	--	--
Attitudinal	22	--	7	13	2

[1] For items with more than 2 response categories the L-fold index of inconsistency was estimated.

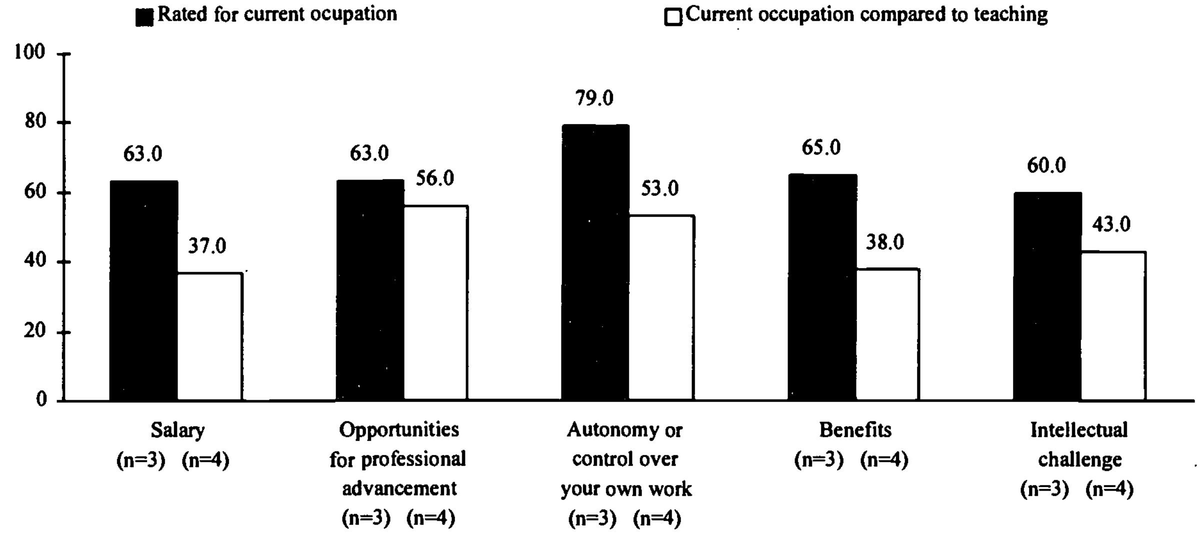
[2] Did not meet the minimum requirements to compute a reliable estimate of the index of inconsistency.

SOURCE: Derived from table 6.3, Jabine, (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS)* (NCES 94-340), p. 6.10.

Former teachers were asked to rate their current occupations on several aspects of job satisfaction both in an absolute and relative sense compared to teaching. Figure 19 presents the indices of inconsistency for these items. Even though the indices were in the moderate to high range for all items, respondents were more consistent in providing comparative ratings on a three-point scale ("Current occupation compared to teaching") than they were in providing absolute ratings on a four-point scale ("Rated for current occupation").

- 26 How would you rate teaching relative to your current primary occupation in terms of each of the following aspects? Please indicate (a) Better in teaching, (b) Better in current position, or (c) No difference...
- 27 How satisfied are you with each of the following aspects of your current job? Are you (a) Very satisfied, (b) Somewhat satisfied, (c) Somewhat dissatisfied, or (d) Very dissatisfied with...

Figure 19. -- IOI for selected opinion items for leavers, 1988-89 TFS reinterview



SOURCE: Derived from table 6.4, Jabine, (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS)* (NCES 94-340), p. 6.11. Based on tables G and H, Royce, (1990), *1989 Teacher Followup Survey (TFS) Reinterview Response Variance Report,* pp. 9-10.

For items on current teachers' satisfaction with their jobs and on former teachers' satisfaction with their current jobs, all of which used a four-point scale, IOIs were re-estimated with the four response categories collapsed into two: satisfied and dissatisfied. The resulting indices in many cases moved from the high to the moderate range.

TFS 1991-92 Reinterview Study

Methodology and Design

In 1992, the TFS Reinterview and Extensive Reconciliation was designed to go beyond measuring the response variance of selected questions. The ultimate goals were: 1) to determine if respondents' answers differed because they were having difficulty comprehending the questions or the response tasks asked of them, and 2) to make recommendations for correcting these difficulties (Jenkins and Wetzel, 1995, p. 3).

Out of the two reinterview samples selected, 678 current teachers (stayers and movers) and 747 former teachers (leavers) were eligible. A total of 629 cases were completed among current teachers and 685 cases completed among former teachers, for response rates of 93 percent and 92 percent, respectively (both higher than 1988-89 results).

Table 33. -- 1991-92 TFS reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
Total	1,425	1,314	92%	23%	Response bias & variance
Current	678	629	93%	14%	Response bias & variance
Former	747	685	92%	49%	Response bias & variance

Source: Derived from table 1, Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), p. 4.

A subset of questions from both the Current Teacher (TFS-3) and the Former Teacher Follow-up Survey (TFS-2) were chosen for reinterview and reconciliation. Data were collected by supervisory field representatives over the telephone. All reinterview questions were asked before any of the reconciliation questions. When there were differences between the reinterview responses and the original answers, reinterviewers were instructed to ask reconciliation probes designed to explain the reasons for differences in respondents' answers. The reconciliation consisted of the following

- Determining the correct answer
- Determining if there was a difference
- Probing with questions to find out the reason for the difference
- Recording and keeping track of the reasons for the differences

To analyze the data from the 1991-92 TFS reinterview and extensive reconciliation, NCES computed three measures: GDR, IOI, and NDR. However, this reinterview study deviated from the assumptions necessary to ideally measure both response variance and bias. Therefore, the estimates of response variance (GDR and IOI) are greatly understated, while the effects on NDR are unknown. For an in-depth discussion of each item, see Jenkins and Wetzel (1995).

Summary of Results

Current teachers. The current teacher survey reinterview contained a total of 42 items for which variance measurements were estimated, when all response categories were included (Question 8 included 16 response categories). Among the 35 items for which an index of inconsistency was calculated, 33 were in the low range, two were in the moderate range (teacher assignment and teaching certificate), and none were in the high range; seven did not meet minimum requirements to compute a reliable estimate of the index of inconsistency. See table 34. Question 31, total combined income, seemed to cause respondents difficulty.

Table 34. -- 1991-92 TFS current teachers: average rank of reinterview subject area measurements

Subject areas	IOI	Items with high IOI (above 50%)
Employment and teaching status	Low	--
Incentives and compensation	Low	--
Background	Low	--

SOURCE: Derived from tables 2-34 , Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), pp. 11-39.

Among the recommendations that arose from this reinterview study were to rearrange and reword the answer categories for the teaching assignment question as follows:

___ You provide instruction at more than one school (e.g., you are an itinerant, traveling, co-op, or satellite teacher).

___ You fill the role of a regular teacher on a long-term basis, but you are still considered a substitute (i.e., you are a long-term substitute teacher).

___ You are a regular full-time or part-time teacher.

This arrangement gives respondents the choices of "itinerant" and "long-term substitute" before the choice of "regular," which should reduce the number of inappropriate choices of "regular." This wording also prominently displays the definitions for itinerant and substitute and minimizes technical terminology.

Another recommendation was to rearrange the questions on the TFS-3 from

- 5a Main teaching assignment field
- 5b Teach classes in other assignment fields
- 6a Teaching certificate in main field
- 6b Type of certificate
- 6c Certificate granted within last 12 months
- 7a Teaching certificate in other field
- 7b Type of certificate
- 7c Certificate granted within last 12 months

to

- 5a Main teaching assignment field
- 6a Teaching certificate in main field
- 6b Type of certificate
- 6c Certificate granted within last 12 months
- 5b Teach classes in other assignment fields

• 7a	Teaching certificate in other field
• 7b	Type of certificate
• 7c	Certificate granted within last 12 months

This new arrangement separates the questions about a teacher's main assignment from the questions about other assignments. A "no" response to question 5b will cause a skip to the next topic (Question 8, "In what grade levels are the students in your classes at THIS school?"). This skip pattern will also improve question 7a by eliminating the need for a "not applicable" answer category (Jenkins and Wetzel, 1995, p. 2).

It was felt that further research was needed before suggestions could be made on improving the income questions. In particular, a better understanding of respondents' use of records is needed.

Former teachers. The former teacher survey contained 21 items. Among the 15 items for which IOI could be calculated, 13 were in the low range and two were in the moderate range (total combined income and person other than spouse or children who are dependent for more than half their financial support); six did not meet minimum requirements to compute a reliable estimate of the index of inconsistency (see table 35).

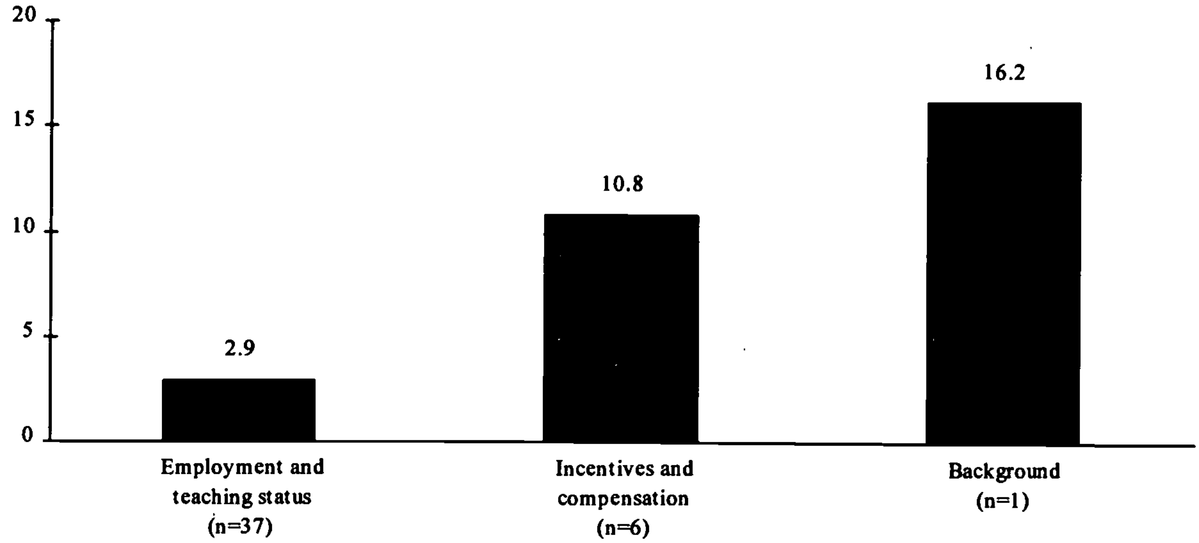
Table 35. -- 1991-92 TFS former teachers: average rank of reinterview subject area measurements

Subject areas	IOI	Items with high IOI (above 50%)
Employment status	Low	--
Educational activities and future plans	Low	--
Background information	Low	--

SOURCE: Derived from tables 35-58, Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), pp. 41-55.

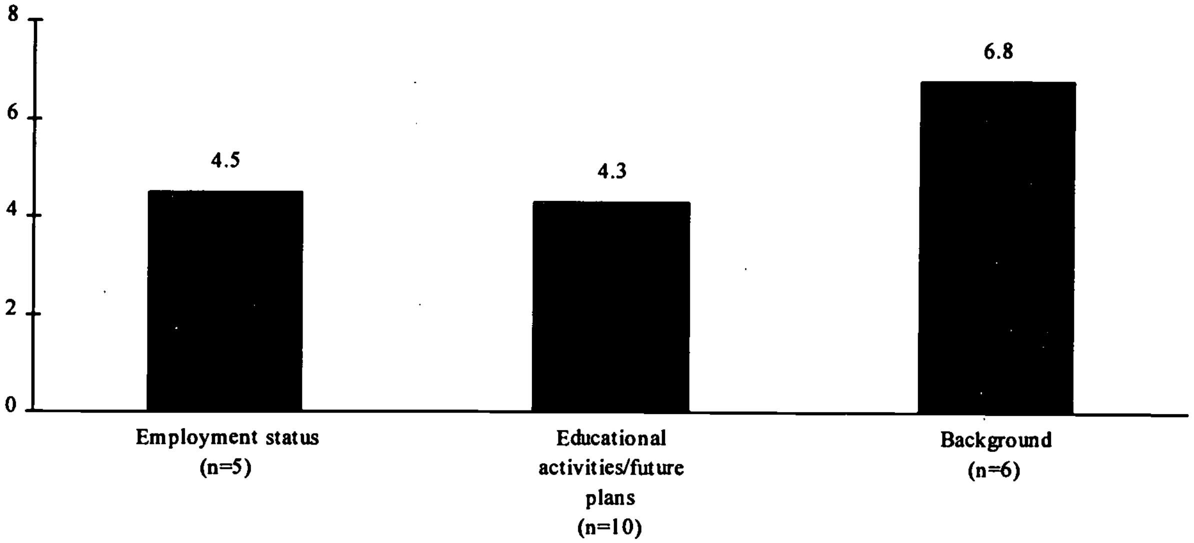
Figure 20 depicts the mean GDRs for current teachers in TFS 1991-92, by question category and figure 21 depicts the mean GDRs for former teachers in TFS 1991-92, by question category.

Figure 20. -- Mean unreconciled GDR, 1991-92 TFS current teachers



SOURCE: Derived from tables 2-34 , Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), pp. 11-39.

Figure 21. -- Mean unreconciled GDR, 1991-92 TFS former teachers



SOURCE: Derived from tables 35-58, Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), pp. 41-55.

Extensive reconciliation. In general, the 1991-92 TFS reinterview and extensive reconciliation did not provide enough differences between the original and reinterview responses to produce many reasons for differences. Jenkins and Wetzel attributed this to the 1991-92 reinterview study's having employed a dependent-type reinterview (i.e., an interview where the original response is transcribed onto the reinterview questionnaire). They believe the results suggest that reinterviewers did not always ignore the original responses. However, other problems were that the reconciliation produced too many open-ended reasons for differences and too many general reasons for differences.

Jenkins and Wetzel (1995, pt. 2, p. 1) proposed three alternatives: 1) an independent reinterview followed by a third-visit small-scale unstructured reconciliation; 2) an independent reinterview with a large-scale extensive reconciliation conducted at the same time using computer-assisted interviewing (CATI or CAPI); and 3) a monitored independent CATI reinterview followed by a third-call reconciliation.

Comparing the 1991-92 Model to the 1988-89 Methodology

Table 36 lists the fourteen questions from the 1991-92 TFS Reinterview and Extensive Reconciliation that are the same as those from the 1988-89 TFS Reinterview. All but two of the 1991-92 questions have before-reconciliation GDRs significantly lower than their 1989 counterparts at the 90 percent confidence level [Main teaching assignment-check box (1988-89 3.8; 1991-92 3.0) and type of certificate (1988-89 10.9; 1991-92 9.0)]. An asterisk (*) in the 1992 column indicates significance.

Lower GDRs in 1992 may have occurred for two reasons. First, the 1989 methodology used an independent reinterview, whereas the 1992 methodology used a dependent-type reinterview combined with reconciliation. Past research has shown that having the original responses visible or available to the reinterviewers results in fewer differences. In addition, respondents' memories may also interfere with independence; they may wish to appear consistent rather than admit they misunderstood something. This may result in artificially depressing reinterview measures.

Second, the 1989 reinterview used field representatives in both the original interview and reinterview. In contrast, 1992 procedures used supervisory field representatives to conduct the reinterview and extensive reconciliation, since field representatives were not trained to conduct a reconciliation. Jenkins and Wetzel hoped that the supervisors would ignore the original response, but the data suggest that this was not the case (1995, pt. 2, p. 6).

Table 36. -- Before reconciliation GDRs: 1988-89 versus 1991-92

Question	GDR percentage	
Title	**1988-89**	**1991-92**
TFS-3R:		
Main teaching assignment - *Field*	11.2	1.6*
Main teaching assignment - *Check box*	3.8	3.0
Teach classes in other fields - *Yes/No*	13.5	3.7*
Teach classes in other fields - *Field*	17.3	3.4*
Certificate in state in main assignment field	7.8	1.5*
Type of certificate	17.6	6.3*
Certificate in state in main assignment field	24.5	13.4*
Type of certificate	10.9	9.0
Teaching is same school	5.1	1.6*
Academic base year teaching salary - *Dollars*	28.0	14.8*
Added compensation from school - *Dollars*	59.0	20.4*
TFS-2R:		
Main occupational status	17.6	6.9*
Lifetime teaching certificate	7.7	3.3*
TFS-3R:		
Still teaching*	6.8	1.3*

*We combined the response counts from the 1992 TFS-3R and 2R questionnaires to correspond to the combined 1989 results for this question.

SOURCE: Derived from table 1, Jenkins and Wetzel, (1995), *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation* (Working Paper 95-10), pt. 2, p. 6.

Baccalaureate and Beyond Longitudinal Study (B&B:93/94) Field Test Reinterview Study

The Baccalaureate and Beyond Longitudinal Study (B&B) was designed to provide data to answer questions about the aspirations, career plans, and achievements of bachelor's degree recipients; access to and progress through graduate and professional programs; the effects of undergraduate and graduate experiences on career histories; and the returns to both individuals and to society of investments in postsecondary education. The B&B sample, which is a subsample of students selected for the 1993 National Postsecondary Student Aid Study (NPSAS), represents all postsecondary students in the United States (including the District of Columbia and Puerto Rico) who received a bachelor's degree in academic year 1992-93. The B&B:93/94 study was the first in a planned series of followup interviews of the same group of respondents (panel survey) to take place over a period of up to 12 years. The first B&B followup collected information one year after respondents had received a bachelor's degree. (Green, Speizer, Campbell and Knepper, 1994).

Methodology and Design

The B&B:93/94 field test included a reinterview study.[2] The reinterview calls, designed for 5 minutes each, focused on items suspected of being unreliable. Reliability was measured in terms of the proportion of cases that had data which disagreed between the two interviews. (The proportion of disagreements is the same as GDR, and GDR is twice the response variance.) Thus, the "level of disagreement" equaled the number of discrepancies divided by the number responding to the item, multiplied by 100. Levels of disagreement in excess of 20 percent were cause for concern. However, Green et al. (1994, p. 52) cautioned that the actual disagreement between similar variables in the two data files might have been larger because the items were embedded in larger "skip patterns."

The reinterview study goal was to complete 100 reinterviews. A sample of 200 respondents was randomly selected from among the cases completed in the telephone center. Green et al. selected such a large sample relative to the number of respondents they intended to interview to reduce the cost of the reinterview and to complete the reinterviews within a short period of time. Indeed, reinterviewing was conducted in a 5-day period (September 18-23, 1993), and the goal of 100 completed reinterviews was met (1994, p. 51). See table 37.

[2] The B&B:93/94 field test was not the only time a B&B reinterview study was done. For example, see Green, Meyers, Giese, Law, Speizer, Tardino and Knepper, (1996), *Baccalaureate and Beyond Longitudinal Study: 1993/94 First Follow-up Methodology Report* (NCES 96-149) for a reinterview study conducted on the full scale survey.

Table 37. -- B&B:93/94 field test reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
B&B 1993/94	200	103	52%*	13%	Response variance

* A reinterview sample was selected from which only a targeted number (100) needed to be completed.

SOURCE: Green et al., (1994), *Baccalaureate and Beyond Longitudinal Study First Followup Field Test Report, 1993* (NCES 94-371), p. 51.

Summary of Results

The reinterview included five series of items: 1) graduate education choices, 2) costs of graduate education, 3) employment history, 4) undergraduate loans, and 5) changes in marital status. These subject areas will be described below along with the reason for including specific items in and the results of the reinterview study.

Graduate education choices. Students were asked about their first and second choice fields for graduate study, and their first and second choice schools within those areas. NCES wanted to ascertain whether students had stable preferences in those areas. In addition, respondents who had not applied for graduate study were asked an open-ended question about their reasons for deciding not to apply so NCES could assess the reliability of field-coding for this item. The following three questions showed levels of disagreement above 20 percent.

- AQ202 How many schools have you applied to for graduate study in <FIELD>? (26.83 percent) Green et al. (1994) did not comment on this item.
- AQ204 And what school was your FIRST/SECOND choice to study <FIELD>? (First choice field of study 21.95; First choice field—second choice school 42.11 percent) Inconsistent coding seemed to be more of a problem than inconsistencies in the respondents' reports. Increased interviewer training was recommended. The choice of a next best school appeared to be difficult for respondents. It was recommended to delete this item.
- AQ191 "Why did you decide not to apply to graduate school?" (77.4 percent) Green et al. suggested that the high level of error might indicate respondents "lack stable, strong reasons for not attending graduate school" (1994, p. 55). However, interviewers also had difficulty categorizing respondents' answers. It was recommended to use an alternate version of the question in the full scale survey.

Costs of graduate education. Respondents enrolled in school were asked about tuition, aid, and the total costs of schooling. The three questions asking specifically about costs all showed levels of disagreement above 20 percent.

- AQ212 Now I'd like you to think about the period from July 1, 1992 to June 30, 1993. During that period, what was your tuition at <GRADSCHOOL>? (58.33 percent)
- AQ213 What were the total costs of attending <GRADSCHOOL> during this same period (July 1, 1992-June 30, 1993). Total costs include tuition, books, transportation, living expenses, and other expenses related to attending school. (61.54 percent)
- AQ214 How much aid did you receive during this time, while enrolled at <GRADSCHOOL>? (53.84 percent)

The July 1 to June 30 time period used in the questions seemed to cause respondents some difficulty, perhaps because it did not correspond to the academic calendar. Green et al. recommended changing the phrase to the "past academic year and the summer preceding it." (1994, p. 56). However, even if the respondents had difficulty with the phrase, they did not usually report amounts in the reinterview that differed largely from their earlier reports. In approximately half the cases, the discrepancies were less than $500.

Respondents were also asked about the type of aid they received. Green et al. (1994, p. 57) state the information was reliably reported except for the category "monthly stipends, fellowships, grants, and scholarships." They decided to delete the word "monthly" from this category in the main data collection.

Employment history. The wording of the employment history items was changed for the reinterview to measure comparability with Recent College Graduates (RCG) survey items. Even though the alternative wording produced approximately the same responses as the original field test questions, almost all of the items were unacceptable by the 20 percent rule. Examination of the responses revealed that respondents in the reinterview sometimes did not mention a brief period of unemployment (or employment) following graduation that they had reported in the field test survey. Green et al. stated that "it was not clear if respondents' recall had diminished or if they had learned to shorten the interview by telling the interviewers less" (1994, p. 57). In light of these inconsistencies, Green et al. suggested that NCES reconsider collecting month-by-month employment histories, adding that if this series were retained, additional training hours would need to be added to familiarize interviewers with the items. Finally, if comparisons to the RCG data were desired, the results indicated that question wording should be the same in the two studies.

Undergraduate loans. One of the goals of B&B:93/94 was to understand how undergraduate debt affected graduates' choices concerning career and further schooling.

The field test included a series of items about education loans and debt repayment. Items in the reinterview were worded identically. Respondents could reliably confirm the amount of money they borrowed, but had some difficulty with the number of loans still outstanding. They had more difficulty supplying reliable information about the amount they still owed to lenders, but it is possible that the amount owed by the respondent actually changed between the two interviews. Two questions had levels of disagreement above 20 percent.

- AQ255 Of the $<Q253> you borrowed for your undergraduate education, how much do you still owe? (53.33 percent)
- AQ256 How many separate loans from undergraduate study are still outstanding, that is, for how many loans is or will a distinct payment be required? (36.73 percent)

Green et al. (1994, p. 62) suggested a number of options to improve the quality of this information. First, reliability might be improved by having the interviewers read each category to the respondent. It might also be useful to explore importing additional NPSAS data on the types of loans respondents have taken. Another option would be to conduct a limited number of brief cognitive interviews to identify why respondents were having difficulty in supplying information about the amount currently owed and the number of loans outstanding.

Changes in marital status. The reinterview tested a newly constructed format. When interviewers asked respondents to confirm their marital status, the interviewers could record all changes on a single CATI screen instead of cycling through a series of items. Interviewers could record responses in the sequence given by the respondent and generally have more flexibility in asking for the information. Green et al. report that the interviewers' response was positive and that the matrix was incorporated into the final instrument (1994, p. 63).

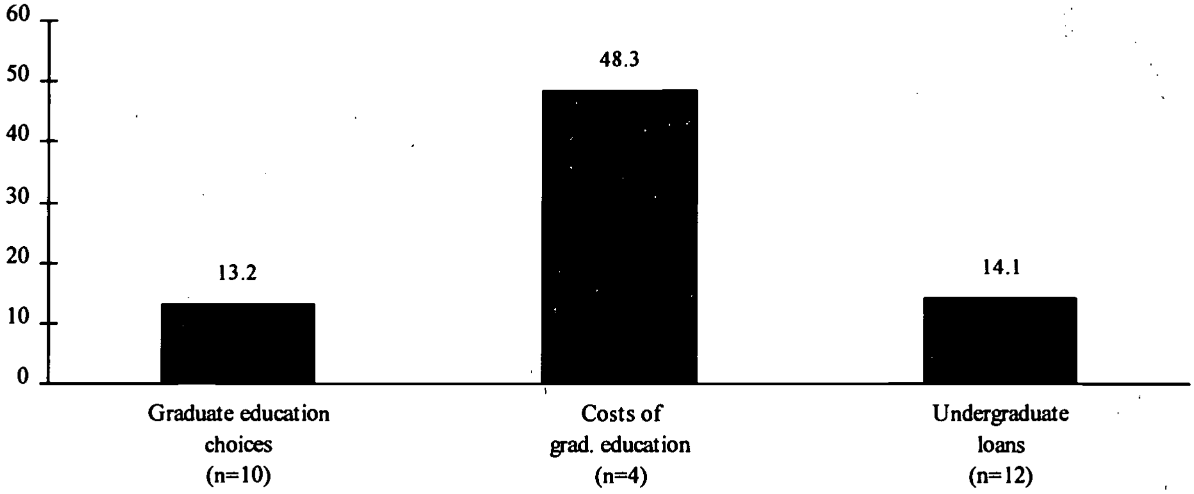
Areas of the survey chosen for the reinterview and the items with levels of disagreement above 20 percent are shown in table 38. Figure 22 shows mean reliability for the subject areas used in the B&B:93/94 reinterview study. The number of items (n) included in each subject area is indicated below the subject area title.

Table 38. -- B&B:93/94 field test reinterview subject areas and items with disagreement levels greater than 20 percent

Subject areas	Items with levels of disagreement above 20 percent
Graduate education choices	AQ196, AQ202, AQ204,
Costs of graduate education	AQ212, AQ213, AQ214
Employment history	--
Undergraduate loans	AQ255, AQ256
Changes in marital status	--

SOURCE: Derived from exhibits 6.3, 6.5, and 6.8, Green et al., (1994), *Baccalaureate and Beyond Longitudinal Study First Followup Field Test Report, 1993* (NCES 94-371), pp. 53-61, and appendix E.

Figure 22. -- GDR (mean values), B&B:93/94 (multiplied by 100)



NOTE: In B&B, GDR was referred to as level of disagreement: N of discrepancies/Percent out of N responding. GDR is equal to twice the response variance.

SOURCE: Derived from exhibits 6.3, 6.5, and 6.8, Green et al., (1994), *Baccalaureate and Beyond Longitudinal Study First Followup Field Test Report, 1993* (NCES 94-371), pp. 53, 56, and 61.

CHAPTER 4
Reinterview Studies: Reliability and Validity

In the psychometric literature "reliability" and "validity" are sometimes assessed using replicated measures, in a test-retest design. Some reinterview studies conducted during NCES full scale studies and field tests calculated reliability measures (from the psychometric perspective) as part of an examination of measurement error. This is what NCES did in the Beginning Postsecondary Students Longitudinal Survey (BPS), National Postsecondary Student Aid Study (NPSAS), and the National Study of Postsecondary Faculty (NSOPF) reinterview studies.

The same measurement error model specified by (3.1) and (3.2) in chapter 3 is applicable when considering how to evaluate reliability and validity. However, there are many cases when the existence of a "true value" is not tenable. Psychometricians distinguish between two types of measurements, or "scores": Platonic and non-Platonic (or classical). A Platonic true score is one for which the concept of a true value is plausible; physical measurements, personal demographic characteristics, and behavioral characteristics are examples. Classical true scores are those such as psychological states, attitudes, or knowledge, for which a true value can not be well defined.

To handle classical measurements, psychometricians assume a different origin for measurement error than for Platonic true scores. They assume that there exists a response distribution for the measurement y_{ti} which is associated with an individual i. Let μ_i denote the mean of the response distribution and let

$$\varepsilon_{ti} = y_{ti} - \mu_i ;$$

that is, ε_{ti} is a "sampling deviation for the t-th response (or measure) obtained from the i-th individual's response (or *propensity*) distribution. Since μ_i is the mean of the response distribution, we have the $E(\varepsilon_{ti}|i) = 0$. If the remaining assumptions of the simple response variance model specified by (3.1) and (3.2) hold, then the psychometricians' model for y_{ti} is equivalent, with the only difference being the interpretation of μ_i. This model has been referred to in the literature as the "classical true score" model (see for example Lord and Novick, 1968).

The psychometricians' objective is to provide "good" measures of a true score or *construct*. The two major criteria for ascertaining the goodness of a measure are *validity* and *reliability*. The concept of validity is complex and numerous types of validity have been proposed. (See chapter 1 for a review of some of these.) Our concern here is with *theoretical validity*. Under the model described above, theoretical validity or "reliability" can be assessed by the correlation of the two observed scores.

$$\mathrm{r}_{y_{i1},y_{i2}} = \frac{\mathrm{Cov}(y_{i1}, y_{i2})}{\sqrt{\mathrm{Var}(y_{i1})\mathrm{Var}(y_{i2})}}$$

Note that the index of inconsistency described in chapter 3 is exactly related to the index of reliability (Biemer and Stokes, 1991, p. 492). Indeed

$$r_{y_{i1},y_{i2}} = 1 - \text{IOI}$$

Thus, the two disciplines (survey statistics and psychometrics) have very similar forms of estimation for variable measurement error.[3]

The specific measures of association used to compute reliability in the NCES reinterview studies described in this chapter were related to the type of data being examined: coefficients of correlation are used with ratio-type data, coefficients of rank correlation with, obviously, ranked data, and coefficients of association or agreement with categorical data (see table 39). These measures are described in more detail below.

Table 39. -- Measures of association used to compute reliability in the BPS, NPSAS, and NSOPF* field test and full scale reinterview studies included in this report

	BPS: 90/92 field	BPS: 90/92 full	BPS: 90/94 field	NPSAS: 92-93 field	NPSAS: 96 field	NSOPF 1993 field
Coefficient of correlation						
Pearson's coefficient of correlation (r)	✓	✓	✓	✓	✓	✓
Coefficient of rank correlation						
Kendall's Tau (τ)			✓		✓	
Spearman's Rho (ρ)	✓	✓			✓	
Coefficient of association/agreement						
Phi coefficient (ϕ)	✓					
Cramer's V statistic (V)			✓		✓	
Cohen's Kappa (κ)		✓				

*NSOPF also used IOI.

SOURCE: Abraham et al., (1994), *1992-93 National Study of Postsecondary Faculty Field Test Report* (NCES 93-390); Abt Associates, Inc., (1993), *The National Postsecondary Student Aid Study, Field Test Report*; Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report: BPS 90/92* (NCES 92-160); Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study First Followup (BPS:90/92) Final Public Technical Report* (NCES 94-369); Green et al., (1994), *Baccalaureate and Beyond Longitudinal Study First Followup Field Test Report, 1993* (NCES 94-371); Loft et al.(1995) *Methodology Report for the National Postsecondary Student Aid Study, 1992-93* (NCES 95-211), Pratt et al., (1994), *Beginning Postsecondary Students Longitudinal Study Second Followup Field Test Report: BPS 90/94* (NCES 94-370); Research Triangle Institute, (1996), *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report*

Coefficient of correlation. A coefficient of correlation is used to analyze items with continuous response categories. The Pearson coefficient of correlation, also referred to as the product moment coefficient of correlation, has as its numerator the first product moment or

[3] This example of the index of inconsistency and the reliability coefficient is typical of other comparisons of terms between survey statistics and psychometric approaches to measurement error. The psychometric approaches typically measure the positive side—validity and reliability; the survey statistics models measure the negative side—measurement bias and response variability.

covariance of the two variates (Kendall and Buckland, 1971, pp. 112 and 119). The coefficient of correlation is defined as

$$r = \frac{\text{Covariance}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

Coefficients of rank correlation. A coefficient of rank correlation is used to analyze items with discrete, ordered response categories. Rank correlation measures the degree of agreement between two sets of rankings or the degree of correspondence between them (Kendall and Buckland, 1971, p. 125). The two principal coefficients of rank correlation are Spearman's Rho (ρ) and Kendall's Tau (τ).

Kendall and Buckland's definition of Spearman's rank order coefficient, ρ, (1971, p. 141) states that if the two rankings are a_i, b_i, and d_i is defined as $a_i - b_i$, where $i = 1, 2, ..., n$, the coefficient is given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

They also explain that Spearman's Rho is the product moment correlation between the rank numbers *a* and *b*.

Kendall's Tau is a coefficient of rank correlation "based on the number of inversions in one ranking as compared with another" (Kendall and Buckland, 1971, p. 78). Pratt et al. (1994, p. 66) state that Kendall's Tau is a measure of consistency of ranks or other data with only ordinal properties (in simplest form reflecting the difference between the proportions of consistent responses and those of reversed response). A Tau value near 0 indicates that consistent and reversed positions are equally likely, reflecting little predictability across the two interviews; a Tau of 1 represents perfect consistency. Spearman's and Kendall's estimators are asymptomatically equivalent.

Coefficients of association/agreement. Coefficients of association and agreement can be used with both nominal and ordinal categorical data. Kendall and Buckland (1971) define association in its most general sense as "the degree of dependence, or independence, which exists between two or more variates whether they be measured quantitatively or qualitatively" (p. 6). However, the term is used in a more narrow sense "to denote the relationship between variates which are simply dichotomized, namely in a 2 x 2 table....If, in a two-fold table, the frequencies of the attributes (*A*, *B*) (not *A*, *B*), (*A*, not *B*) and (not *A*, not *B*) are respectively *a, b, c, d,* the association between *A* and *B* is said to be positive if

$$a > \frac{(a+b)(a+c)}{a+b+c+d}$$

within sampling limits, and negative in the contrary case; if the inequality becomes an equality the attributes are independent" (p. 6).

The two coefficients of association named in the field test reinterview studies we discuss in this report are the Phi coefficient (ϕ) (Kendall and Buckland, 1971, p. 32) and Cramer's V statistic (Kendall and Stuart, 1979, p. 588).

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n(q-1)}}$$

where $q = \min(r, c)$, when r equals number of rows and c equals number of columns.

Cramer's statistic is a simple function of a chi-square statistic, normalized to vary between 0 and 1. For items with only two categorical outcomes, Cramer's V statistic is equal to the Phi coefficient for 2 x 2 tables.

Cohen's Kappa, a measure of agreement, is defined in the following way by Agresti (1990, p. 366).

Let π_{ij} denote the probability of classification in the *i*th category at the first interview and the *j*th category at the reinterview. Then

$$\Pi_o = \sum \pi_{ii}$$

is the probability the two interviews agree. Perfect agreement corresponds to $\Pi_o = 1$. If the two interviews are statistically independent, $\pi_{ii} = \pi_{i+}\ \pi_{+i}$, then the probability of agreement equals

$$\Pi_e = \sum \pi_{i+}\pi_{+i}$$

Thus, $\Pi_o - \Pi_e$ is the excess of the interview agreement over that expected purely by chance (i.e., if interviews were statistically independent).

Cohen's Kappa is

$$\mathrm{K} = \frac{\Sigma\pi_{ii} - \Sigma(\pi_{i+})(\pi_{+i})}{1 - \Sigma(\pi_{i+})(\pi_{+i})} = \frac{\Pi_o - \Pi_e}{1 - \Pi_e}$$

The denominator replaces Π_o by its maximum possible value of 1, corresponding to perfect agreement. Kappa equals 0 when the agreement equals that expected by chance, and it equals 1.0 when there is perfect agreement. The stronger the agreement, the higher the value for a given pair of marginal distributions.

Categories of Reliability Measurements

In chapter 3, when looking at IOI, high measurements indicate there may be problems.[4] High correlations, on the other hand, indicate a high degree of reliability, which is supportive of the validity of the instrument.

The range of reliability measurements are rated as follows

- Less than 0.5, the reliability is low
- Between 0.5 and 0.8, the reliability is moderate
- Greater than 0.8, the reliability is high

The remainder of this chapter will describe the methodology, design, and results of the following reinterview studies.

- **Beginning Postsecondary Students Longitudinal Survey (BPS)**
  First Followup Field Test, BPS:90/92
  First Followup, BPS:90/92
  Second Followup Field Test, BPS:90/94

- **National Postsecondary Student Aid Study (NPSAS)**
  NPSAS:92-93 Field Test
  NPSAS:96 Field Test

- **National Study of Postsecondary Faculty (NSOPF) 1993**

For each reinterview study, tables and figures display summary statistics and results. Where the information is available, tables show the reinterview sample size, the number of completed reinterviews, and the response rate, as well as how the reinterview sample size compares with the number of completed interviews in the original survey (sampling percentage).

Whereas the tables presented in chapter 3 listed the mean score for IOI in each subject area and drew attention to items with high IOI as being problematic, the tables in this chapter will show median values and list the items with reliability measures in the low range.

Beginning Postsecondary Students (BPS) Longitudinal Study Reinterview Studies

The Beginning Postsecondary Students (BPS) Longitudinal Study was begun to complement the high school cohort longitudinal studies and to improve data on participants in postsecondary education. BPS includes not only the "traditional" students, recent high school

[4] The ranges for IOI are less than 20 (the impact of measurement error is low), between 20 and 50 (the impact of measurement error is moderate), and greater than 50 (the impact of measurement error is high).

graduates, but the "nontraditional" older students, making BPS representative of all beginning students in postsecondary education. By starting with a cohort which has already entered postsecondary education, BPS is able to address issues of persistence, progress, and attainment, as well as issues related to transitions between undergraduate and graduate education and transitions between postsecondary education and work.

The BPS sample is based on the National Postsecondary Student Aid Study (NPSAS). BPS followed NPSAS:90 beginning students starting in 1992. About 8,000 students who began their postsecondary education career in the 1989-90 academic year responded to NPSAS:90 and were included in the first BPS (BPS:90/92) in the spring of 1992 and the second BPS (BPS:90/94) in the spring of 1994.

BPS includes reinterview studies as part of its methodology for field tests and full scale surveys.[5] The reinterview studies discussed in this report assessed the reliability of responses over short periods of time (i.e., a matter of weeks). BPS also conducted validation reinterviews, but since they included followup questions using paraphrasing techniques, the results are discussed in chapter 7.

BPS:90/92 First Followup Field Test Reliability Reinterview

Methodology and Design

The reinterview of CATI operations was administered randomly to a subset of BPS respondents to assess the short-term reliability of selected items. The items chosen were generally important to the study and the responses were not expected to change much between interviews. This was useful for assessing whether interview responses contained sizable measurement errors that were unstable over relatively short time frames. These errors could be a result of inattention, inaccuracy of recall, or difficulties in understanding survey questions. The analysis used three measures of association: Pearson's coefficient of correlation (r) for data such as number of terms, beginning or ending months or years, or dollars; Spearman's Rho (ρ) for data such as rankings; and the Phi coefficient (ϕ) for dichotomous (yes/no) data for selected items. In addition, proportions of responses that were exactly the same in the main interview and the reinterview were calculated separately for each item (Burkheimer, Forsyth, Wheeless, Mowbray, Boehnlein, Knight, Veith, and Knepper, 1992, pp. VI-5 and VI-6).

The reinterview study lasted slightly more than 7 weeks. The sample size was 125, that is, 11 percent of the field test sample; the response rate was 92 percent (see table 40).

[5] Documentation for the BPS:90/94 full scale reinterview came out too late to be included in this report. See Pratt, Whitmore, Wine, Blackwell, Forsyth, Smith, Becker, Veith and Bobbitt (1996), *Beginning Postsecondary Students Longitudinal Study Second Follow-up (BPS:90/94) Final Technical Report* (NCES 96-153).

Table 40. -- BPS:90/92 first followup field test reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
BPS:90/92 field test	125	115	92%*	11%	Reliability

* A reinterview sample was selected from which only a targeted number needed to be completed.

SOURCE: Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report: BPS 90/92* (NCES 92-160), pp. VI-4 and VI-5.

Summary of Results

The analysis looked at items in six subject areas: terms at the NPSAS school, terms at other schools, information about terms since February 1989, satisfaction with school programs, financial aid, and work experience. Table 41 summarizes the results, as well as listing the item numbers which had measures of association in the low range. Each of the subject areas is discussed below.

Table 41. -- BPS:90/92 first followup field test reliability reinterview study median subject area measurements

Subject areas	Median percentage exact agreement	Measures of association*	
		Median	Items in the low range
Terms at the NPSAS school	90.0	High r	--
Terms at other schools	79.2	High r	B.2 for "last term ending month" (r)
Information about terms since February 1989	80.0	High r / Moderate ϕ	B.7.i for "last term" (ϕ)
Satisfaction with school programs	63.8	Moderate ρ	B.9(j) (ρ)
Financial aid	69.4	High r / Moderate ϕ	--
Work experience	64.5	Moderate r / Low ϕ	D.1 (ϕ)

*Pearson's coefficient of correlation (r), Spearman's Rho (ρ), and the Phi coefficient (ϕ).

SOURCE: Derived from tables VI.E.1-6, Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report: BPS 90/92* (NCES 92-160), pp. VI-7 to VI-11.

Terms at the NPSAS school. Students were asked to correct or update information already on file about terms at the NPSAS school. Only 76 percent of the students gave data for the same number of terms in the two interviews although the correlation between the numbers of terms given was quite high (r = 0.90). The agreement between the beginning and ending dates for the first terms at the NPSAS school was also quite high (r = 0.79 - 1.0). Agreement between the beginning and ending dates for the last reported term was also high (r = 0.86 - 0.89) (Burkheimer et al., 1992, VI-6 and VI-7).

Terms at other schools. Students were asked to update information about terms at schools other than the NPSAS school. Agreement was higher for the first term (r = 0.73 - 0.90) than for the last reported term (r = 0.47 - 0.81), and was higher for year (r = 0.76 - 0.90) than for months (r = 0.47 - 0.82). The following item had the low 0.47 score.

- B.2 Now I want to ask you about any other schools and the terms in which you may have gone to these schools. [school name/start month/start year/end month/end year] (r = 0.47 for last term ending month)

Information about terms since February 1989. Students were asked about features of the terms of enrollment since February, 1989. Only 68 percent of the students responded for exactly the same number of terms in the two interviews, although the correlation between the number of terms reported was high (r = 0.83). Other reinterview items asked about the number of courses, the number of credits for which the students was enrolled, whether or not the students were working toward a license or certificate, and if so which one, and whether or not those students completed work toward the certificate or license. Burkheimer et al.(1992) found the reliability of these data was generally acceptable, particularly for the number of courses enrolled in during a term, credits enrolled for, and completion of work on a certificate or license (r = 0.77 - 0.89). Other information about licenses and certificates was not systematically reliable in terms of the Phi coefficient (0.49 - 0.65); however, exact agreements were all 80 percent or greater (pp. VI-8 and VI-9). The following item had a low coefficient.

- B.7.i Were you working toward a certificate or a license? (ϕ = 0.49 for "last term")

Satisfaction with school programs. Students were asked to rate their satisfaction with various services, programs, and features at the NPSAS school. About 65 percent of the students gave identical ratings in the two interviews. The Spearman's Rho correlations for the three items in this subject area were 0.72 (job placement), 0.62 (financial aid counseling), and 0.45 (career or job counseling). Burkheimer et al. (1992) commented that the "general low reliability of these 'nonfactual' items is well known and the results are not particularly surprising" (p. VI-9). The full wording for the item with a low correlation is

B.9 While you were enrolled in (name of first school/college), how satisfied were you with the following? [very dissatisfied/somewhat dissatisfied/somewhat satisfied/very satisfied/NA]

- (j) Career or job counseling (ρ = 0.45)

Financial aid. Students were asked about education expenses. While exact agreement was not uniformly high, the correlation in the amounts of aid received and the amount currently owed was quite high (r and ϕ coefficients greater than 0.8). As Burkheimer et al. (1992)

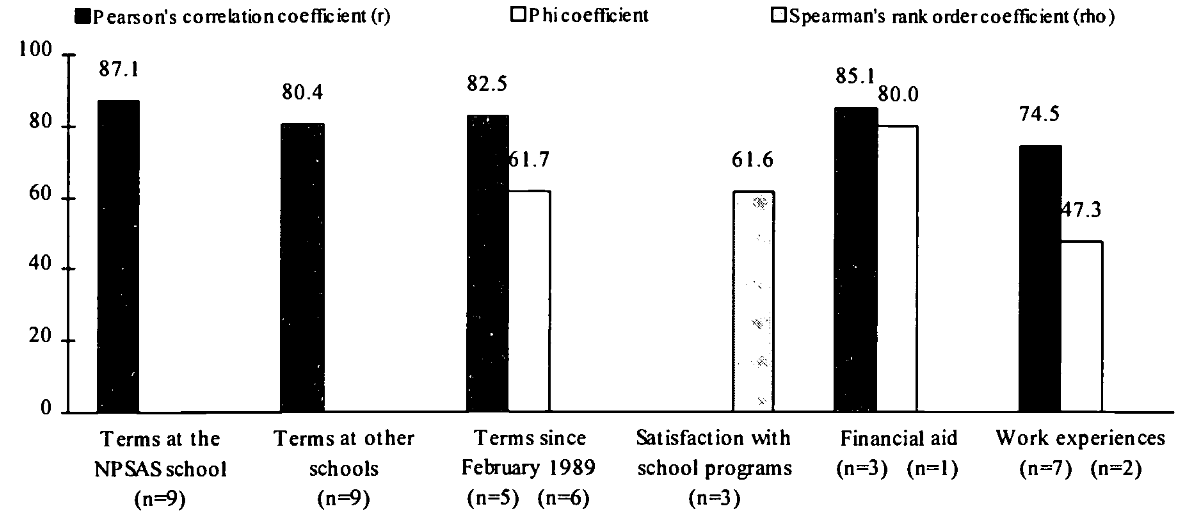
expected, the data for the more recent 1990-91 school year had a higher correlation (r = 0.95) than for the 1989-90 school year (r = 0.83) (pp. VI-9 and VI-10).

Work experience. Students were asked about all jobs held since February 1989. Reinterview items also collected salary information, but this information was not compared with original survey data because students could legitimately change units (hours worked per day, days worked per week). Burkheimer et al. (1992) found the job information was somewhat less reliable than they would have liked. For example, there was only 60 percent agreement as to the number of jobs held since February 1989 (r = 0.75). The reliability for items about the first job held since February 1989 ranged from 0.53 to 0.86 (r and ϕ coefficients). To improve the reliability of these items for the main study, Burkheimer et al. recommended implementing a summary/verification screen for listing jobs similar to that used for the enrollment data (pp. VI-9 to VI-11). The one item with a low coefficient follows.

- D.1 Have you held a job for pay at any time (including co-ops, work study, summer jobs, and part-time jobs such as in the National Guard or military reserve), either full-time or part-time, since February 1989? [yes/no] (ϕ = 0.28)

Figure 23 illustrates the median values of the measures of association by subject areas. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 23. -- Measures of association (median values), BPS:90/92 field test (multiplied by 100)



SOURCE: Derived from tables VI.E.1-6, Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report: BPS 90/92* (NCES 92-160), pp. VI-7 to VI-11.

BPS:90/92 First Followup Full Scale Reliability Reinterview

Methodology and Design

The methodology and design of the reliability reinterviews conducted for the BPS:90/92 full scale study were similar to those employed for the field test reinterviews. The reliability reinterviews again assessed the short-term (typically two to six weeks) reliability of selected items, and the analyses focused on data items that were important to the study and not expected to demonstrate much real change between interviews. The BPS:90/92 full scale reinterview study also computed measures of association for each selected item, and, as in the field test reinterview study, three measures of temporal stability were used: Pearson's coefficient of correlation (r) and Spearman's rank order coefficient, or Spearman's Rho (ρ), as before, but the full scale study used Cohen's Kappa statistic (κ) instead of the Phi coefficient.

Proportions of agreeing responses across the main interview and the reinterview were calculated separately for each item. They were calculated as follows: (1) for nominal and ordinal variables, the proportions were computed based on the number of responses that were exactly the same across the main interview and reinterview, and (2) for continuous variables, based on the number of responses that were within one standard deviation unit of each other across the main interview and the reinterview.

Table 42. -- BPS:90/92 first followup full scale reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
BPS:90/92 full scale	--*	191	--	--*	Reliability

* The sample size is not documented; however, only a target number needed to be completed. The technical report states that 9,011 initial sample members were fully (8,495) or partially (516) interviewed, thus becoming eligible for sampling for the reinterview study.

SOURCE: Burkheimer et al., (1994), *Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:90/92) Final Public Technical Report* (NCES 94-369), pp. 53, 85.

Summary of Results

The analysis looked at items in the same six subject areas as in the field test study: terms at NPSAS school, information about other schools, information about terms since February 1989, education services at the NPSAS school, factors related to education financing, and work experience. Table 43 summarizes the results, as well as listing the item numbers which had measures of association in the low range. Each of the subject areas is discussed below.

Table 43. -- BPS:90/92 full scale reinterview median subject area measurements

	Median proportion	Measures of association*	
Subject areas	agreement	Median	Items in the low range
Terms at NPSAS school	0.98	High r	--
Information about other schools	0.99	High r	--
Information about terms since February 1989	0.86	Low r / Moderate κ	Number of courses for first and last terms; other last term information—complete academic degree
Education services at the NPSAS school	0.64	Low κ / Moderate ρ	Use/satisfaction with remedial instruction, academic counseling, and career counseling; service provision for remedial format, career format, remedial provider, and career provider
Factors related to education financing	0.95	High κ and r	First and last NPSAS term—employer benefits, relative loan
Work experience	0.82	Moderate r and κ	--

*Pearson's coefficient of correlation (r), Spearman's Rho (ρ), and Cohen's Kappa statistic (κ).

SOURCE: Derived from tables IV-5 through IV-11, Burkheimer et al., (1992), *Beginning Postsecondary Students Longitudinal Study First Followup (BPS:90/92) Final Public Technical Report* (NCES 94-369), pp. 85-92.

Enrollment at NPSAS schools. Respondents were asked to correct and/or update base-year data about terms at the NPSAS school and to provide information about additional terms at that school since the base-year study. The results were generally comparable to those for the field test. The Pearson's correlations for items on the number of school terms at NPSAS schools and the starting and ending dates for first and last term at a NPSAS school were between 0.82 and 0.99.

Enrollment at other schools. Respondents were asked to update/correct any prior information about terms at schools other than the NPSAS school. The stability statistics were high (r = 0.88 - 0.98), generally much higher than results obtained in the field test (r = 0.47 - 0.90). Burkheimer et al. attributed this to changes made in the CATI program to obtain these data in a more straightforward manner (1994, p. 86).

Information about terms since February 1990. In this section, information was collected about first and most recent terms of enrollment at the NPSAS school since February, 1990. The test-retest correlations were low for the first (r = 0.35) and most recent (r = 0.44) NPSAS school terms although the percent agreements were high (94 and 95 percent, respectively). Burkheimer et al. (1994, p. 87) strongly suspected that the low correlations were caused by modifying the way the term of reference was identified in the production and reliability interviews. The measures of association for these items in the field test, when the data were collected exactly the same way in both production and reliability interviews, were moderate to high (r = 0.77 - 0.89, ϕ = 0.50 - 0.89).

Education services at the NPSAS school. In the field test, the items in this subject area asked about satisfaction with services without asking how respondents used the services. For the full scale study, interviewers asked not only about user satisfaction, but how often respondents used the services and the type of service provided. Question 9a reads

- I am now going to ask you about your satisfaction with certain school features and services at (name of NPSAS school). For the services I mention, please first indicate whether or not you used the service, and then indicate your satisfaction. (1) Very Dissatisfied, (2) Somewhat Dissatisfied, (3) Somewhat Satisfied, or (4) Very Satisfied [(5) Didn't use (where applicable)]. (The services included in the question were special tutoring or remedial instruction, academic counseling, and career or job counseling.)

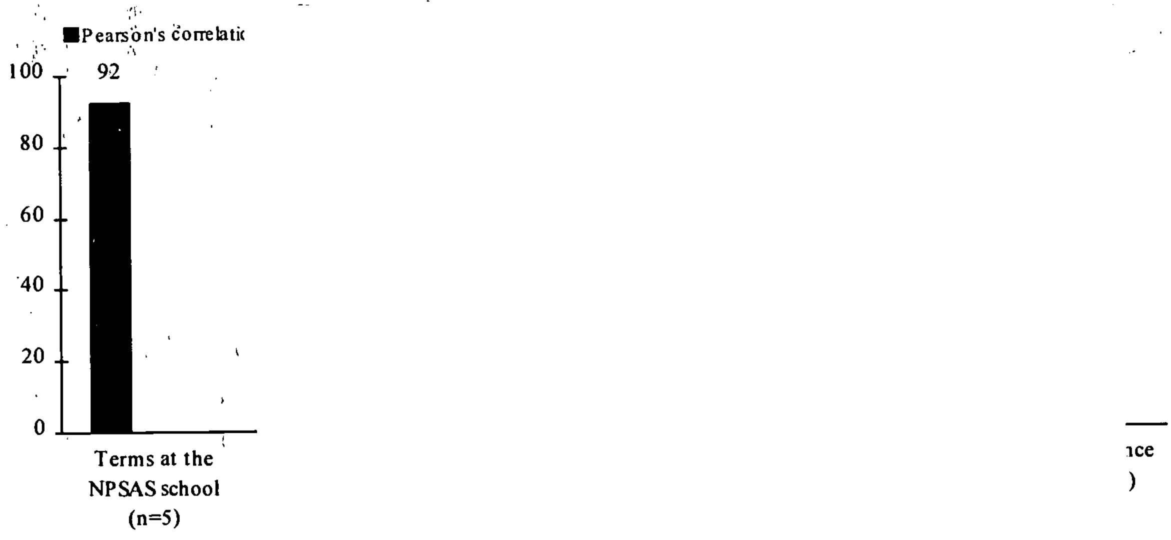
The measures of association were low to moderate ($\kappa = 0.10 - 0.42$, $\rho = 0.57 - 0.63$). Burkheimer et al. (1994) believe that at least part of the unreliability of these items was because respondents were asked about use and satisfaction levels in the same item. The field test results for items asking only about satisfaction were ($\rho = 0.45 - 0.72$. Burkheimer et al. believe this indicates that a better presentation of these items in subsequent followup studies could improve the reliability of the data collected (1994, p. 88).

Factors related to education financing. The reliability indices for items in this subject area were generally quite high (median $\kappa = 0.87$; the sole $r = 0.96$). Burkheimer et al. felt that the emphasis placed on obtaining good educational financing data during interviewer training and during subsequent monitoring and supervision contributed to this (1994, pp. 90-92). The emphasis was not as great during the field test; the indices obtained were $r = 0.83 - 0.95$, $\phi = 0.80$.

Work experience. Burkheimer et al. felt the agreement proportions (56 - 92 percent, median 82 percent) were comparable to field test results (45.9 - 89.1 percent, median 64.5 percent) and were generally acceptable.

Figure 24 illustrates the median values of the measures of association by subject areas.

Figure 24. -- Measures of association (median values), BPS:90/92 full scale (multiplied by 100)



SOURCE: Derived from tables [illegible] Burkheimer et al., (1994), *Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:90/92) Final Public Technical Report* (NCES 94-369), pp. 85, 87, 89, 91-92.

BPS:90/94 Second Followup Field Test Reliability Reinterview

As in previous BPS data collections, the BPS:90/94 field test study included a reinterview study to evaluate short-term reliability of BPS interview responses. Each new reliability reinterview is designed to build on previous analyses by targeting revised or new items, and items not previously evaluated. BPS reinterview analyses also generally focused on data items expected to be stable for the relatively short time period between the initial interview and the reinterview.

Methodology and Design

The reinterview goal was to complete 100 interviews, and a sample of 113 respondents was selected. Reinterview respondents were contacted four to eight weeks after completing the initial interview: analyses were based on the 95 respondents who completed reinterviews (or applicable subsets thereof). Four major areas were examined: 1) education experiences, including primary school information and grades, tests, and expectations; 2) work experiences, including principal job information, job search activities, satisfaction with most recent primary job, and factors in employment goals; 3) marital status; and 4) finances (educational and personal).

Pratt, Burkheimer, Forsyth, Wine, Veith, and Knepper used three relational statistics in their analyses: Cramer's V statistic (V), Kendall's Tau (τ), and Pearson's coefficient of correlation (r). Proportions of agreeing responses across the main interview and the reinterview were

calculated the same way they were for the BPS:90/92 full scale reinterview study: (1) for nominal and ordinal variables, the proportions were computed based on the number of responses that were exactly the same across the main interview and reinterview, and (2) for continuous variables, based on the number of responses that were within one standard deviation unit of each other across the main interview and the reinterview.
(1994, p. 66).

Table 44. -- BPS:90/94 second followup field test reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
BPS:90/94 field test	113	95	84%	11%	Reliability

SOURCE: Pratt et al., (1994), *Beginning Postsecondary Students Longitudinal Study Second Followup Field Test Report: BPS 90/94* (NCES 94-370), p. 66.

Summary of Results

This analysis looked at items in 10 subject areas: primary (in the sense of "main") school information; grades, tests, and expectations; general job information; principal job information; job search activities; satisfaction with most recent primary job; factors in employment goals; marital history; education finances; and personal finances. Table 45 summarizes the results; each of the areas is then discussed below.

Table 45. -- BPS:90/94 second followup field test reliability reinterview study median subject area measurements

Subject areas	Median percent agreement	Measures of association*	
		Median	Items in the low range
Education experiences:			
Primary school information	68.2	Moderate τ and V	B33D c, e (τ)
Grades, tests, and expectations	71.8	High τ / Moderate V	--
Work experiences:			
General job information	89.5	Moderate V / High r	C02A, C90A (V)
Principal job information	80.0	Moderate V	C51C, C54A (V)
Job search activities	99.0	Low V	C53B (V)
Satisfaction with most recent primary job	76.4	Moderate τ	--
Factors in employment goals	72.1	Moderate τ	C92A a, c, e; C93A d; C94A a, b (τ)
Marital history	97.9	High τ and r	--
Education finances	90.6	High r / Moderate V	FxxC (r)
Personal finances	92.6	Moderate V and r	G0aC, G0aE, G0aG (combined) (V)

* Cramer's V statistic (V), Kendall's Tau (τ), and Pearson's coefficient of correlation (r).

SOURCE: Derived from tables V.2 - V.11, Pratt et al. (1994), *Beginning Postsecondary Students Longitudinal Study Second Followup Field Test Report; BPS:90/94* (NCES 94-370), pp. 67-76.

Education experiences: primary school information. Respondents were asked about the schools they attended and about their satisfaction with specified features of school climate at their principal school. The percent agreement and relational statistics were higher for primary school identification (97.0 to 97.8 percent; V = 0.64 - 0.72) than for satisfaction with school climate features (52.3 to 69.8 percent; τ = 0.41 - 0. 64). Pratt et al. suggested two general revisions for the satisfaction items. Question wordings should be revised to emphasize and anchor the question time frame. Additionally, the number of response categories could be reduced from four ratings of satisfaction to two (1994, p. 67).

Two items had measures of association in the low range

B3bD As an undergraduate at [principal undergrad school] since 1 July 1991 how satisfied were you with... [very dissatisfied/somewhat dissatisfied/somewhat satisfied/very satisfied]

- c. Your intellectual growth (τ = 0.41)
- e. The prestige of the school (τ = 0.43)

Education experiences: grades, tests, and expectations. Respondents were asked about undergraduate grades ($\tau = 0.80$), their expected level of overall education completion ($\tau = 0.87$), and whether any graduate admissions tests were taken ($V = 0.80$). The results indicated generally high levels of short-term reliability for responses to these questions. This was particularly noteworthy for the item on education expectations, which was, "Considering all practical constraints, what is the highest level of education you ever expect to complete?" Pratt et al. stated that "tying the item to practical expectations seemed to provide more response stability than was possible in 'aspiration' versions of this question" (1994, p. 67).

Work experiences: general job information. The percentage agreements for the items on reports of any job since February 1991, number of jobs since February 1991, and first and most recent job start dates were high (89.5 - 95.9 percent) and the reliability indices were moderate to high ($V = 0.51$-0.78; $r = 0.87 - 0.95$). However, Pratt et al. were somewhat concerned with the response reliability of a 5-category item asking about the primary role of the sample member while attending school. It is listed below.

C02A Since you were employed before your last reported term in school, how did you view your primary role in relation to education and work?

- a student who works to help pay expenses while in school/college
- a student who works to earn extra spending money while in school/college
- an employee who attends school/college to gain skills necessary for job advancement
- an employee who attends school to expand new career possibilities
- an employee who attends school to expand personal knowledge/skills not related to employment

Pratt et al. tried to examine whether the low percent agreement and reliability index could be attributed to the relatively fine distinctions within the "primarily student" and "primarily employee" response categories by computing a second set of consistency measures for a variable collapsing responses into a dichotomous student or employee variable. The percent agreement improved (from 61 - 88 percent); however, the value of Cramer's V statistic increased only slightly (0.38 - 0.41). They recommended that response options distinguish respondents who were "primarily students" from "primarily employees" and that the relevant time period be more clearly defined in the item (1994, pp. 68-69).

Item C90A, about working plans, also had a relatively low Cramer's V statistic (0.49).

- C90A Five years from now (spring of 1998), do you intend to be working for pay, either full-time or part-time? If you are not sure of the answer, please give your best estimate. [yes—full-time/yes—part-time/no/don't know/refuse]

Pratt et al. felt this reflected a basic unreliability among dual purpose items (e.g., asking in one item both whether one plans to be working in five years and, if so, whether full- or part-time). Research has suggested these items are difficult for respondents to interpret and answer correctly (1994, p. 69).

Work experiences: principal job information. The data element for type of company yielded a high percentage of agreeing responses (91.0), as did the data element for education need to get job (94.0), but the results were not as good for the remaining items analyzed for this section, none of which had been previously examined. The two items listed below caused considerable concern.

- Cy1C Please tell me which of the following statements (which I will read to you) apply to your job with (principal job employer)
  I did not use tools/equipment I was trained to use (V = 0.11—not significantly different than zero)

  Other response categories: I was able to apply most of what I learned in school, the job was different from my education/ training, I could have gotten this job without my training/ education

- Cy4A Was additional education or training required for advancement in your job? [yes/no/don't know/refuse] (V = 0.12—not significantly different than zero)

Pratt et al. decided that the low reliability indices might have been due to confusing or vague question wording. They felt that the "tools" question should be reworked for the full scale study, and suggested that the revised question be re-evaluated. For the second question, on additional education, Pratt et al. suggested using wording that more clearly defined the types of training respondents should consider and the types of advancement relevant to answering the items (1994, pp. 69-70).

Work experiences: job search activities. Information on job search activities was collected through an "open ended" question.

- Cy3B What were the primary things you did to find this job?
  Among the responses, the "resume" category had a Cramer's V statistic of 0.38, "interviewed about opportunities" 0.04 (not significantly different than zero), and "read want ads" 0.49.

The interviewers categorized the responses and asked the respondents to verify the categorization. If respondents did not mention a specific activity, the interviewer marked "not pursued." As Pratt et al. stated, "different responses over time could reflect situational recall and interviewer interpretation as well as changes in question interpretation and response selection strategies on the part of the respondent" (1994, p. 71). Therefore, if these items were kept in the full scale instrument, they recommended interviewers be specially trained in evoking appropriate recall and in coding these responses. The interviewers need a better understanding of what should be included and what should be excluded from codes for each job search activity, and the response options need clearer and more explicit definitions.

Work experiences: satisfaction with most recent primary job. Items requesting reports of satisfaction with specified aspects of the respondent's most recent job had generally high consistency measures. No items had measures of association in the low range. Pratt et al. speculated that the higher consistency for job satisfaction items might have resulted, in part, from the greater recency of the most recent primary job (1994, p. 72). Therefore, they felt that the job satisfaction items in the full scale study questionnaire should clearly define and anchor the question time frames in order to improve reliability. The response options were another issue. These items used (1) satisfied, (2) neutral or no opinion, and (3) dissatisfied. The members of the Technical Review Panel recommended eliminating the "neutral" response option, since such an option is always available to the respondent by answering "I don't know."

Work experiences: factors in employment goals. Respondents were asked to rate the importance of several factors in determining lifelong work in three items. The Kendall's Tau values suggested only marginally acceptable temporal stability (0.23 - 0.66); the percent agreements ranged from 61.0 to 87.4. As these were attitudinal items, Pratt et al. believe the response inconsistency may have reflected real changes in attitudes across the two interviews. Some of the unreliability was also probably due to complexities in some question wording. For example, "good income or fringe benefits to start or within a few years" compounds income and benefits and time frame. One of the recommendations for the full scale study was to use separate items for income and benefits, "possibly concentrating on the general importance of these two aspects of work, rather than introducing the compound time frame" (1994, p. 73). Items about the importance of social factors were even less reliable. Based on this, the Technical Review Panel recommended eliminating the full set or correcting particularly vague wordings in this set of items. Listed below are those portions of the items with measures of association in the low range (less than 0.5).

C92A In determining the kind of work you plan to be doing for most of your life, how important are each of the following general factors? [not important/ somewhat important/very important]

- a. Previous work experience in the area ($\tau = 0.42$)
- c. Freedom to make your own decisions ($\tau = 0.49$)
- e. Work with increasing responsibilities over time ($\tau = 0.45$)

Other response categories: b. Work that seems important and interesting to you, d. Work where most problems are quite difficult and challenging

C93A In determining the kind of work you plan to be doing for most of your life, how important are each of the following factors? [not important/somewhat important/very important]

- d. Opportunity for further education and/or training ($\tau = 0.38$)

Other response codes: a. Good income or fringe benefits to start or within a few years, b. Job security and permanence, c. Opportunity or promotion

C94A In determining the kind of work you plan to be doing for most of your life, how important are each of the following lifestyle factors? [not important/ somewhat important/very important)

- a. Meeting and working with sociable people ($\tau = 0.23$)
- b. Work that has high status and prestige ($\tau = 0.45$)

Other response categories: c. Work that lets you establish roots & not have you move, d. Work that leaves lots of time for other things in life, e. Work that allows a great deal of travel

Marital history. Respondents were asked for detailed information on marital history. The reliability indices were generally high: percent agreements from 95.7 to 100.0; Cramer's V statistics of 0.76 to 1.00, and a Pearson's coefficient of 0.95. Pratt et al. believe that the high reliability is at least partly due to marital status not changing during the 4-year period covered by the questions. They suggest that these items be re-evaluated in subsequent followups since there are likely to be more changes in marital status as the BPS sample ages.

Education finances. Respondents were asked a series of items about financial aid at the school they had identified as their principal school for the 1991-92 academic year. The percent agreement measures were generally high (80 - 100 percent). The lowest agreement was 80 percent for the question on receipt of "other" aid, which also had a Cramer's V statistic that did not differ significantly from zero. Pratt et al. (1994) explain that "most of the 20 applicable respondents answered, in both interviews, that they had *not* received "other" aid; however, the small number of cases reporting to have received such aid typically did so inconsistently. The Cramer's statistics magnifies inconsistency within small groups" (p. 74).

FxxC While attending (principal school), as an undergraduate student, between 1 July 19xx and 30 June 19xx, did you receive

- other financial aid ($V = 0.11$)

Other response categories: grants/scholarships/student loans/prepaid tuition benefits/reimbursed tuition benefits

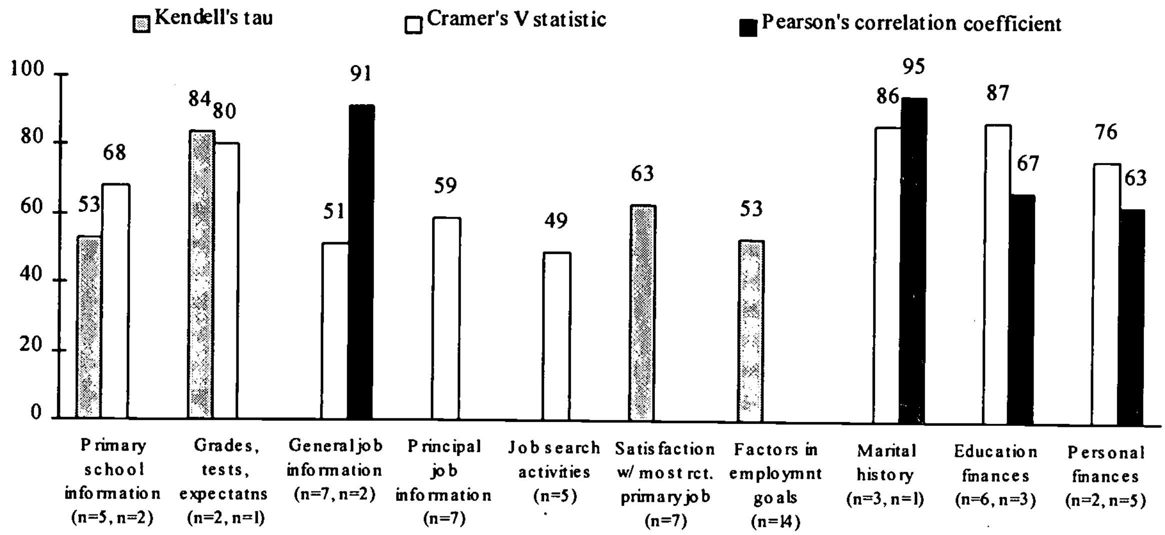
Personal finances. Indices were high for owning or renting a residence and, among those owning or renting, for amount of monthly payments (percent agreements of 92.6 and 94.5, V = 0.90, and r = 0.89). Indices for owning or leasing a vehicle were noticeably lower (percent agreement of 87.4, r = 0.62); however, among those consistently reporting having a vehicle, the monthly payment amount was quite reliable (percent agreement of 96.7, r = 0.96). Monthly expenditures for nonrecreational items was the least stable of the expenditure amounts (percent agreement of 83.0, r = 0.60). Pratt et al. speculated that this was due to the inclusion/exclusion rules for nonrecreational items varying over time (and probably over interviewer) and because such expenditures also vary by month much more than home and car payments (1995, p. 76).

Reported total income for 1992 showed low and moderate correlations (Total income, 1992, all responses, r = 0.40; Total income, 1992, open-ended, r = 0.63). Obtaining reliable income data has been consistently problematic. Interviewers asked about income in a series of items.

- G0aC Was your total personal income in 1992 about the same as, more than, or less than $30,000?
- G0aE I will read some dollar ranges, please tell me the range that best estimates your personal income in 1992: less than or equal to $3,000; $3,001-$6,000; $6,001-$9,000; $9,001-$12,000; $12,001-$15,000; $15,001-$18,000; $18,001-$21,000; $21,001-$24,000; $24,001-$27,000; $27,001-$30,000.
- G0aG I will read some dollar ranges, please tell me the range that best estimates your personal income in 1992: $30,001-$40,000; $40,001-$50,000; $50,001-$60,000; $60,001-$70,000; $70,001-$80,000; $80,001-$90,000; more than $90,000.

Figure 25 illustrates the median values of the measures of association for the subject areas.

Figure 25. -- Measures of association (median values), BPS:90/94 field test (multiplied by 100)



SOURCE: Derived from tables V-2 through V-11, Pratt et al., (1994), *Beginning Postsecondary Students Longitudinal Study Second Followup Field Test Report; BPS:90/94.* (NCES 94-370), pp. 67-76.

National Postsecondary Student Aid Study (NPSAS) Field Test Reinterview Studies

The National Postsecondary Student Aid Study (NPSAS) is a comprehensive, nationwide study of students enrolled in less-than-2-year institutions, community and junior colleges, 4-year colleges, and major universities located in the United States and Puerto Rico. Undergraduate, graduate, and first-professional students who receive financial aid, as well as those who do not receive aid, participate in NPSAS. Data are gathered from institutional records and student and parent interviews. The study collects information on student demographics, family income, education expenses, employment education aspirations, parental demographic characteristics, parental support, and how students and their families meet the costs of postsecondary education. In addition to describing characteristics of students enrolled in postsecondary education, the results are used in part to help determine future federal policy regarding student financial aid.

The first NPSAS was conducted during the 1986-87 school year. There were additional waves in 1989-90, 1992-93, and 1995-96. The next wave after 1995-96 is scheduled for 2000-01. This report discusses reinterview studies conducted as part of the 1992-93 and 1995-96 NPSAS field test studies.

NPSAS:1992-93 Field Test Reliability Reinterview

Methodology and Design

The NPSAS:1992-93 field test reinterview study was conducted to evaluate data from its telephone interviews. A subset of the student sample was reinterviewed between one and three months after the initial interview. The same question wordings were used in each of the two interviews. However, the data collection agents changed the mode of administration between the field test and the reinterview data collection, and they alert researchers that this could affect the accuracy of the information (1993, appendix E, p. 2). The field test data were obtained using a CATI system where the respondents' answers were directly entered into a computer database and were subjected to range checks, skip pattern checks, and other logical tests. The reinterview data were entered onto paper forms by telephone interviewers and then were key-entered for computer analysis. Furthermore, the reinterview data sets did not receive the thorough editing which might have discriminated between valid outlier values and erroneous entries. Reinterviews were conducted with 237 students.

Table 46. -- NPSAS:1992-93 field test reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
NPSAS:1992-93 field test	--	237	--	--*	Reliability

*The NPSAS 1992-93 methodology report states there were 7,417 eligible student records in the field test (Loft et al., 1995, pp. 8-21).

SOURCE: Abt Associates, Inc., (1993), *The National Postsecondary Student Aid Study, Field Test Report*, appendix E, p. 2.

Pearson's coefficient of correlation (r) was used to calculate a measure of reliability. Due to concern that outliers might have an inordinate effect on the correlations, the responses to financial aid items were plotted against the original responses and the plots were overlaid with both a linear regression line and the 95 percent confidence intervals for the regression (Abt Associates, Inc., 1993, appendix E, p. 2).

Summary of Results

The subject areas for the NPSAS 1992-93 field test reinterview included questions about postsecondary degrees, use of school facilities and services, financial aid, employment, enrollment, satisfaction with various aspects of school services and facilities, plans for education, employment and volunteer work, demographic characteristics, and miscellaneous questions. Information on 51 items common to both instruments was reported.

The mean values of the Pearson's coefficient of correlation for the subject areas ranged from lows of 0.57 for questions about degrees and about use of facilities and services to a high of 0.90 for demographic items such as gender, race, and the number of dependents. Abt Associates found the percentage of matching entries particularly interesting in the cases where the estimated coefficient was low but the proportion of exact matches was high. For example, question RA015, about the degree to which student course work was leading, had a coefficient of only 0.19, but 78.1 percent of cases had exactly the same answers in both surveys. They pointed this out to alert researchers to use correspondence information along with the estimated reliabilities to assess how well the items performed.

Since the financial aid questions were critical for NPSAS:93, that was the category on which research focused. The mean value of the Pearson's coefficients of correlation for financial aid, 0.60, was moderate, while the mean percent of exact matches for financial aid items was very low (7.7). Even items with high measures of association had low percentages of exact matches.

Abt Associates did not determine why the measures of association were so low for many of the financial aid items. However, it is possible that students might not have known the actual amounts or sources of different types of aid they received. Abt Associates suggested adding an item to the main data collection asking students how certain they were of the amounts they reported. Students who were certain about the amounts of financial aid they received could be separated from those who were not if it became necessary to compare institutional and student-reported amounts in the final survey data (1993, p. 21).

Table 47 shows the median values of the Pearson's coefficients of correlation and of the percent exact matches by subject area.

Table 47. -- NPSAS:1992-93 field test reliability reinterview study median subject area measurements

	Mean percent exact	Pearson's coefficient of correlation	
Subject areas	match	Median	Items in the low range
Enrollment	67.9	High	RA012, RA014
Degrees	37.7	Moderate	RA015, RA016B
Financial aid	7.7	Moderate	RC111U, RC112, RC118, RD018
Employment	46.9	Moderate	--
Used facilities/services	24.9	Moderate	RF011B
Satisfaction	18.4	Moderate	RF011FF
Plans	68.5	Moderate	--
Demographics	52.4	High	--
Miscellaneous	41.8	Moderate	--

Note: The median values for percent exact match are as follows: enrollment, 86.5; degrees, 32.1; financial aid, 8.7; employment, 44.7; used facilities/services, 25.7; satisfaction, 13.1; plans, 80.1; demographics, 55.7; miscellaneous, 41.8.

SOURCE: Derived from table 1 and table A1, Abt Associates, Inc., (1993), *The National Postsecondary Student Aid Study, Field Test Report*, Draft, pp. 6-8, 26-7.

The items with measures of association in the low range are listed below by subject area. Only the variable label was available in the Abt Associates report.

Enrollment

- RA012 Transfer to sample school in 92-93 (r = 0.47)
- RA014 Last level in last term at sample school (r = 0.31)

Degrees

- RA015 To which degree were your courses leading (r = 0.19)
- RA-16B Year completed sample school degree (r = 0.20)

Financial aid

- RC111U Amount borrowed for undergraduate school (r = 0.50)
- RC112 Amount borrowed with federal loans (r = 0.30)
- RC118 Amount still owed in federal loans (r = 0.49)
- RD018 Parental loans for school (r = 0.12)

Used facilities/services

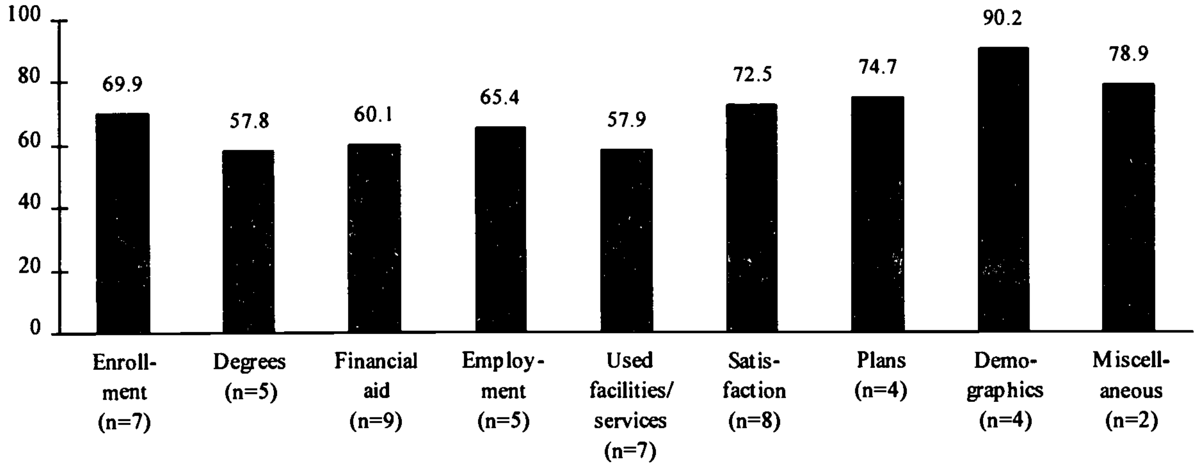
- RF011B Used academic counseling services (r = 0.36)

Satisfaction

- RF011FF Satisfied with cultural/music/art/drama (r = 0.27)

Figure 26 illustrates the mean values of the measure of association. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 26. -- Pearson's coefficient of correlation (mean values), NPSAS:1992-93 field test (multiplied by 100)



Note: For reference, the median values of the subject areas (multiplied by 100) are as follows: enrollment, 80.7; degrees, 69.1; financial aid, 58.7; employment, 65.3; used facilities/services, 56.4; satisfaction, 73.5; plans, 75.7; demographics, 93.9; miscellaneous, 78.9.

SOURCE: Derived from table 1, Abt Associates, Inc., (1993), *The National Postsecondary Student Aid Study, Field Test Report,* Draft, pp. 6-8.

NPSAS:96 Field Test Reliability Reinterview

Methodology and Design

A subset of the eligible sample members who completed the NPSAS:96 field test interview was selected to participate in a reliability reinterview by a random selection algorithm programmed directly into the CATI instrument. A total of 252 students were selected, 249 agreed to participate, and 226 were reinterviewed, for a response rate of 90.8 percent (see table 47). The reinterview sample was fairly representative of the total respondent group in respect to institutional control and student stratum: out of all field test respondents, 49.1 percent attended public institutions, 39.3 percent private, non-profit institutions, and 11.6 percent private, for-profit institutions; 1.4 percent were potential full-time beginning students (FTBs), 30.3 percent were other undergraduates, and 28.3 percent were graduate/first professionals. The reinterviews were generally conducted 2 to 3 weeks after the initial interview (Research Triangle Institute, 1996, pp. III-23 to III-25).

Table 48. -- NPSAS:96 field test reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
Total	252	226	91%	7%	Reliability
Student stratum: potential FTB	103	91	90%	7%	Reliability
Student stratum: other undergraduate	78	68	87%	7%	Reliability
Student stratum: grad./ first professional	71	67	96%	7%	Reliability
Institutional control: public	133	121	92%	7%	Reliability
Institutional control: private, non-profit	97	85	89%	7%	Reliability
Institutional control: private, for profit	22	20	91%	5%	Reliability

SOURCE: Research Triangle Institute, (1996), *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report* (Working Paper 96-17), pp. II-7 to II-8 and III-23 to III-25.

The reinterview study was designed to assess the short-term reliability of selected items on the assumption that lack of agreement (or low correlation) between responses from the same individuals would identify items susceptible to measurement error, items that might need to be revised or deleted. Items were selected based the following criteria

- Items not selected for prior NPSAS or BPS reliability reinterview studies
- Items that, taken together, would be broadly representative of the student interview
- Items that have been problematic in prior NPSAS surveys
- Items for which response should not change over time

Percent agreement and correlational analyses were used to analyze the reinterview results. Pearson's coefficient of correlation (r) was used for continuous measures, such as year of graduation or dollar amounts; Cramer's V statistic (V) for items with discrete, unordered response categories; and both Spearman's Rho (ρ) and Kendall's Tau (τ) for items with discrete, ordered response categories.

Summary of Results

The items chosen for the reinterview covered educational experiences, education expenses and finances, work and community service experiences, and participation in school-related activities.

Educational experiences. Educational experiences in the reinterview covered reports of high school completion and enrollment at the NPSAS postsecondary school. The reliability indices for reports of type and date of high school completion were very high (V = 0.97, r = 1.00), as were reports of first postsecondary school attended and date of first attendance (V = 0.91, r = 0.94). The consistency of student reports of type of degree program enrolled in during the first term at the NPSAS school was lower, but still in the moderate range (V = 0.66). The index for undergraduate responses to the question on level in the program was much higher (ρ = 0.84) than the graduate responses (ρ = 0.32).

Education and living expenses. The results from questions about annual educational expenses and monthly living expenses were fairly consistent with results of prior investigations of similar items and respondent groups, and indicate that students' reports for items dealing with dollar amount estimates (of expenses, awards, earnings) are generally somewhat less stable across time than are their reports of events and activities. The measures of association for all questions but one in this subject area were in the low range (r = 0.24 - 0.43). That one question asked about the number of children or other dependents living with the respondent or receiving at least 50 percent of their support from the respondent (r = 0.84).

Loans. The question on loans asked for three types of information: whether students got loans, the source of the loans, and the amount of the loans. Since few students in the sample received loans, the analysis was based only on whether loans were received. The reliability index was low (V = 0.34); however, "generally, reliability of these data is acceptable and consistent with prior investigations...the relational statistic of 0.34 reflects sensitivity to small systematic changes in the distribution of responses examined" (Research Triangle Institute, p. VI-3).

Employment and community service. In general, the consistency of responses for items asking about students' employment status, participation in a college work study program or assistantship, and performance of community service was high (percent agreements from 95.3 - 98.2 percent; V = 0.69 - 0.77).

Participation in school-related activities (FTBs only). The reinterview included questions asking students how frequently they participated in school-related activities. Researchers were interested in whether the 10-point scale (0 to 9, with 9 indicating 9 or more times) provided better information than the three-point response scale (1 = never, 2 = sometimes, 3 = often) used in prior BPS interviews. While the results were low to moderate (τ = 0.36 - 0. 56), they were also consistent with those of prior BPS studies for similar items. "Part of the problem might stem from vague or unclear item wording, which can be corrected for the full scale. For example, only 32 students provided a scale response for 'participation in student assistance center/programs;' more than half of the FTBs asked about this activity responded 'don't know,' indicating that they were not sure what was meant by this question" (Research Triangle Institute, 1996, p. VI-5).

Table 49 shows the median percent agreement and median values of the measures of association by subject matter area.

Table 49. -- NPSAS:96 field test reliability reinterview study median subject area measurements

Subject areas	Median percent agreement	Measures of association*	
		Median	Items in the low range
Educational experiences	95.6	High V and r / Moderate ρ	A_FSTLVG (ρ)
Education and living expenses	85.8	Low r	B_ED_EXP (COMMUT, TRHOME); B_LIVEXP (FOOD, TRANS, PRSEXP, OTEXP) (r)
Loans	85.3	Low V	C_OTHLON (V)
Employment and community service	96.0	Moderate V	--
Participation in school-related activities (FTBs only)	47.1	Low τ	F_PARTIC (ADVSR, ACDMTG, SOCIAL, STDYGP, ASTCTR, EVENT) (τ)

* Pearson's coefficient of correlation (r), Kendall's Tau (τ) and Spearman's Rho (ρ), and Cramer's V statistic (V).

SOURCE: Derived from tables VI.A1-VI.A5, Research Triangle Institute, (1996), *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report* (Working Paper 96-17), pp. VI-2 to VI-5.

The following are the items with measures of association in the low range.

Educational experiences

- A_FSTLVG During the first term you were enrolled at [fill school] in the 1994-1995 school year, what was your level in the program? [9 categories—first year to ninth year, and beyond tenth year] (ρ= 0.32)

Education and living expenses

B_ED_EXP For the 94-95 school year, how much did you spend for...

- COMMUT ...commuting to class, such as bus fare and gasoline? [range=$0-$5,000] (Do not include the cost of car insurance and maintenance.) (r = 0.35)
- TRHOME ...other educational expenses, (such as transportation to your permanent home or dependent care while attending classes)? [range=$0-$20,000] (r = 0.27)

B-LIVEXP Between July 1, 1994, and June 30, 1994, how much were your average [r]monthly[n] expenses for...

- FOOD ...food, including meals in restaurants and meal plans? [range=$0-$2,000] (r = 0.43)
- TRANS ...car loans, car maintenance, and insurance? [range=$0 $5,000] (Please exclude costs for commuting to school.) (r =0.37)
- PRSEXP ...personal expenses such as clothing, dry cleaning, recreation? [range=$0-$5,000] (r = 0.31)

- OTEXP ...other expenses, such as telephone bills, child support, life or health insurance, or repayment of other loans? [range=$0-$5,000] ($r = 0.24$)

Loans

- C_OTHLON Not including any loans you may have received from the federal government, state government, your school, or your employer, did you receive any loans from parents, relatives, banks, credit unions or other sources for the 94-95 school year? If yes, where did you get the loan and how much did you receive? (1 = parents or guardians, 2 = other relatives or friends, 3 = personal loans secured through your bank, savings and loan, credit union, 4 = other loan, 5 = other loan) ($V = 0.34$—analysis based on students' responses of whether or not loans from these sources were received.)

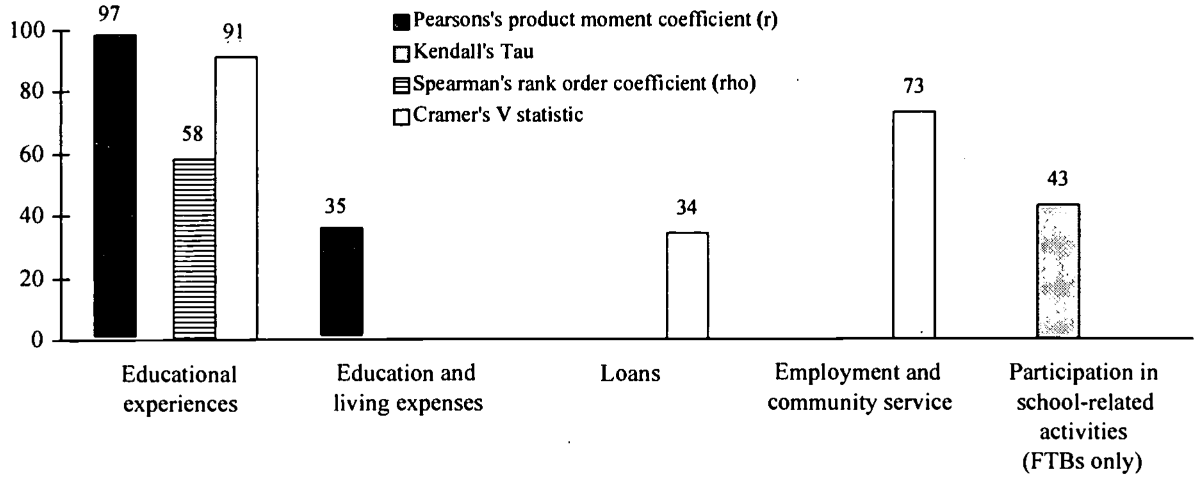
Participation in school-related activities (FTBs only)

F_PARTIC [BPS FTBs only.] I am now going to read you a list of school related activities that you may have participated in during the 1994-95 school year, while you attend [fill school]. Please indicate for each activity how often you participated in the activity.

- ADVSR Talk with faculty about academic matters outside of class time? ($\tau = 0.40$)
- ACDMTG Meet with advisor concerning academic plans? ($\tau = 0.45$)
- SOCIAL Have informal or social contacts with advisor or other faculty members outside of classrooms and offices? ($\tau = 0.41$)
- STDYGP Participate in study groups with other students outside of the classroom? ($\tau = 0.44$)
- ASTCTR Participate in one or more student assistance centers/programs? ($\tau = 0.40$)
- EVENT Attend academic or career-related lectures, conventions, or field trips? ($\tau = 0.36$)

Figure 27 illustrates the median values of the measures of association for the subject areas.

Figure 27. -- Measures of association (median values), NPSAS:96 field test (multiplied by 100)



SOURCE: Derived from tables VI.A1-VI.A5, Research Triangle Institute, (1996), *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report* (Working Paper 96-17), pp. VI-2 to VI-5.

National Study of Postsecondary Faculty (NSOPF) Reinterview Study

The National Study of Postsecondary Faculty (NSOPF) was designed to provide data about faculty and instructional staff to postsecondary education researchers, planners, and policymakers. The data it collects include employment characteristics, academic and professional background, institutional responsibilities and workload, job satisfaction, compensation, and sociodemographic characteristics. Respondents were asked to report information about their activities during the 1992 fall term at the institution listed on the label on the back cover of the questionnaire.

NSOPF was conducted by NCES for the first time in the 1987-88 academic year. The study had three major components: a survey of institutional-level respondents; a survey of a eligible instructional faculty members within the participating institutions; and a survey of eligible department chairpersons. The second cycle of NSOPF, NSOPF-93, gathered information from faculty and institution-level respondents.

Methodology and Design

The goals of the NSOPF-93 field test reliability reinterview were "to identify faculty questionnaire items that yield low quality data and to identify characteristics of items, such as question wording, context, and unclear or ambiguous response categories, that caused response problems...[thus providing] a basis for revising questionnaire items prior to implementation in the full scale study" (Abraham, Suter, Spencer, Johnson, Zahs, Myers and Zimbler, 1994, p. 106). To accomplish these goals, the reinterview questionnaire included a

°subset of the same items that were administered in the original interview, items selected in part because they were identified as being potentially problematic for respondents.

A subsample of 117 out of the 495 faculty who responded to the original interview were reinterviewed. All field test faculty were initially asked to complete a self-administered questionnaire; a small number of respondents who failed to complete a self-administered questionnaire completed a computer-assisted telephone interview (CATI). The reinterviews were conducted via telephone (Abraham et al., 1994, p. 106).

Table 50. -- NSOPF-93 field test reliability reinterview study

Reinterview study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
NSOPF 1993 (faculty)	--	117	--	24%	Reliability

SOURCE: Abraham et al., (1994), *1992-93 National Study of Postsecondary Faculty Field Test* Report (NCES 93-930), p. 125.

The reinterview instrument included items on instructional duties, principal activities, field or discipline, degrees and honors, previous jobs, publications and presentations, funded research, allocation of time, and salary. Unlike the studies discussed previously, Abraham et al. presented their analysis of the reinterview data by item type—categorical or continuous variables—rather than by subject areas. Their analysis of categorical variables used percent inconsistent and inconsistency indexes. For continuous variables, they presented the interview and reinterview means and Pearson correlation coefficients.

Summary of Findings

For most of the categorical and continuous variables analyzed, Abraham et al. state that the level of consistency between data sources was relatively high. They found a 70 percent consistency between the original and the reinterview responses for most of the categorical questions, and a 0.7 correlation between responses for most of the continuous variables. Abraham et al. judged only six of the interview-reinterview items to have unacceptably low correlations (1994, pp. 107 and 124).

Eight categorical variables were evaluated: instructional duties (Question 1), credit or noncredit courses (Question 1A), principal activity (Question 2), principal field (Question 14), last degree (Question 18), employment sector, last main job (Question 19B), level of students in classes (Question 23), and funded research (Question 29). Abraham et al. conducted a detailed analysis of Question 19 because it had the highest percent inconsistent responses (28.0 percent with a standard error of 4.7 percent) and the highest inconsistency index (36.0 with a standard error of 6.0). For the analysis, they collapsed the 11 categories (see below) into five: Ph.D. granting institution; other four-year; elementary or secondary; consulting; and all other. They found that the inconsistencies appeared to be "fairly evenly distributed across possible combinations" and concluded that the high number of response categories and the involvement of some faculty in more than one job sector were plausible reasons for the high rate of inconsistency (1994, p. 111).

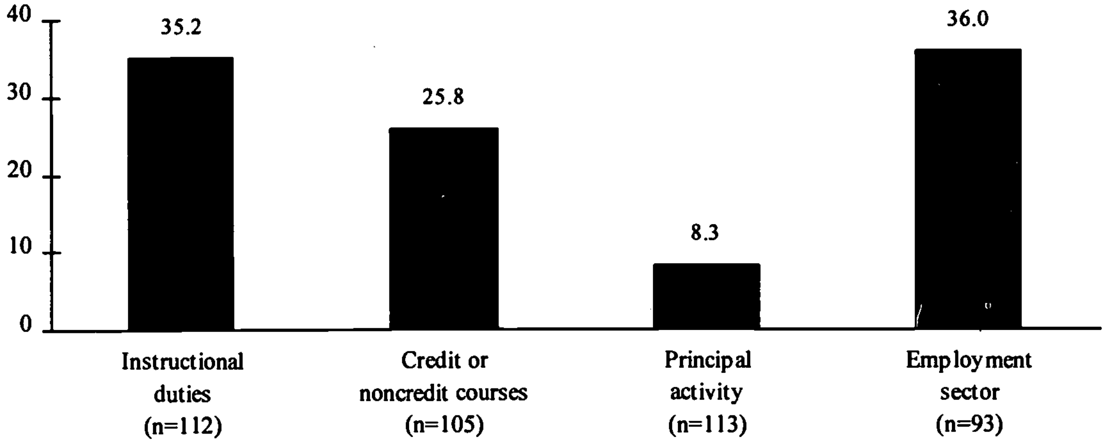
19. Next I'm going to ask about the last significant and main job that you held previous to your current position at (institution).

- B. What was the employment sector for that job? Was it (a/an) ...
  Doctoral degree granting university or college, including professional schools
  Other 4-year college or university
  2-year postsecondary institution
  Less-than-2-year postsecondary institution
  Elementary or secondary school
  Hospital or other health care or clinical setting
  Consulting, freelance work, or self-owned business
  Foundation or other nonprofit organization
  For-profit business or industry in the private sector
  Federal government position, including military, or state or local government
  Other place

However, as they state, "given the high standard errors associated with a sample of 117 cases, we do not have evidence of poor reliability" (Abraham et al., 1994, p. 107).

Figure 28 illustrates the four categorical variables for which inconsistency indices were calculated. In this case, the n equals the number of cases.

Figure 28. -- Mean IOI, NSOPF-93



SOURCE: Derived from exhibit 10-3, Abraham et al., (1994), *1992-93 National Study of Postsecondary Faculty Field Test* Report (NCES 93-930), p. 111.

For the 19 continuous variables examined in the reliability reinterview, 11 had high correlations, 3 had moderate correlations, and 5 had low correlations. While Abraham et al. again pointed out that the level of precision possible with 117 or fewer sample cases is not high, they were reassured by how high most of the correlations were (four were greater than 0.91, another 4 were greater than 0.81). The items which had the lowest correlations were those asking for retrospective reporting of numbers that were small fractions of dollars or hours, and they asked for summary statistics on activities that were likely to fluctuate over time. Studies have shown that responses to these types of questions tend to be unreliable (1994, pp. 111-13). The five variables with low correlations were

37. Now I'm going to ask some questions about how you spent your time during the 1991 Fall Term. On the average, how many hours per week did you spend in (activity) during the 1991 Fall Term? (average number hours per week during the 1991 Fall Term)

- c. Unpaid or pro bono professional service activities (r = 0.31)

Other response categories: a. All activities including teaching, research, administration, etc., b. Any other paid activities, such as consulting, working on other jobs

38. The next question is about what percentage of those (Question 37 number) hours you spent during the 1991 Fall term in teaching, professional growth, research or scholarship and in non-teaching activities such as administration or service. As I ask you about each activity, make note of the percentage; the number you give me for all activities should add up to 100 percent. (First/next), I will ask you about (four teaching/two professional growth/four research or scholarship/six non-teaching) activities.

- Professional Growth (r = 0.13)
  e. Taking courses, pursuing an advanced degree
  f. Other professional development activities, such as practice or activities to remain current in your field
- Research/Scholarship (r = 0.29)
  g. Research, that is time spent in activities that will lead to a concrete product, such as an article, grant proposal, software development, etc.
  h. Reviewing or preparing articles or books; attending or preparing
  for professional meetings or conferences; reviewing proposals
  i. Seeking outside funding, including proposal writing
  j. Giving performances or exhibitions in the fine or applied arts, or giving speeches
- Administrative/Service/Other Non-teaching (r = 0.47)
  k. Administrative activities, including paperwork, staff supervision, serving on in-house committees such as the academic senate, etc.

l. Providing legal or medical services or psychological counseling to clients or patients

m. Outside consulting or freelance work, working at a self-owned business

n. Paid or unpaid community or public service, civic, religious, etc.

o. Service to professional societies/associations

p. Any other non-teaching activities? (What________________)

51. For the calendar year 1991, estimate your gross earnings before taxes from...

- h. Outside consulting, consulting business or freelance work (r = 0.40)

Other response category: a. basic salary

CHAPTER 5
"Multiple Indicators" Studies

Survey respondents

> *may not always possess accurate information or for other reasons may provide inaccurate answers. Good survey practices require the examination of the quality of the data collected. Assessment of data quality leads to better analysis and interpretation of the data and improvements in the designs of future studies* (Fetters, Stowe and Owings, 1984, p. v.)

One means of examining the quality of the data in this case is to use multiple indicators of the same characteristic administered in a single survey. NCES surveys typically have multiple components that provide multiple indicators of the same characteristic. For example, the High School and Beyond (HS&B) has items that are common to both the student component and the parent component. In addition, HS&B also includes a transcript component which provides external record check data. Although the results of the transcript record check study would be more appropriately described in the next chapter we have included it in this chapter where we first describe HS&B measurement error studies.

The most striking difference between this approach and those mentioned in earlier chapters is that an explicit model of correspondence between two or more different survey questions is posited. We introduce this notion by dealing with the simplest (and most constraining model), *parallel measures*, when two questions measure the same underlying characteristic with the same degree of precision. In notation, the simplest set of parallel measures is

$$\begin{aligned} y_{ikm_1} - y_{ikm_2} &= (X_{ik} - X_{ik}) + (e_{ikm_1} - e_{ikm_2}), \\ &= e_{ikm_1} - e_{ikm_2} \end{aligned}$$

where y_{ikm_1} = indicator m_1 of the *k*th underlying characteristic for the *i*th unit;

y_{ikm_2} = indicator m_2 of the *k*th underlying characteristic for the *i*th unit;

X_{ik} = true value of the *k*th characteristic for the *i*th unit;

e_{ikm_1} = random error terms for indicator m_1

e_{ikm_2} = random error terms for indicator m_2

Each of these measurements has the simple classical true score assumptions (reviewed in chapter 1). Another assumption gives a boost to assessing the reliability of y_{ikm} as an indicator of X_{ik}. Just as psychometricians were forced to assume independence between measurement error in a test-retest situation, a similar assumption is required here in order to get an unbiased estimate of the correlation of scores over replications. So in this situation, we assume that the error committed on one indicator is not correlated to the error on the other; that is, $\mathrm{Cov}(e_{im_1}, e_{im_2}) = 0$. This makes the existence of two parallel measures even better than the test-retest situation, because there is no threat of change in true values (X_i) with parallel measures administered. *Conditional on the two being parallel measures and errors independent*, the reliability of y_{ikm_1} or y_{ikm_2} , is

$$r_{y_{ikm_1},\, y_{ikm_2}} = \frac{\mathrm{Cov}(y_{ikm_1}, y_{ikm_2})}{\mathrm{Var}(y_{ikm_1})\mathrm{Var}(y_{ikm_2})}$$

Other models of correspondence between two indicators permit more complex assessment of reliability. Alwin and Jackson (1980) review different models.

The remainder of the chapter describes the results of some of the "multiple indicators" studies conducted by NCES. Some of the characteristics of the studies included in this report are summarized in table 51 below. NPSAS has also conducted some "multiple indicators" studies. NPSAS collects information from the parents of a subsample of students and has conducted studies examining the correlations between student and parent responses.

Table 51. -- "Multiple Indicators" studies source of data, sample size, number of matched pairs, and match rate

Source of data	Study Sample size	Matched pairs	Match rate
HS&B base year[1]			
Parent survey (sophomores)	3,654	3,367	92%
Parent survey (seniors)	3,547	3,197	90%
Twin component	636 pairs	511 pairs	80%
Transcript study (sophomores)	18,152	15,941	88%
NELS:88 base year[2]			
Student	24,599	22,651	92%
Parent	22,651	22,651	100%
NSOPF-93[3]	417	333	80%

[1] A random sample of 312 of the schools that fully participated in HS&B was chosen for the parent survey. In each parent survey school, simple random samples of 12 sophomores and 12 seniors were selected from those students who had completed HS&B questionnaires and taken the HS&B tests. A total of 30,030 completed questionnaires were collected from sophomores and 28,240 from seniors.

[2] The number of logical student-parent pairs depended primarily on the skip pattern of previous items and whether the mother or the father responded to the parent questionnaires. Therefore, all 22,651 student-parent pairs were used in the analysis of some items, while other items (e.g., father's educational expectations for the student) were based on much smaller logical sample sizes. See table 56 for item-by-item sample sizes.

[3] The data collection agent re-contacted all institutions whose faculty list counts were discrepant from their institutional questionnaire counts by 10 percent or more.

SOURCE: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

High School and Beyond (HS&B)

High School and Beyond (HS&B) provides information on the educational, vocational, and personal development of young people as they move from high school into postsecondary education or into the work force and then into adult life. It was the second longitudinal study NCES initiated, and was designed to complement the first, the National Longitudinal Study of the High School Class of 1972 (NLS-72). HS&B studied the sophomore and senior high school students of 1980. There was group administration of questionnaires and tests to some 30,000 sophomores and 28,000 seniors, including more than 500 sets of twins, in more than 1,000 public and private schools. The longitudinal design called for followup surveys of substantial subsets of the two cohorts at two-year intervals. Data collection for the first followup began in spring 1982.

Methodology and Design of HS&B Base Year Quality of Responses Study

The HS&B Quality of Responses Study (Fetters, Stowe and Owings, 1984) looked not only at response bias, but was an early attempt by NCES to test response variance. High school student responses to group-administered questionnaires were judged against the standards of parent responses and transcript data, and the responses of both members of more than 500 twin pairs were compared. Finally, the quality of data was evaluated as a function of item type and the age, sex, race/ethnicity, and other characteristics of the respondents.

The parent responses and transcript data were used as standards because they were assumed to be more accurate than the student responses. However, the farther apart the collections of data are in time, the more likely a difference between the two responses reflects true change rather than unreliability. In this context, it must be remembered that there was a six-month lag between the collection of data from the parents and from the students, which makes the standards against which accuracy was judged somewhat unreliable.

NCES also compared data from twins. Reliability is generally determined by the consistency of repeated independent measurement of a fixed value with the same instrument. Thus, studies of response reliability have used the test-retest approach, where the expected value of the response is assumed to be the same for both repetitions. However, as stated above, the farther apart the repetitions are in time, the more likely a difference between the two responses reflects true change rather than unreliability. Hence it was felt that

> *the correlation between responses of twins who were administered questionnaires at the same session should be a better estimate of the reliability of certain items than the estimate resulting from the test-retest approach. The test-retest approach, on the other hand, provides reliability estimates for attitudinal variables while the approach utilizing twin data does not. The twin data are employed in this study to estimate the reliability coefficients for home and family background variables and certain present and past school experiences. All twin pairs were attending the same school*

> *and living in the same household in spring 1980. The estimates of reliability coefficients assume, for some variables, that both members of the pair shared common school experiences (e.g., kindergarten attendance) and home experiences at earlier stages of their lives as well. Estimated coefficients also are presented for a few school variables (e.g., instructional methods used in courses taken) for which the true value might not always be the same for both members of the twin pair. For these variables, the correlation between twin responses sets a lower bound for the reliability coefficient* (Fetters, Stowe and Owings, 1984, p.11).

The Quality of Responses study used two measures of association to compute reliability: Pearson's coefficient of correlation and Cramer's V statistic. Since these statistics are discussed in the beginning of chapter 4, all that will be said here is that they are rated as low (less than 0.5), moderate (0.5 to 0.8), and high (greater than 0.8). A high coefficient indicates a high degree of reliability, which is supportive of the validity of the instrument. The standard errors calculated for this study are summarized in two tables at the end of this section.

The difference in means between student responses and those obtained from the standard—parents or transcripts—was used to measure bias. Therefore, the total error of student reports of parental income, for example, may be thought of as the root mean square of the sums of two components.

Sources of Data

Twin component of student survey. The base-year survey of HS&B used a two-stage stratified sample. Public and private schools were stratified according to several key variables. Schools within each stratum were then selected with probabilities proportional to estimated average grade 10 and/or grade 12 enrollment. Certain types of schools were oversampled. In the second stage, simple random samples of 36 sophomores and 36 seniors, school size permitting, were chosen from each selected school. Efforts were made to identify twins among those students. If the co-twin of a twin selected to participate attended the same high school and had not been already randomly selected into the sample, the co-twin was asked to take part. Questionnaires were completed by both members of 276 sophomore and 235 senior twin sets.

Parent survey. A random sample of 312 of the schools that fully participated in HS&B was chosen for the parent survey. In each parent survey school, simple random samples of 12 sophomores and 12 seniors were selected from those students who had completed HS&B questionnaires and taken the HS&B tests, for a total of 3,654 sophomores and 3,547 seniors. Data were collected from the parents of these students by a combination of mailed-out questionnaires, telephone interviews, and personal interviews. Data collection occurred between October 10, 1980 and December 31, 1980. The response rate was 91 percent: 3,367 forms were completed by parents of sophomores and 3,197 by parents of seniors.

Transcript study. The sample for the transcript study consisted of 18,152 selections from the 1980 sophomores who were eligible to participate in the first followup survey of HS&B. Transcripts were provided by high schools in fall 1982 for 15,941 (88 percent) of the individuals for whom they were requested.

Table 52. -- HS&B base year quality of responses study

Source of data	Sample size	Completed questionnaires or collected transcripts	Response rate
Sophomores	35,723	30,030	84%
Seniors	34,981	28,240	81%
Twin component	636 pairs	511 pairs	80%
Parent study	7,201	6,564	91%
Transcript study	18,152	15,941	88%

SOURCE: Fetters, Stowe and Owings, (1984), *High School and Beyond, A National Longitudinal Study for the 1980's, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), pp. 4-6. Jones et al. (1986), *High School and Beyond 1980 Senior Cohort Second Follow-up (1984) Data File User's Manual*, pp. 17-18. National Center for Education Statistics, (1982), *High School and Beyond: Twins and Siblings' File Users' Manual*, p. 3.

Treatment of the Data

Cases were omitted from the analysis when data were missing due to instrument or item nonresponse or when data were reported in an invalid way for one or both of the sources being compared. In addition, cases were omitted whenever at least one member of a child-parent or twin/co-twin pair answered "don't know" or "not applicable" to a variable measured on a regular scale.

Both members of twin pairs answered the same questions. The format of some parent questionnaire items differed for the same items answered by students, but parent responses were transformed into the student questionnaire format since NCES was evaluating the quality of student responses.

Summary of Results for the HS&B Study

There were four subject areas: family background items, school-related items, grade-related items, and financial and college items. Family background items included parents' education levels, family income, and family possessions. School-related items asked about attendance, number of school changes, school rules, disciplinary problems, etc. Grade-related items included questions on grade averages, parental aspirations for children, and future plans, as well as course work and courses. Financial and college items asked what respondents thought college would cost, who would pay the costs, etc.

Comparing Data from Twins

The twin data were used primarily to estimate the coefficients for home and family background variables and certain present and past school experiences. The parent questionnaire data were used as a standard in estimating the validity of family background items, the grade-related items having to do with grades, expectations, and attitudes, and financial and college items. Transcript data were used as the standard when estimating coefficients to test the validity of grade, course work, and courses items.

Using the twin files allowed for an extensive investigation of data quality. Both twins completed all of the items on the student questionnaires whereas many student questionnaire items were not included in the parent questionnaire. There were three major findings.

- The coefficient of correlation measuring reliability for responses regarding family background, estimated from twin data, were largely correlated with (r = 0.92) and slightly larger (0.05, on the average) than coefficient of correlation measuring validity estimated by comparing student and parent responses.

- When using students as sources of information about their schools with regard to judgmental questions—rule enforcement, disciplinary problems, and several other matters—the measures of association were only between 0.20 and 0.40. Although most measures for individual items were low, many items could be aggregated to the school level, and the composites and school means had much higher measures of association than did individual items or students.

- Within the file, 42 percent of the measures of association are at least 0.75 in value while 44 percent are between 0.50 and 0.74.

Items[*] having low measures of association follow. See table 55 for standard errors.

Family background

36 Which of the following people live in the same household with you?

- d Mother (sophomore, r = 0.44)

103/104 Which of the following do you have in your home? [have/do not have]

- a Specific place to study in home (sophomore, r = 0.33; senior, r = 0.30)
- c Encyclopedia, other reference books in home (sophomore, r = 0.37; senior, r = 0.49)

[*] When an item was on both the sophomore and senior questionnaires, both item numbers are given: sophomore questionnaire item number/senior questionnaire item number.

- g More than 50 books in home (sophomore, r = 0.36; senior, r = 0.32)
- i Pocket calculator in home (sophomore, r = 0.19; senior, r = 0.40)

School-related items

3 How often has each of the following been used in the courses you are taking this year? [used never/seldom/fairly often/frequently]

- a Listening to the teacher's lecture (senior, r = 0.25)
- b Participating in student-centered discussions (senior, r = 0.28)
- c Working on a project or in a laboratory (senior, r = 0.30)
- d Writing essays, themes, poetry, or stories (senior, r = 0.33)
- e Having individualized instruction (senior, r = 0.19)
- f Using teaching machines or computer-assisted instruction (senior, r = 0.05)

19 To what extent are the following disciplinary matters problems in your school? [often happens/sometimes happens/never happens]

- Students don't attend school (sophomore, r = 0.36)
- Students cut classes, even if they attend school (sophomore, r = 0.38)
- Students talk back to teachers (sophomore, r = 0.23)
- Students refuse to obey instructions (sophomore, r = 0.19)
- Students get in fights with each other (sophomore, r = 0.30)
- Students attack or threaten to attack teachers (sophomore, r = 0.18)

20 Listed below are certain rules which some schools have. Please mark those which are enforced in your school:

- School grounds closed to students at lunch time (sophomore, r = 0.38)
- Students responsible to the school for property damage (sophomore, r = 0.26)
- Hall passes required (sophomore, r = 0.43)
- "No smoking" rules (sophomore, r = 0.36)

52 How much has each of the following interfered with your education at this school? [not at all/somewhat/a great deal]

- a Courses are too hard (senior, r = 0.23)
- b Find it hard to adjust to school routine (senior, r = 0.14)
- c Poor teaching (senior, r = 0.24)
- d Poor study habits (senior, r = 0.18)
- e Courses are too easy (senior, r = 0.22)

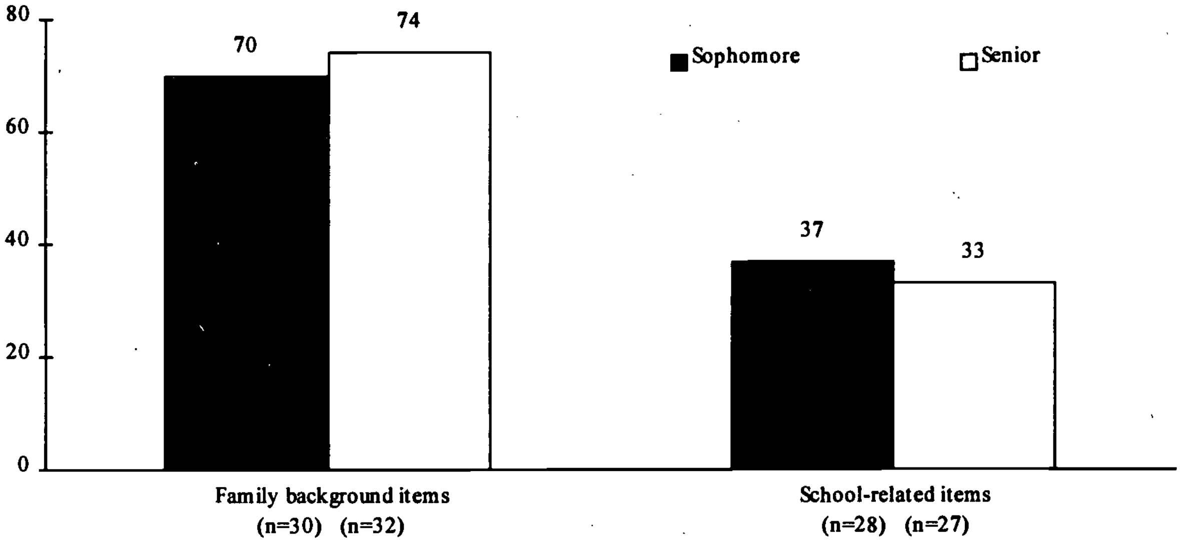
57/53 Please rate your school on each of the following aspects [poor/fair/good/ excellent/don't know]

- a Condition of buildings and classrooms (sophomore, r = 0.39)
- b Library facilities (sophomore, r = 0.29; senior, r = 0.39)
- c Quality of academic instruction (sophomore, r = 0.20; senior, r = 0.32)
- d Reputation in the community (sophomore, r = 0.23; senior, r = 0.44)
- e Teacher interest in students (sophomore, r = 0.23; senior, r = 0.35)
- f Effective discipline (sophomore, r = 0.21; senior, r = 0.26)
- g Fairness of discipline (sophomore, r = 0.21; senior, r = 0.24)
- h School spirit (sophomore, r = 0.31; senior, r = 0.38)

- 58 Does your high school have a minimum competency or proficiency test—that is, a special test that all students must pass in order to get a high school diploma [yes/no/don't know] (sophomore, r = 0.45)

Figure 29 shows the median values of the measures of association calculated when correlating responses of students in the twin component. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 29. -- Pearson's coefficient of correlation (median values) for sophomore and senior responses using twin file component, HS&B base year quality of responses study (multiplied by 100)



SOURCE: Derived from tables A.3 and A.6, Fetters, Stowe and Owings, (1984), *High School and Beyond, A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), pp. 43 and 46.

Comparing Student Responses to Parent Questionnaire Data

Using parent questionnaire data as the standard, there were three major findings:

- The quality of HS&B student questionnaire data generally was high for contemporaneous, factual information. For example, the measures of association were almost 0.90 for father's educational attainment, although only about 0.60 for father's occupation. The validity of the socioeconomic status (SES) composite that was often employed in HS&B data analyses was found to be at least 0.80.

- The quality of retrospective information tended to decline with the passage of time. For example, when students were asked about which periods of time their mothers worked, the correlations were about 0.71 for during high school, about 0.64 for during elementary school, and 0.53 for prior to elementary school.

- The validity of attitudinal items tended to be lower; e.g., about 0.60 for the mother's aspirations for her child's education.

Items having low measures of association follow. See table 55 for standard errors.

Family background

- 41/41 Please describe below the job most recently held by your mother (stepmother or female guardian) even if she is not working at present. (sophomore, V = 0.44; senior, V = 0.45)

103 Which of the following do you have in your home? [have/do not have]

- a Pocket calculator in home (sophomore, r = 0.39)
- c More than 50 books in home (sophomore, r = 0.35)
- g Encyclopedia, other reference books in home (sophomore, r = 0.35)
- i Specific place to study in home (sophomore, r = 0.21)

The study also examined the consistency between parent and child attitudinal variables; that is, the degree to which parents and children share the same perceptions on variables for which the parental response could not be taken as a factual standard.

Grade, expectation, and attitude items

49 How much has each of the following persons influenced your plans for after high school? [not at all/somewhat/a great deal]

- a Your father (senior, r = 0.21)
- b Your mother (senior, r = 0.18)

78/81 At what age do you expect to...[don't expect to do this/have already done this/ under 18/18/19/20/21/22/23/24/25/26/27/28/29 /30 or more]

- a Marry (sophomore, r = 0.42)
- b Have first child (sophomore, r = 0.39)

- c Start regular job (sophomore, r = 0.32; senior, r = 0.48)
- d Live in own home (sophomore, r = 0.27; senior, r = 0.40)
- e Finish full-time education (sophomore, r = 0.43; senior, r = 0.48)

68/62 Write in here the name of the job or occupation that you expect or plan to have when you are 30 years old. (sophomore, V = 0.31; senior, V = 0.37)

72/68 Did you expect to go to college when you were in the following grades? [yes/ no/was not sure/hadn't thought about it]

- 72a-b Grade 6 or 7 (sophomore, r = 0.43)
- 72c-d/68a-b Grade 8 or 9 (sophomore, r = 0.45; senior, r = 0.40)
- 68c Grade 10 (senior, r = 0.42)
- 68d Grade 11 (senior, r = 0.46)

- 73/69 Whatever your plans, do you think you have the ability to complete college? [yes, definitely/yes, probably/not sure/I doubt it/definitely not] (sophomore, r = 0.40; senior, r = 0.42)

63 How do you feel about each of the following statements? [agree strongly/ agree/disagree/disagree strongly]

- A working mother of pre-school children can be just as good a mother as the woman who doesn't work (sophomore, r = 0.18)
- b It is usually better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family (sophomore, r = 0.21)
- c Most women are happiest when they are making a home and caring for children (sophomore, r = 0.18)

Financial and college choice items

- 76 About how much money do you expect to spend on living expenses (such as room and board and clothing) next year? (i.e., Sept. 1980 to Aug. 1981) [almost none—I plan to live at home/none, for other reasons/less than $1,000/$1,000 to $1,999/$2,000 to $2,999/$3,000 to $3,999/$4,000 to $4,999/$5,000 to $6,999/$7,000 to $10,000/more than $10,000] (senior, r = 0.22)

79 How do you plan to pay for the living expenses and schooling expenses (if any) you may have next year? For each source listed below, indicate how much money you expect to receive in the year beginning July 1980 and ending June 1981 [none/under $300/$300 to $599/$600 to $1,199/$1,200 to $2,000]

79A My family

- a Parents (senior, r = 0.47)

- b Husband or wife (senior, r = 0.12)
- c Other relatives (senior, r = 0.16)

79B Myself

- a Summer earnings (senior, r = 0.18)
- b Earnings Sept. 1980 to Aug. 1981 (senior, r = 0.27)
- c Savings (senior, r = 0.17)

79C Other sources

- a State scholarship or grant (senior, r = 0.33)
- b Federal scholarship or grant (senior, r = 0.33)
- c Other scholarship or grant (senior, r = 0.38)
- d State loan (senior, r = 0.17)
- e Federal loan (senior, r = 0.16)
- f Other loan (senior, r = 0.07)
- g Social security or Veterans Administration benefits (senior, r = 0.47)

111/111 How much do you think each of the following kinds of schooling would cost for a year? [Under $500/$500-$1,000/$1,001-$2,000/$2,001-$3,000/$3,001-$5,000/$5,001-$7,000/don't know]

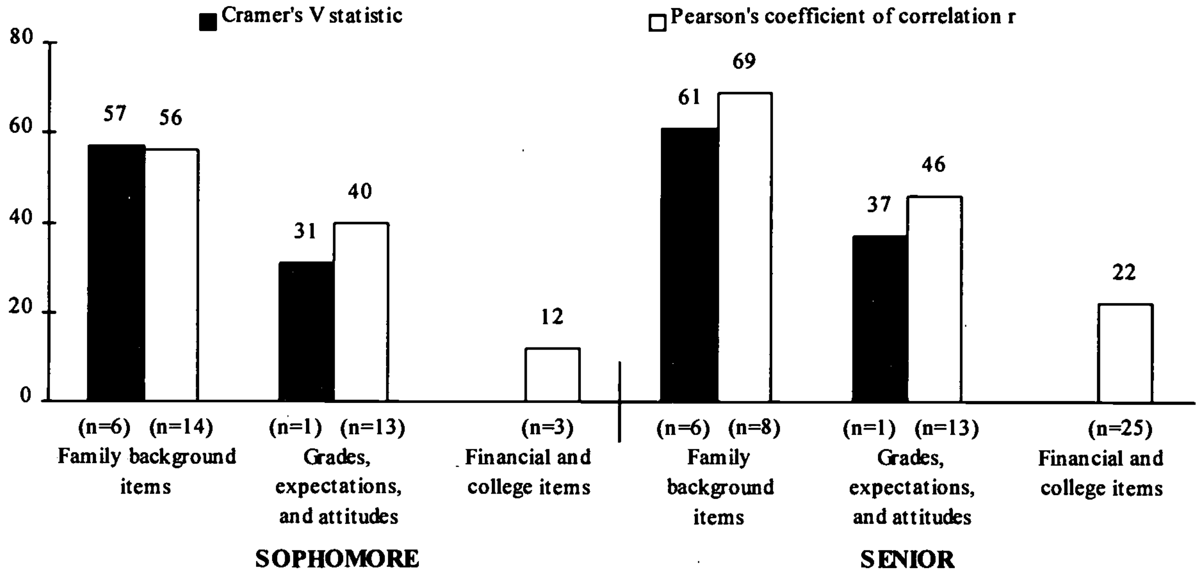
- a Public junior or community college (sophomore, r = 0.14; senior, r = 0.25)
- b State four-year college or university (sophomore 0.12, r =; senior, r = 0.15)
- c Private four-year college or university (sophomore, r = 0.12; senior, r = 0.06)

116 How important are each of the following in choosing a college you plan to attend? [not important/somewhat important/very important]

- a College expenses (tuition, books, room and board) (senior, r = 0.24)
- b Availability of financial aid (such as a school loan, scholarship, or grant) (senior, r = 0.44)
- c Availability of specific courses or curriculum (senior, r = 0.11)
- d Reputation of the college in academic areas (senior, r = 0.15)
- e Reputation of the college in athletic programs (senior, r = 0.32)
- f Social life at the college (senior, r = 0.18)

Figure 30 shows the median values of the measures of association calculated when correlating responses of students and parents.

Figure 30. -- Measures of association (median values) for sophomore and senior responses using parent file, HS&B base year quality of responses study (multiplied by 100)



SOURCE: Derived from tables A.1, A.7, and A.9, Fetters, Stowe and Owings, (1984), *High School and Beyond, A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), pp. 39, 47, and 50.

Comparing Student Responses to Transcript Data

Using transcript data as the standard, there were two major findings.

- The grade point averages reported by seniors correlated well (0.77) with grade point averages computed from their transcripts. Seniors reported their grades to be somewhat higher (about one-fourth of a letter), however, than shown by their transcripts. [Note: This is an example of why using the correlation to measure response bias is not a good method, since one can get very high correlation and still have bias.]

- The correlations between senior reports and transcript data were high (in the 0.80s) for amount of course work in specific foreign languages (French and German, 0.87; Spanish 0.86) and for whether geometry (0.85), physics (0.80), or chemistry (0.87) ever was taken. The correlations were somewhat lower, ranging from 0.63 to 0.70, for amounts of course work in mathematics (0.66) and science (0.70) and for whether second year algebra (0.68), trigonometry (0.63), and calculus (0.67) were ever taken. They were lower yet for the two areas (history or social studies, 0.39, and English or literature, 0.28) that show relatively little variation from student to student in amount of course work taken. Seniors tended to report they had taken more course work in most areas than reflected by their transcripts. Their claims were greatest for mathematics (about one semester) and science (about 1/2 semester).

Items having low measures of association follow. See table 55 for standard errors.

Course work, 10th grade through end of 1981-82 school year (semesters):

6/4 (Sophomore: During the tenth grade, including all of this school year,...) (Senior: Starting with the beginning of the tenth grade and through the end of this school year,...) how much course work will you have taken in each of the following subjects? [none/1-2 year/1 year/1 1/2 years/2 years/2 1-2 years/3 years/more than 3 years]

- b English or literature (all students 0.28)
- f History or social studies (all students 0.39)

Table 53. -- HS&B base year quality of responses study median subject area measurements

Subject areas	Twin component	Standard—Parent survey	Standard—Transcript	Items in the low range
Family background items				36; 41; 104a, c, g, i
Sophomore	Moderate r	Moderate r & V	--	
Senior	Moderate r	Moderate r & V	--	
Grade-related items				4b, f; 49a-b; 62; 68a-d; 69; 81a-e
Sophomore	--	Low r & V	--	
Senior	--	Low r & V	Moderate r	
School-related items				3a-f, 53a-h, 52a-e
Sophomore	Low r	--	--	
Senior	Low r	--	--	
Financial and college items				76, 79 (all), 111a-c, 116a-f
Sophomore	--	Low r	--	
Senior	--	Low r	--	

SOURCE: Derived from tables A.1, A.3, A.4, A.6, A.7, and A.9, Fetters, Stowe and Owings, (1984), *High School and Beyond, A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), pp. 39, 43-4, 46-7, and 50.

Subgroup Comparisons

The quality of responses (i.e., the accuracy, as measured by agreement coefficients) was found to be better for some groups of students. There were four major findings.

- Seniors provided a higher degree of concordance than sophomores. The average coefficient for 12 family background items when correlating student and parent responses was 0.64 for sophomores, 0.67 for seniors, and the average coefficient for 31 family background items when using data from the twin component was 0.66 for sophomores and 0.70 for seniors. Looking at twin data with regard to school-related variables, the average measures of association were 0.44 for sophomores and 0.54 for seniors.

- Female students provided a slightly higher degree of concordance than males. The overall average measures of association when correlating student and parent responses were 0.64 for females and 0.61 for males for family background items, and 0.72 for females and 0.69 for males for high school grades and course work. The difference in mean values is small (about 0.26), but highly significant.

- White students showed a higher degree of concordance than Hispanic or black students. For example, the mean coefficients when correlating student and parent responses for family background items were 0.61 for whites, 0.57 for Hispanics, and 0.54 for blacks; and for high school grades, the coefficients were 0.80, 0.66, and 0.65, respectively. Also, Hispanic and black students overstated to a greater degree than white students the amount of course work they had taken in mathematics and science.

- Students who performed well academically on the cognitive tests in HS&B had a higher degree of concordance in their responses about family background characteristics than students who performed poorly, with a difference in mean coefficients of 0.107. From lowest to highest test score quartile, the mean measure of association when correlating student and parent responses rose from 0.56 to 0.67 for family background items and from 0.47 to 0.73 for amount of course work taken. Some student questionnaire and transcript mean values also agreed less closely for low- than for high-scoring students. The most extreme example was the amount of mathematics taken, where the mean calculated from questionnaires exceeded that calculated from transcripts by 1.8 semesters for students in the lowest test score quartile, but by only 0.6 semester for those in the highest quartile.

Table 54 displays results of subgroup comparisons for selected characteristics.

Table 54. -- Measures of association (mean values) for family background variables of 1980 sophomores and seniors, by selected student and parent characteristics

Subgroup	Number of coefficients averaged	Measure of association (mean value)
Cohort		
Sophomore	12	0.635
Senior	12	0.671
Sex of students		
Male	35	0.613
Female	35	0.639
Race/ethnicity of students		
White	29	0.612
Hispanic	29	0.570
Black	29	0.544
Cognitive test performance		
Lowest quartile	35	0.562
Highest quartile	35	0.669

SOURCE: Derived from table 3.2, Fetters, Stowe and Owings, (1984), *A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), p. 15.

Table 55. -- Approximate standard errors of measures of association* for the HS&B base year quality of responses study

Subgroups	Sophomores			Seniors		
	r = 0	r = 0.50	r = 0.90	r = 0	r = 0.50	r = 0.90
Student-transcript comparisons	--	--	--			
Total population	--	--	--	0.019	0.015	0.004
Male or female students	--	--	--	0.028	0.021	0.005
Test or SES quartile subgroups	--	--	--	0.042	0.032	0.008
White students	--	--	--	0.025	0.019	0.005
Black students	--	--	--	0.053	0.043	0.010
Hispanic students	--	--	--	0.043	0.032	0.008
Parent-child comparisons						
Total population	0.027	0.020	0.005	0.028	0.021	0.005
Male or female students	0.040	0.030	0.008	0.040	0.030	0.008
Test or SES quartile subgroups	0.055	0.041	0.011	0.055	0.041	0.011
White students	0.032	0.024	0.006	0.033	0.025	0.006
Black students	0.072	0.054	0.014	0.079	0.059	0.015
Hispanic students	0.084	0.063	0.016	0.091	0.068	0.017
Female respondents (Parent questionnaire)	0.034	0.026	0.006	0.036	0.027	0.007
Male respondents (Parent questionnaire)	0.045	0.034	0.009	0.045	0.034	0.009
Twin/co-twin comparisons						
Total population	0.094	0.071	0.018	0.102	0.077	0.015

*The standard errors presented in the table were calculated from the formula $\mathrm{var}(r) = D(1 - r^2)^2/(.8n)$, where r is the estimated measure of association, D is a survey design adjustment factor, and n is the sample size (Kendall & Stewart, 1958). Conservative values of D were employed: 2.0 for parent-child and twin/co-twin comparisons and 4.0 for comparison based on transcript data (Tourangeau et al., 1983). Sample sizes were reduced by 20 percent to adjust for cases not usable in the analysis due to item nonresponse, "don't know" replies, etc.

The analysis for a few items was restricted to subsets of cases, viz., language in home, mother had job, mother's aspirations for child's education, estimated cost of school, and college choice factors. For the first item, the tabled standard errors should be doubled; for the remaining ones, they should be increased by 30 percent.

SOURCE: Derived from table A.11, Fetters, Stowe and Owings, (1984), *A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), p. 52.

Response Bias

In the final effort of the Quality of Responses study, HS&B identified the direction of response error based on the parents' responses as the "true" or correct response. This is an example of an early attempt by NCES to estimate response bias.

Students overreported, for example, that their parents had a high school education only (45 versus 36 percent, for sophomore mothers), but underreported that their parents had had some but less than 2 years of postsecondary education (10 versus 24 percent, sophomore parents). The size of the bias in student reports of parental education tended to be smaller for the father's than for the mother's education level, and smaller for seniors than for sophomores.

Students tended to classify their father's occupation as clerical and craftsman less than they should have (2 versus 5 percent and 17 versus 23 percent, respectively), but to claim their fathers were farmers and laborers more than they should have (5 versus 2 percent and 10 versus 5 percent). Similarly, they underclassified the occupations of their mothers as clerical (31 versus 35 percent, sophomores), but overclassified their mother's occupation as a laborer (3 versus 1 percent) and professional or school teacher (23 versus 17 percent).

When student and parent responses on family income were examined, 68 percent of sophomores but only 40 percent of their parents were found to report income values between $7,000 and $25,000 per year. Family income was underestimated by an average of about $3,000 (12 percent). The income ranges used were 0 to $6,999; $7,000 to $11,999; $12,000 to $15,999; $16,000 to $19,999; $20,000 to $24,999; $25,000 to $37,999; and $38,000 and above.

Students who indicated they came from a foreign-language background tended to overreport that the language was Spanish (23 versus 17 percent, sophomore) and that there was a second language, which was Italian, French, or German (18 versus 10 percent, sophomore). For the most part, biases were smaller for seniors than for sophomores.

Table 56. -- Approximate standard errors of bias estimates (percentage points)— parents as standard*

Measure of association		Percentage (P)	
	P = 50	P = 20 (or (80)	P = 10 (or 90)
0.2	2.0	1.6	1.2
0.4	1.7	1.4	1.0
0.6	1.4	1.1	0.8
0.8	1.0	0.8	0.6

*The size of the standard error is a function of the correlation (r) between child and parent responses and the percentage values for child (P1) and parent (P2). Computation of standard errors were based on the equation $Var(B) = Var(P_1) + Var(P_2) - 2cov(P_1P_2)$, where $b = P_1 - P_2$. When $P_1 = P_2 = P$, $Var(b) = 2(1 - r)Var(P)$. Estimates were adjusted for item nonresponse, "don't know" responses, etc., by reducing n by 20 percent and for survey design effect by increasing the simple random sampling estimates by 40 percent. The standard error estimates may be used for either cohort.

The analysis for a few items was restricted to subsets of cases, viz., language in home, mother had job, mother's aspirations for child's education, estimated cost of school, and college choice factors. For the first item, the tabled standard errors should be doubled; for the remaining ones, they should be increased by 30 percent.

SOURCE: Derived from table A.12, Fetters, Stowe and Owings, (1984), *A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items* (NCES 84-216), p. 53.

National Education Longitudinal Study of 1988 (NELS:88)

The National Education Longitudinal Study of 1988 (NELS:88) was the third major longitudinal study sponsored by NCES, after the National Longitudinal Study of 1972 (NLS-72) and High School and Beyond (HS&B). The earlier studies surveyed high school seniors (and sophomores in HS&B) through high school, postsecondary education, and work and family formation experiences. NELS:88 expanded this base of knowledge by following young adolescents starting at an earlier age (8th grade) and by updating information throughout the 1990s.

NCES never considered retesting or reinterviewing NELS:88 respondents because NCES considered such activities too heavy a burden on the respondents. Moreover, the reinterview studies that were conducted by HS&B were believed to be relevant to NELS. Instead, NELS concentrated its efforts on validity evaluation and cognitive research. NLS-72 and HS&B studies had concluded that students were relatively good sources of information about family background variables. However, accuracy of student reporting was often systematically affected by the way the questions were asked, the specific information sought, and the characteristics of the student. In *Quality of the Responses of Eighth-Grade Students in NELS:88* (Kaufman, Rasinski, Lee and West, 1991), NCES sought to compare NELS:88 data quality with HS&B data quality (see previous section).

Methodology and Design of NELS:88 Base Year Quality of Responses Study

NELS:88 assessed the quality of base-year eighth-grade student data by judging selected student responses against the standard of parent or teacher responses and by examining their consistency with other student items. The analyses were conducted without the use of the weights associated with the NELS:88 database. That is, errors in the responses to questionnaire items were directly linked to the wording of particular items, placement of items in the questionnaire, and conditions under which the questions were administered. This was done because the HS&B Quality of Response study used unweighted data. However, Kaufman et al. (1991) conducted both weighted and unweighted analyses for a sample of survey items. The results indicated that use of one or the other produced few differences in the data quality indicators used in their study. Kaufman et al. likewise used the same measures of association employed in the HS&B study: Pearson's coefficient of correlation and Cramer's V statistic.

Table 57. -- NELS:88 base year quality of responses study

Variable	Number of valid pairs	Percent missing
Race-ethnicity	22,651	1.6
Number of siblings	22,651	2.4
Number of older siblings	21,300	4.0
Father's education	22,222	17.1
Mother's education	19,184	13.2
Father's occupation	18,796	4.1
Mother's occupation	22,600	10.8
Mother home	22,651	5.0
Father home	22,651	9.1
Other adult home	22,651	14.9
Father's expectations for student's education	4,190	11.4
Mother's expectations for student's education	18,300	12.8
Language usually spoken at home	3,635	5.3

SOURCE: Derived from table 3.2, Kaufman et al. (1991), *Quality of the Responses of Eighth-Grade Students in NELS:1988* (NCES 91-487), p. 14.

Correspondence between Student and Parent Responses

Kaufman et al. used three statistics to assess correspondence between student and parent responses: what they termed the validity coefficient, that is the correlation of student and parent responses; the percentage of students whose response identically matched their parent's response; and the relevant bias in student responses. For this study, the bias in student responses was expressed as "the difference between the mean of the parent response and the mean of the student response divided by the mean of the parent response" (1991, p. 6).

Two measures of association were used to calculate the validity coefficients: Pearson's coefficient of correlation (r) was used with ordinal variables, Cramer's V statistic with nominal variables. (Note that although Pearson's r is frequently used with ordinal-level data, some measurement theorists and statisticians oppose its use with data that is not at least at the interval level, and some measurement theorists vary on whether Likert-type data is considered ordinal- or interval-level data. These statistics are described in more detail at the beginning of chapter 4.)

Kaufman et al. felt that examining the correlation between parent and student responses alone could be misleading because "the marginal distributions of a pair of variables can have a dramatic impact on the size of correlation between them" (1991, p. 14). Therefore, they also looked at the percentage of cases where student responses matched their parents' responses. Used together, these statistics provide a clearer picture of data quality than either statistic provides on its own. A high correlation and a high percentage of matched cases indicate high quality for the student responses, at least as judged by how they correspond to the parent responses. It follows that a low correlation and a low percentage of matched cases indicates low quality data. If the correlation between student and parent responses is low, but the percentage of matched cases is high, then the distributional properties of the variables should be investigated. If the correlation between student and parent responses is high, but the percentage of matched cases is low, then student responses are almost certainly biased to some degree. The size of the bias, however, depends in part on the units of the original items.

Kaufman et al. (1991) had other caveats to make about the correspondence between student parent response on school experience items. The assumption is that one source contains the true values. Parents, however, may not be the most accurate reporters of school-related information. Moreover, several of the school experience items were attitudinal in nature. For those items, the correlations are more appropriately regarded as measures of consistency between parent and student responses than measures of validity. Finally, some of the parent and student items were not exact matches. For example, some student questionnaire items referred to the first semester while the corresponding question on the parent questionnaire referred to the entire school year.

These three statistics were used in one more way. The HS&B study had found that the quality of responses was better for some groups of students than for others. For example, seniors had provided higher quality data than sophomores. Therefore, Kaufman et al. generated the validity coefficients, the percentage of matched cases, and the relative bias statistics for the whole sample of students and for various subgroups to assess whether the data quality was constant across all students or varied systematically in relation to student characteristics. The student characteristics investigated were sex, race-ethnicity, family income, socioeconomic status, and reading level.

Inter-item Consistency of Student Responses

Kaufman et al. defined inter-item consistency as "a measure of the reliability of student responses from one item to the next" (1991, p. 10). They examined inter-item consistency in terms of how reliably students reported on similar factual items and on how consistently student responded on less factual or subjective items.

Reliability of Scales

Finally, Kaufman et al. (1991) assessed the reliability of several scales created from the student, teacher, and school administrator data files. These scales, or composite variables, included teacher engagement, academic press, discipline climate, and student behavior. Kaufman et al. tested both the inter-item reliability of these variables (using Cronbach's Alpha) and looked at how reliable the scales were for different subgroups of students.

Cronbach's Alpha measures how well items in a scale correlate with one another. It should be interpreted as an estimate of the degree to which all items within a scale correlate with each of the other items. Thus, it is virtually a type of average across all of the correlations that exist within a given set of items. As Kaufman et al. define it, if the items in the scale are standardized to have the same variance, alpha can be computed using the following formula

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$

where K is the number of items in the scale and $\bar{r}$ is the average correlation between items (1991, p. 37).

Summary of Results for the NELS:88 Study

The NELS:88 Quality of Responses study looked at items in two subject areas: family background and school experiences. Kaufman et al. (1991) found that generally the student responses on family background items were reliable and accurate. The correspondence between student and parent responses was within conventional standards of validity, although lower than those in HS&B. However, since NELS:88 asked questions of eighth-grade students while HS&B questioned high school sophomores and seniors, Kaufman et al. had expected lower results. The validity of most of the school-related items was not as high as those for the family background items. This had also been seen in the HS&B study. However, the percentage of matched cases demonstrated a much better correspondence between the student and parent responses to these items than had the validity coefficients. Furthermore, the inter-item consistency check showed that the majority of students were answering the items consistently. More detailed information is presented below.

Family background items. These items included race-ethnicity, parents' occupations and education, the number of siblings, the language usually spoken at home, and the people at home after school. The correlations ranged from a low of (r =) 0.41 for "father's expectations for the student's education" to a high of (r =) 0.85 for the "number of older siblings." Like the students in the HS&B study, the eighth-grade students in NELS:88 showed higher correlations with their parents' responses when providing factual information than attitudinal information. This can be seen for the following items, for example: race-ethnicity (V = 0.77, percentage matched = 91.6), number of siblings (r = 0.83, percentage matched = 82.2), number of older siblings (r = 0.85, percentage matched = 86.4), but father's expectations for student's education (r = 0.41, percentage matched = 47.5) and mother's expectations for student's education (r = 0.43, percentage matched = 43.1). However, Kaufman et al. suggest several possible explanations for the low correlations for parents' expectations for their children's education. For example, since the children were only in eighth grade, the parents might not yet have discussed higher education with their children, or even thought much about it.

Comparison of student and parent responses to items about the parents' education demonstrated the importance of using a variety of statistics to analyze quality of data. The validity coefficients for father's and mother's education were high (father, r = 0.82; mother, r = 0.76), but the percentage of matched cases was moderate (father = 61.0, mother = 62.5). The relative bias showed that students systematically overestimated the level of their father's education by about seven percent (0.066) and underestimated their mother's by about eight percent (-0.082).

Items[*] having low measures of association follow.

Family background items

- 4b/34b & 37b What kind of work does she normally do? That is, what is the job called? (V = 0.42, percent matched = 47.8, relative bias not applicable)

40/72 Are any of the following people at home when you return home from school? [usually/sometimes/rarely/never]

- c. other adult relative (r = 0.48, percent matched = 60.5, relative bias = -0.029)

Other categories were a. your mother or female guardian, b. your father or male guardian, d. a sitter, e. an adult neighbor, f. older brother or sister, g. younger brother or sister, h. no one is home. Results were reported only for a, b, and c.

48/76 How far in school do you think your father and your mother want you to get? BE SURE TO ANSWER BOTH A AND B BELOW. [less than high school graduation/ graduate from high school, but not go any further/go to vocational, trade, or business school after high school/attend college/graduate from college/ attend a higher level of school after graduating from college/don't know]

- a. Father (or male guardian) (r = 0.41, percent matched = 47.5, relative bias = 0.062)
- b. Mother (or female guardian) (r = 0.43, percent matched = 43.1, relative bias = 0.078)

School experience items. These items asked if students discussed school experiences with their parents, whether students were enrolled in gifted or bilingual programs, whether the school was safe, and if parents were contacted by the school.

As mentioned earlier, Kaufman et al. (1991) had several caveats to make about the school experience items. The assumption is that one source contains the true values, but parents may not be the most accurate reporters of school-related information. Moreover, for attitudinal items—which several of the school experience items are—the correlations are more appropriately regarded as measures of consistency between parent and student responses than measures of validity. Finally, some of the parent and student items were not exact matches. For example, some student questionnaire items referred to the first semester while the corresponding question on the parent questionnaire referred to the entire school year.

[*] The questions are given as they were asked on the student questionnaire, but both the student questionnaire number and the corresponding parent questionnaire number are listed, student number first.

That said, the correlations for school experience items had a median value of r = 0.20, ranging from r = 0.08 for "student enrolled in a bilingual program" to r = 0.51 for "student enrolled in a gifted class." The percentage of cases matched ranged from 47.1 for "school is safe" to 92.9 for "student enrolled in a bilingual program" (median value, 62 percent).

The item on whether a student was enrolled in a bilingual program had unusually disparate numbers (r = 0.08, percent matched 92.9). Almost all of the students and parents (19,018 out of 20,477 valid student-parent pairs) agreed that the student was not in a bilingual class. However, among the parents who said their child was in a bilingual class, 86 percent of their children said they were not. Among the children who said they attended a bilingual class, 91 percent of their parents disagreed. Kaufman et al. speculated that differences in question wording may have been part of the problem, but that it is also possible neither parents nor students may have known what a bilingual or bicultural class was (1991, p. 24).

Student questionnaire item

68 Are you enrolled in any of the following special programs/services? [yes/no]

- b. Special instruction for those whose first language is not English—for example, bilingual education or English as a second language (not regular English classes)

Parent questionnaire item

49 Is your eighth grader currently enrolled in any of the following special programs/services? [yes/no/don't know]

- a. Bilingual or bicultural education program
- b. English as a second language program

Items[*] with low measures of association follow.

School experience items

- 36/66 Since the beginning of the school year, how often have you discussed the following with either or both of your parents or guardians? [not at all/once or twice/three or more times] a. selecting courses or programs at school, b. school activities or events of particular interest to you, c. things you've studied in class. (r = 0.16, percent matched = 51.1, relative bias = -0.138)

55/57 During the first semester of the current school year, has any of the following things happened to you? [never/once or twice/more than twice]

- D. My parents received a warning about my grades (r = 0.19, percent matched = 47.8, relative bias = -0.465)

[*] The questions are given as they were asked on the student questionnaire, but both the student questionnaire number and the corresponding parent questionnaire number are listed, student number first.

- E. My parents received a warning about my behavior (r = 0.44, percent matched = 71.9, relative bias = -0.580)

59/74 How much do you agree with each of the following statements about your school and teachers? [strongly agree/agree/disagree/strongly disagree]

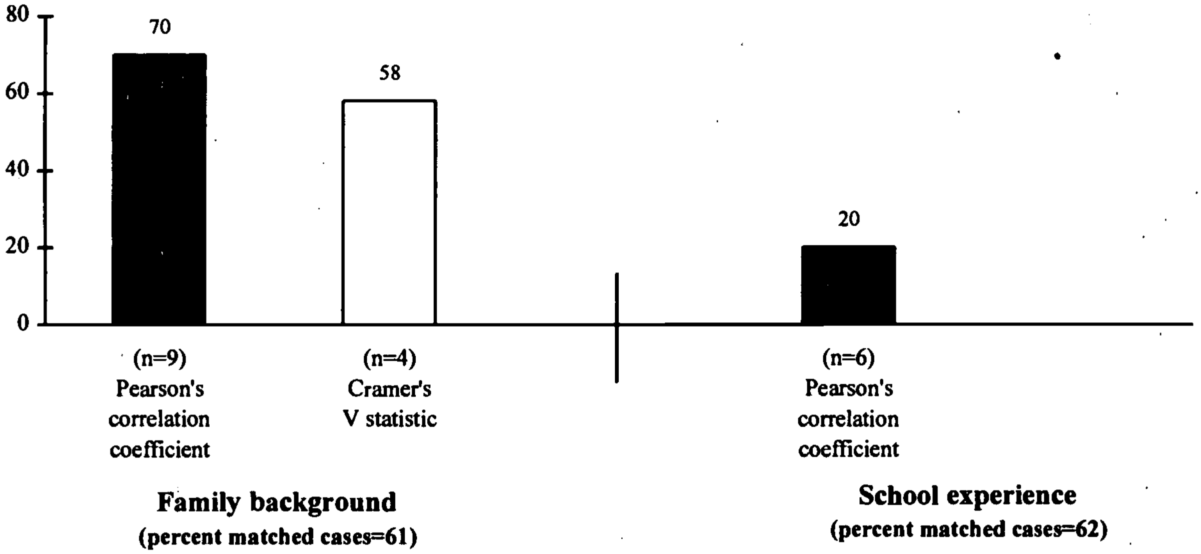
- K/I. I don't feel safe at this school. (r = 0.20, percent matched = 47.1, relative bias = 0.289)

68/49 Are you enrolled in any of the following special programs/services? [yes/no]

- B. Special instruction for those whose first language is not English—for example, bilingual education or English as a second language (not regular English classes) (r = 0.08, percent matched = 92.9, relative bias = -0.008)

Figure 31 shows the median values of the measures of association calculated for the family background and school experience subject areas. The number of items (n) included in each subject area is indicated below the subject area title.

Figure 31. -- Measures of association (median values), NELS:88 base year quality of responses study (multiplied by 100)



SOURCE: Derived from tables 3.2 and 3.9, Kaufman et al., (1991), *Quality of the Responses of Eighth-Grade Students in NELS:88*, pp. 14 and 24.

Subgroup Comparisons

As stated earlier, Kaufman et al. (1991) examined subgroups based on the characteristics of sex, race-ethnicity, family income, socioeconomic status, and reading level. The quality of responses was found to be better for some group of students than others. Specifically, students from high socioeconomic backgrounds, those with higher reading ability, white or Asian students, and female students were more likely to give more valid responses than their peers. However, the differences in the validity coefficients and percentage of cases matched were small.

Family background items. There were no practical differences among the validity coefficients, percentage of matched cases, and relative bias for items when comparing male and female responses. Race-ethnicity did yield some differences. The validity coefficients and percentage of matched cases were generally higher for Asians and whites than for Hispanics and blacks. The mean validity coefficients and mean percentage of matched cases for the family background items were as follows: for Asians, r = 0.65, percent matched = 63.1; for whites, r = 0.64, percent matched = 65.5; for Hispanics, r = 0.59, percent matched = 59.8; and for blacks, r = 0.53, percent matched 56.5. Interestingly, examining the relative biases showed that blacks tended to underestimate their father's educational attainment and to overestimate the mother's, whereas Asian, Hispanic, and white students tended to do the opposite.

Students of lower socioeconomic status (SES) and those with lower reading abilities tended to have lower correlations to the family background items than students of higher SES and those with better reading skills. The lowest SES quartile and the lowest reading quartile had average validity coefficients of 0.53 and 0.52, respectively, compared to 0.60 for the highest SES quartile and 0.61 for the highest reading quartile. The average percent of cases matched showed similar disparities: 60.1 and 57.3 for the lowest SES quartile and the lowest reading quartile compared to 67.9 and 69.1 for the highest SES quartile and the highest reading quartile. However, there was some variation among the subgroups on individual items.

School experience items. While the validity coefficients showed that females agreed more often with their parents on three items—"enrolled in a gifted class," " school is safe," "and "discuss school with parent"—the percentage of cases matched showed that females agreed more often with their parents on all items than did the males. The average scores for the females were r = 0.27, percent matched = 69.0 and for the males, r = 0.24, percent matched = 62.9.

The averages by race-ethnicity group were as follows: for Asians, r = 0.27, percent matched 66.7; for whites, r = 0.26, percent matched = 67.3; for Hispanics, r = 0.23, percent matched = 63.0; and for blacks, r = 0.20, percent matched = 61.4. Kaufman et al. (1991) found it interesting that the validity coefficients were low even for Hispanics and Asians on the bilingual education item (r = 0.15 and 0.09, respectively) although they were higher than for the blacks and whites (r = 0.02 and 0.03, respectively). However, the percentage of case matched on this item showed the parent and student responses agreed more among blacks and whites than among Hispanics and Asians: Hispanics = 88.1, Asians = 87.9, blacks = 90.0, and whites = 94.7.

The correspondence between parent and student responses by SES background and reading ability followed a pattern similar to that shown with the family background items. Contrasting the averages for the highest and lowest SES and reading quartiles shows, for highest: SES r = 0.25, percent matched = 67.6; reading r = 0.26, percent matched = 69.5; for lowest: SES r = 0.22, percent matched 66.0; reading r - 0.18, percent matched - 59.7.

Reliability of scales

The scale, or composite, variables proved to be reliable, particularly the school-administrator level scales. The reliability of the school-administrator scales—school problems, teacher involvement, academic press, school security, and discipline climate—ranged from a low of 0.708 for academic press to a high of 0.881 for the school problems scale.

The student-level scales—school problems, locus of control, self-concept, teacher quality—ranged from a low of 0.572 for locus of control 1 to a high of 0.920 for school problems. Kaufman et al. presented two locus of control and two self-concept scales. The first of each were more comparable to the scales in HS&B, but the second versions were able to take advantage of NELS:88 having more relevant items to create measures designed to be more stable (1991, p. 37). For example, locus of control 1 used 3 items; its reliability was 0.572. Locus of control 2 used 6 items and its reliability was 0.678. However, it should be noted that reliability of student-level scales did differ somewhat by subgroup. In particular, the reliability of both the self-concept scales was lower for black students and for students with low reading abilities than it was for other students.

National Study of Postsecondary Faculty (NSOPF)

The National Study of Postsecondary Faculty (NSOPF) provides data on faculty and instructional staff including employment characteristics, academic and professional background, institutional responsibilities and workload, job satisfaction, compensation, and sociodemographic characteristics. A reinterview study conducted during the NSOPF-93 Faculty field test was discussed in chapter 4. The 1996 study discussed here was titled a retrieval, verification, and reconciliation. This study was designed to reconcile estimates of the total number and the number of full- and part-time faculty and instructional staff from two different sources effort (Selfa, Suter, Myers, Johnson, Zahs, Kuhr, Abraham and Zimbler, 1997).

Background

The Chief Administrative Officer (CAO), as identified for the institution on the Integrated Postsecondary Education Data System (IPEDS) Institutional Characteristics (IC) Survey, appointed someone to be responsible for providing NCES with a list of faculty and instructional staff. This was the list from which the NSOPF-93 sample was taken. The CAO also identified an institutional respondent who completed the Institution Survey for NSOPF-93. One of the primary issues addressed on the Institution Survey was academic turnover. For this reason, the institutional respondent was asked to provide a count of faculty and instructional staff by employment status, full-time versus part-time, and presence or absence of instructional responsibilities. The weighted estimates based on the faculty lists and the institution questionnaires raised some concern, however, because several patterns emerged that were contrary to expected results.

Although some variance in the estimates based on the lists and the institution questionnaires was expected, the large magnitude of the difference was not. This, in and of itself, was not seen as a problem since the estimates were from two different sources. Less plausible were the trends in the estimates of part-time faculty between NSOPF-88 and NSOPF-93. The Institution Survey showed a 5 percentage point increase in the estimate of part-time faculty between the fall of 1987 and the fall of 1993. The Faculty Survey, based on the lists of faculty and instructional staff provided by the institution, showed no change in the percentage of part-time faculty between the two points in time.

A second pattern that raised concern was that the weighted estimates based on the lists also showed a 37.5 percent decrease in the number of health sciences faculty and instructional staff from the fall of 1987 to the fall of 1992. Third, closer inspection of the estimates revealed that more than one-half (450 out of 817) of the institutions had discrepancies between the two sources of greater than 10 percent.

NSOPF-93 Retrieval, Verification, and Reconciliation

NCES conducted a retrieval, verification, and reconciliation in an attempt to discover if there were any evidence of systematic nonsampling errors in the list collection. We must assume there is one correct response. Bias is the difference from that true response, and bias has been shown to have a discernible pattern compared to the true response. For example, if respondents failed to list faculty and instructional staff from a particular discipline, say medical faculty, then the estimated total number of faculty would be expected to be lower than the true number. Lacking an external source that could provide the true number, NCES took the reconciled estimates as the "true" response.

Methodology and Design

The study involved a one-stage sample of the institutions who provided an estimate of faculty on the list (LIST) which differed from the estimate reported on the Institution Survey (QUEX) by 10 percent or more. A total of 509 institutions were recontacted, 450 institutions showing this 10 percent difference and 59 institutions NCES designated as operating medical schools or hospitals. Responses were received from 492 institutions for a response rate of 96.6 percent (see table 58). The field period for this effort lasted from January 3 through February 21, 1996.

Table 58. -- NSOPF-93 retrieval, verification, and reconciliation

"Multiple Indicators" study	Sample size	Completed reinterviews	Response rate	Sampling percentage	Primary study purpose
NSOPF-93	509	492	97%	62%	Response bias

SOURCE: Selfa et al., (1997), *1993 National Study of Postsecondary Faculty Methodology Report* (NCES 97-467).

Five telephone interviewers were selected and trained to conduct the reinterview and reconciliation. According to Forsman and Schreiner's (1991) optimal requirements for a reinterview designed to estimate response bias, the choice of respondent should be the person most knowledgeable, not necessarily the original respondent. The interviewers were instructed to begin the reinterview with the Institution Survey respondent because that was the last person contacted at each institution, but they were instructed to contact the individual that provided the list, if necessary, and any others suggested as more knowledgeable, including staff in human resources or personnel, or the director of institutional research.

Based on the reinterview, interviewers marked whether the most accurate data were found on the LIST, QUEX, or neither and made corrections to the LIST and/or QUEX estimates. Interviewers also asked the institution respondents to provide the most common reasons for the discrepancies. These reasons were entered on a coded list.

Summary of Results

According to the results of the reconciliation, 280 institutions (63 percent) identified the Institution Survey questionnaire estimate (QUEX) as being the most accurate response, and 122 (24.8 percent) stated the list estimate (LIST) as being more accurate. Fifty-six institutions (11.4 percent) provided a new estimate, and five institutions (1 percent) chose IPEDS as the most accurate estimate. Twenty-nine institutions (5.9 percent) could not verify any of the estimates and therefore accepted the original LIST estimates.

As mentioned above, institutions were asked to explain why there were discrepancies between the LIST and QUEX. Institutions were allowed to offer up to three reasons. Three hundred and seventy-four institutions provided at least one explanation. Their responses are shown in table 59.

Table 59. -- Explanations institutions gave for discrepancies between LIST and QUEX

Reason for Discrepancy	Percent of total reasons (all valid answers) (n=464)
Different academic base years for LIST and QUEX	1.7
Different academic terms used for LIST and QUEX	11.4
Layoffs or downsizing	1.9
All part-time or adjunct faculty excluded from LIST	4.7
All part-time or adjunct faculty excluded from QUEX	1.7
Some part-time or adjunct faculty excluded from LIST	22.4
Some part-time or adjunct faculty excluded from QUEX	6.9
Some full-time faculty excluded from LIST	16.4
Some full-time faculty excluded from QUEX	3.2
Higher QUEX figure is an aggregate of all campuses	3.4
Higher LIST figure is an aggregate of all campuses	1.5
Medical school excluded from LIST	0.6
Medical school excluded from QUEX	1.5
Unpaid/honorary faculty excluded	1.3
Ineligible faculty included in error	5.4
Data entry error by institution	2.6
Different definitions of full-time faculty used	3.9
Different definitions of part-time faculty used	4.3
Full-time equivalents (FTEs) used instead of headcount	0.6
Other	4.3

SOURCE: Derived from exhibit 3.4, appendix R, Selfa et al., (1997), *1993 National Study of Postsecondary Faculty Methodology Report* (NCES 97-467).

The majority of discrepancies were the result of excluding some full- or part-time faculty either on the LIST or QUEX. Another factor was the time interval between when the LIST was compiled and the time the QUEX was completed. Downsizing also affected faculty counts at several institutions. Selfa et al. (1997) point out that about 118 of the reconciled institutions could not provide a specific reason for the discrepancies.

Some of the reasons institutions gave had not been expected. For example, some institutions provided "full-time equivalents" (FTEs) on the institution questionnaire instead of an actual headcount of part-time faculty. This was only observed at three institutions, but Selfa et al. (1997) suggest this may highlight a bias towards underreporting part-time faculty members. They speculate that some institutions may prefer to report FTEs because the number of instructional faculty an institution employs can be a sensitive issue.

Sometimes, part-time faculty were overreported. The reason tended to involve confusion between the pool of part-time or temporary staff employed by or available to the institution during the course of the academic year, and the number actually employed during the fall semester. Another reason for faculty overreporting was an inability to distinguish honorary/unpaid part-time faculty from paid faculty and teaching staff.

This study also confirmed that a small number of institutions excluded medical school faculty from their lists of faculty. In those cases, the institutions considered their medical schools as separate from their main campuses.

While these results indicated that there may have been some evidence of bias in the NSOPF-93 sample, no measure of the potential bias, such as the net difference rate (NDR), was computed. Instead, the reconciliation prompted NCES to apply a poststratification adjustment to the estimates based entirely on the "best" estimates obtained during the reinterview.

CHAPTER 6
Record Check Studies

Another means of examining the quality of data is to use an external data source that is not subject to the same sources of measurement error as the survey data. Repeated measurement studies that compare results from different sources are called record check studies. In these studies, researchers obtain external data on individual persons in an NCES survey. Since these external data are considered to be true, record check studies are used to estimate response bias.

To recapitulate what was said in chapter 1, there are three kinds of record check studies: reverse record check, forward record check, and full design record check. In the *reverse record check study*, the researcher goes back to the records which were the source of the sample to check the survey responses. The survey data are compared with the record data to estimate measurement error. The weakness of reverse record check studies is that they cannot by themselves measure errors of overreporting. They can only measure what portion of the records sample corresponds to events reported in the survey and whether the characteristics of the events are the same on the records as in the survey report.

In a *forward record check study*, the researcher obtains the survey data first and then moves to new sources of record data for the validity evaluation. Thus, in this design, the sample is drawn from a separate frame. Some surveys may be designed to include questions asking about where records containing similar information on the sample person can be found. Forward record check studies work well for measuring overreports in a survey, but they are not commonly used because they generally entail contacting several different record-keeping agencies and may require asking the respondents for permission to access their record files from the different agencies. They are also limited in their measurement of underreporting. "They learn about the failure to report events only when mention of those events appears on records corresponding to other events which *are* reported. Records are not searched for those respondents who fail to report any event" (Groves, 1989, pp. 301-302).

In the *full design record check* study, the survey sample comes from a frame covering all persons of the population (reverse record check design) and researchers seek records from all sources relevant to those persons (forward record check design). Thus, researchers measure survey errors associated both with underreporting and overreporting by comparing all records corresponding to the respondent. However, this design requires a database that covers all persons in the target population and all events corresponding to those persons.

All validity evaluation designs share three limitations. As mentioned earlier, there is the assumption that the record systems do not contain errors of coverage, nonresponse, or missing data. Second, it is also assumed that the individual records are complete and accurate, without any measurement errors. The third limitation involves matching errors—difficulties matching respondent survey records with the administrative records—and these could affect the estimation of measurement errors. As Groves explains, "*If mismatches* *occur at random within the subsets,* the expected mean difference between interview responses and mismatched records will be equal to that of the expected mean difference between interview responses and correctly matched records. However, even under such restrictive assumptions, the variance in response errors will be overestimated with the possibility of mismatching and the regression of measured response error on the matched record value will have a smaller slope than that of correct response error on the correct record value" (Groves, 1989, p. 302).

Record Check Model

To organize the discussion, we first ignore measurement variability and focus on the bias or fixed errors (Lessler and Kalsbeek, 1992). Thus, we will describe models that assume a fixed-bias for individuals. In the following models one assumes that the record check values are the true values. Of course, it is recognized that these record check values in many cases are not "true values" in any ultimate sense, but are to be treated as preferred values in the context.

For the i^{th} element in the population, let

Y_i = the measurement obtained from the survey for the i^{th} element
X_i = the record check value for the i^{th} element
$B_i = Y_i - X_i$ = individual fixed bias for the i^{th} element.

These individual fixed biases may or may not have net effect on the survey estimate for the population mean, $\overline{X} = \frac{1}{N}\sum X_i$, because the average bias, $\overline{B} = \frac{1}{N}\sum_{i=1}^{N} B_i$ may be negligible. The variance of the fixed bias is defined as

$$Var(B) = \frac{1}{N-1}\sum_{i=1}^{N}(B_i - \overline{B})^2 .$$

We will consider three different record check methodologies below. We assume simple random sampling in each case. For more complex designs, the estimates will need to be modified. Finite population correction factors have been omitted throughout.

Case 1: Record check for each element in the sample

A simple random sample of size *n* is drawn and measurement *(y_i)* obtained for the elements in the sample. A record check *(x_i)* is performed for all the elements in the sample.

The estimate of the average bias, $\overline{B}$, is obtained from the sample as

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \text{, where } b_i = y_i - x_i.$$

The estimate for the variance of the estimated average bias, $V(\bar{b}) = \frac{S_B^2}{n}$, can be calculated from the sample as $\frac{s_b^2}{n}$ where

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(b_i - \bar{b})^2 .$$

And an improved estimate of the true population mean, $\bar{X}$, can be obtained from the record check values as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Case 2: Record check using an independent sample

A simple random sample of size *n* is drawn and measurements *(y_i)* obtained for the elements in the sample. This is followed by drawing an independent simple random sample of size m and measurements *(y_i')* are obtained. A record check *(x_i')* is determined for all the elements in the independent sample of size *m*.

The estimate of the average bias, $\overline{B}$, is obtained from the independent sample as

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i \text{, where } b_i = y_i' - x_i'.$$

The estimate for the variance of the estimated average bias, $V(\bar{b}) = \frac{S_B^2}{m}$, can be calculated from the sample as $\frac{s_b^2}{m}$ where

$$s_b^2 = \frac{1}{m-1}\sum_{i=1}^{m}(b_i - \bar{b})^2 .$$

And an improved estimate of the true population mean, $\overline{X}$, can be obtained as

$$\bar{x} = \bar{y} - \bar{b},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is obtained from the original sample and $\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$ from the independent sample. The variance of $\bar{x}$ which can be written as

$$\mathrm{Var}(\bar{x}) = Var(\bar{y}) + Var(\bar{b})$$

is estimated from the two samples as $\frac{s_y^2}{n} + \frac{s_b^2}{m}$ where $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Case 3: Record check using a subsample of the original sample

A simple random subsample of size n_r is drawn and measurements *(y$_i$)* obtained for the elements in the subsample. A record check *(x$_i$)* is determined for all the elements in the subsample of size n_r.

The estimate of the average bias, $\overline{B}$, is obtained from the subsample as

$$\bar{b} = \frac{1}{n_r}\sum_{i=1}^{n_r} b_i\ ,$$

where $b_i = y_i - x_i$ is the bias measured for the i^{th} element in the record check subsample.

The estimate for the variance of the estimated average bias, $V(\bar{b}) = \frac{S_B^2}{n_r}$, can be calculated from the subsample as $\frac{s_b^2}{n_r}$ where

$$s_b^2 = \frac{1}{n_r - 1}\sum_{i=1}^{n_r}(b_i - \bar{b})^2.$$

And an improved estimate of the true population mean, $\overline{X}$, can be obtained from the subsample record check values as

$$\bar{x} = \bar{y} - \bar{b},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is obtained from the original sample and $\bar{b} = \frac{1}{n_r}\sum_{i=1}^{n_r} b_i$ from the subsample.

The variance of $\bar{x}$ is

$$\text{Var}(\bar{x}) = \frac{S_X^2(1-\rho^2)}{n_r} + \frac{\rho^2 S_X^2}{n} - \frac{S_X^2}{N}$$

where $S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ and can be estimated from the subsample by

$$s_x^2 = \frac{1}{n_r - 1}\sum_{i=1}^{n_r}(x_i - \bar{x})^2$$

and ρ is the true correlation coefficient between X and Y, which can be estimated from the subsample as

$$\rho = \frac{Cov(y_i, x_i)}{s_x s_y}.$$

Also see the introduction to chapter 4 for the relationship between the correlation coefficient and the index of inconsistency (IOI).

Findings from NCES Record Check Studies

The two NCES studies described in this chapter, the RCG:91 Validity Evaluation and the SASS Teacher Transcript Study, are forward record checks. NPSAS has also conducted record check studies, although they are not discussed in this report. For example, the *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report* (Working Paper 96-17) describes the results of matching student reports of receiving a Pell Grant against the Department of Education Pell File, which includes one or more records for each Pell Grant recipient or awardee.

The RCG:91 Validity Evaluation and the SASS Teacher Transcript Study calculated measures of simple response variance (GDR) and response bias (NDR) as described in chapter 3.

Details of the findings from these NCES record check studies follow.

Recent College Graduates (RCG)

The 1991 Survey of Recent College Graduates (RCG:91) provided data on the occupational and educational outcomes of bachelor's degree and master's degree recipients one year after graduation. The reinterview study conducted on this data, described in chapter 3, tried to estimate response variance and bias. Brick, Cahalan et al. (1994) decided that their use of reconciled reinterview data to estimate response bias could be improved by a record check that compared certification data provided by graduates to data collected from state certification agencies.

Methodology and Design of the RCG:91 Validity Evaluation

The RCG:91 validity evaluation examined the number of graduates certified to teach and the kind of certification the graduates obtained. These were key variables for the RCG estimates related to the number of new teachers who had graduated from higher education institutions. The study used a two-stage design. In the first stage, 10 states were selected with probability proportionate to the number of sampled education majors who graduated within each state. In the second stage, a simple random sample was selected within each sampled state from graduates who reported they were certified to teach in that state and had been interviewed on August 1, 1991 or later. In all, 326 graduates were sampled.

The survey form used to collect certification data from the state agencies used the same question wording and response categories used for the sampled graduates. The survey questions included whether the graduate was certified (question 53), the kind of certification (question 56), the grades certified to teach (question 54), and the subjects certified to teach (question 59). All 10 states returned all of their survey forms for a 100 percent response rate (i.e., 326 out of 326).

Data from state agencies were assumed to be correct and unbiased and therefore had no response variance. All of the validity data were categorical. The analysis of the kind of certificate, the grades certified to teach, and the subjects certified to teach was presented in terms of gross difference rate (GDR) and net difference rate (NDR) (a detailed discussion of these statistics is presented in chapter 3).

Brick, Cahalan et al. (1994) studied only a subgroup of the population; that is, only those who reported being certified. Their validity evaluation design allowed them to estimate the ratio of those whom agencies confirmed to be certified to those who reported being certified; however, it was not possible to estimate the net effect of reporting errors on certification status because the responses were not validated for any respondents who said on the survey they were not certified.

Summary of Findings

Certification to Teach. The first item examined was whether the graduate-reported certification was confirmed by the state certification agency. Overall, 94.5 percent of the graduates in the validity study sample had their certification confirmed. Of the 10 states in the study, 5 states confirmed 100 percent of their sampled graduates as certified. Of the remaining 5 states, the confirmation rate varied from 80 to 96 percent.

Graduates could have confused being eligible with being certified, or graduates could have applied for certification but not yet received it. The lower confirmation rate for graduates who reported certification before 1989 may have been partially caused by graduates who were certified at one time but had not renewed their certification. An emergency or temporary certificate may have been held at the time of the interview but not at confirmation. Finally, differences could also have been due to interpretations of what should be included as alternative, emergency, or temporary certification.

Brick, Cahalan et al. (1994, p. 5-12) speculated that simply asking states to search a second time would result in more matches and higher confirmation rates. For the most part, the assumption that the external source is error free is very weak in actual implementation.

Kind of Certification. Both graduates and states were asked to choose one of the following categories for kind of certification: initial or provisional certificate leading to regular or standard certificate; regular or standard; alternative, emergency, or temporary; and other (specify).

The overall GDR (i.e., an aggregate over all 10 states) was 42.5. In general, the state GDRs were relatively large (median GDR = 45.2). The most common situation where the graduate and state reports did not agree was when the graduate reported having an "initial or provisional" certificate but the state reported the graduate as having a "regular or standard" certificate (this situation was the source of difference in 24 percent of all the cases, and in 56 percent of the cases reported differently by graduate and state). The difference may be partially due to the different time periods in which the data were collected and to different interpretations of the reporting categories.

Certification Grades. Graduates and agencies were asked to report all grades the graduate was certified to teach—prekindergarten, each grade from kindergarten through twelfth, ungraded (captures special education), all grades—and subject certified. The GDR for all sampled states for all grades combined was 8.9. The GDRs for the different grades ranged from 4.9 to 17.0. The GDRs were highest in the transitional grades between elementary, middle school, junior high, and high school. This was probably related to grades being grouped differently across the states. For example, elementary certification was sometimes reported through sixth grade and sometimes through eighth grade.

Certification Subject Fields. The last RCG:91 question included within the validity evaluation study was subject fields graduates were certified to teach. There were 35 subject fields: "any elementary field, general, or specialized;" 25 specialized fields (e.g., business, mathematics, social science/social studies, etc.); 7 categories of special education; vocational education (other than business, home economics, or industrial arts); and other. For one analysis, responses were coded so that only real differences in certification would appear in the difference rates. For the coded responses, the GDRs and NDRs were almost all below 4 percent, indicating that there was close agreement between the states and the graduates on certification fields. Since almost all the NDRs were positive, most of the differences were due to graduates claiming certification in more fields than state records showed.

Difference rates were also calculated using the uncoded responses. NDRs were relatively large (-1.6 to 29.5) and almost always positive, which again showed the tendency of graduates to overstate the subjects they were certified to teach. GDRs ranged from 1.6 to 29.5. Since GDRs were only slightly larger than the NDRs, and the GDR and NDR were the same in 15 cases, Brick, Cahalan et al. (1994) concluded that the measurement errors were primarily from the response bias (p. 5-18).

Implications and Recommendations

The findings showed that the response bias due to overreporting being certified to teach was less than 5.5 percent. Since errors in matching the graduate at the state level may have overstated response bias, the net response bias could be less than 5 percent. Based on those results, estimates should be considered fairly accurate for most purposes.

The one area of particular concern was where the response biases for the subjects certified to teach were relatively large. Data users should consider producing estimates that avoid the overestimates to the extent possible instead of using coarse adjustments. Brick, Cahalan et al. (1994) suggest, as an example, that the estimates of subjects certified to teach could be restricted to the subset of graduates who reported they were not certified in the category "any elementary fields," since the problem is largely associated with the elementary grade teachers. At a minimum, estimates of the subjects that have the largest estimated response biases should be noted in any analysis and the reasons for the overestimation should be discussed.

Schools and Staffing Survey (SASS)

A SASS record check study compared teachers' self-reports of their academic qualifications, as provided on SASS:90-91 survey questionnaires, with the use of data from teachers' college transcripts (Chaney, 1994). Reliability was measured in terms of the proportion of cases that had data which agreed between the two interviews. Thus, the "level of agreement" equaled the number of "correct" responses divided by the number responding to the item, multiplied by 100. The major results are summarized below.

Methodology and Design of the 1990-91 SASS Teacher Transcript Study

The 1990-91 SASS Teacher Transcript Study was designed to determine the best method for obtaining teachers' background information—using teachers' self-reports of their academic qualifications, as provided on the survey questionnaires, or teachers' college transcripts. It was assumed the transcripts would provide the most accurate account because they were neither subject to reporting bias nor dependent on teachers' recall.

The transcript study used a multi-stage design. In the first stage, 200 schools were chosen in 10 states: 50 public elementary schools, 50 public secondary schools, 50 private elementary schools, and 50 private secondary schools. A total of 174 schools were determined to be eligible for the study and agreed to participate. A sample of 867 teachers was next selected, with no more than 5 teachers from any one school. However, 32 of the selected teachers were later determined to be out-of-scope, leaving 835 eligible teachers. A total of 637 interviews were completed, for a final response rate of 76 percent. An additional 45 teachers either refused to participate in the transcript portion of the study or failed to supply any information on which college they attended. Any teacher who refused to participate was left out of the transcript study, leaving 592 teachers. (See Chaney, 1994, p. 2 for a full description of reasons why teachers may have refused to participate.)

After requesting teachers to provide a list of all colleges attended, a total of 1,985 separate transcripts were identified. A total 150 transcripts were unavailable due to teacher refusal (n=130), attendance at a foreign institution (n=14), and unlocatable school (n=6). A total of 1,524 transcripts were received, including 1,356 transcripts among those that were originally requested, and an additional 168 transcripts reflecting an undergraduate or graduate enrollment that had not been indicated on the SASS questionnaire. Additionally, 134 schools responded in ways other than sending a transcript, including a teacher not attending school (n=53) and inability to locate teacher transcripts (n=81). Thus, the total number of school responses was 1,658 out of 2,003 identifiable transcript requests (1,835 original requests, plus 168 transcripts received that had not been anticipated), or 83 percent. (See Chaney, 1994, p. 3 for a full description of response rate calculations.)

Summary of Findings

Teacher's Self-reports of the Schools They Attended. The SASS questionnaire asked teachers to list every college or university they had attended, whether or not they obtained a degree at the college. This question was asked to facilitate the collection of college transcripts rather than to verify teachers' accuracy in reporting; nevertheless, the teachers' responses could be examined for accuracy. When the teacher self-reports were compared to the school transcripts, there were 53 cases (from 44 teachers, or 8 percent of the total number of teachers for who at least one transcript was collected) in which a teacher reported attending a college, but the college stated the teacher never attended. In seven of these cases, the college erred, while 46 cases (from 38 teachers, or 7 percent) might indicate false reports by the teachers. There were an additional 81 cases (from 67 teachers, or 12 percent) in which colleges were unable to locate teachers' transcripts; 75 of those cases (from 61 teachers, or 11 percent) might represent false reports, possibly due to differences in definitions of college attendance.

Teacher's Self-reports on the Degrees They Earned. Teachers were asked if they had obtained five different types of degrees: bachelor's, master's, associate's, education specialist or professional diploma, and a doctoral or first professional degree. They were then asked to state their major(s) and the year the degree was received.

Teachers' self-reports on the type of degrees they had earned were most accurate for bachelor's degrees; there were almost no cases of incorrect data. The self-reports could not be confirmed for only 22 respondents. For 80 percent of respondents, the reports were confirmed directly through the college transcripts. For another 14 percent, reports of a bachelor's degree could only be inferred from transcript showing a higher degree or a transcript showing graduate level work. A greater number of errors was found in teachers' reports on master's degrees, somewhere between 3 and 12 percent. Two percent of teachers failed to indicate they received a master's degree. An additional 11 percent could not have their degrees confirmed. These two types of problems also occurred when teachers reported associate degrees. Four percent failed to report an earned associate degree, while 3 percent failed to have a self-reported degree confirmed. Five teachers reported receiving doctoral degrees; of those, four degrees were confirmed, while only partial transcript data were available on the fifth.

Teachers' self-reports on the year their degree was earned were more subject to error than their reports on the degrees themselves. The proportion of errors ranged from 12 to 32 percent. Among teachers with identified errors, teachers' reports were off by one year. Eighty-eight percent of teachers for whom the year of receiving a bachelor's degree was available reported the same graduation year as in the transcript, while 12 percent made a reporting error. The discrepancies were relatively evenly split between teachers who made an error of 1 year and those who were off by more than 1 year, and between teachers who stated a year that was too recent and teachers who stated a year that was too early. A greater proportion of errors occurred for master's and associate degrees, though most were off by only 1 year.

Chaney suggested some legitimate reasons for errors in reporting the year that a degree was earned. The official award of a degree may be delayed until the next scheduled graduation ceremony, even though all requirements for a degree may have been met the previous year. A degree may also be awarded conditionally so that a student may participate in a graduation ceremony with his/her peers, even though some requirements must be met before the degree is actually awarded. Chaney observed there exists a higher error rate associated with teachers with a master's and associate degrees than for bachelor's degrees. He hypothesized that an associate's degree was not as salient for the teachers because the teacher may have received a higher degree at a later date.

Teachers were asked to write the major field of study for each degree they had earned, using two-digit codes provided on the SASS questionnaire. For 65 percent of those teachers earning bachelor's degrees whose self-reports could be compared with the transcript record of their majors, the subject matter was correctly coded. Another 10 percent showed discrepancies only in whether the subject area was listed as a separate discipline or as an area of education. In about 12 percent of the cases, there was a discrepancy when teachers and transcripts both reported majors within education, but within different specialties. Most of these problems might be classified as coding errors or differences in interpretation.

Teacher's Reports on the Courses They Took. The SASS questionnaire asked teachers about other courses they had taken in teacher education, their main teaching assignment, and their second teaching assignment. Teachers were also asked whether the courses were taken using a semester system, a quarter system, or both.

A sample of roughly one-half of the teachers (280 teachers with undergraduate degrees and 250 teachers with graduate degrees) were asked the number of undergraduate and graduate courses they had taken in each field, while the other half (228 with undergraduate degrees and 196 with graduate degrees) were asked the number of credits. Two-thirds of the teachers gave responses that matched their transcripts at the undergraduate level. The highest rate of accuracy was in the category *4 or more courses*, with 81 percent giving responses that could be directly confirmed. Only 3 percent gave responses that were contradicted by the transcripts. Sixteen percent could not have their responses confirmed or contradicted. Teachers reporting they had taken *no courses* in teacher education also showed a high accuracy level. Among the other two categories (i.e., from one to three courses reported by the teacher), a majority of the teachers understated the number of teacher education courses they had taken. Several reasons may explain the high rate of errors, including the questionnaire's requiring an unusually high level of precision, teachers having difficulty remembering those areas where they took only a small number of courses, and teachers failing to follow the instruction to include courses such as mathematics education within the teacher education category.

Roughly the same pattern of responses was found in teachers' reports of the number of credit hours taken, and in their reports on graduate courses in education. However, one difference was teachers were less likely to underestimate the number of graduate teacher credit hours taken, and more likely to overestimate the amount.

Broader categories were used for collecting data on courses and credits in teacher main teaching assignment. They corresponded roughly to requirements for majors and minors. Only 53 percent of the responses could be directly confirmed at the undergraduate course level: 35 percent of teachers overestimated the number of courses, while 11 percent gave an underestimate. Among graduate teachers, 42 percent of responses were confirmed, 55 percent were shown to be overestimates, and 3 percent were underestimates. The results for teachers' self-reports on credit hours were not substantially different from those on the number of undergraduate courses taken. However, teachers were somewhat less accurate in their counts of graduate courses, with more overestimates and fewer underestimates.

Few teachers reported a second teaching assignment. Only 30 teachers with undergraduate degrees and 22 teachers with graduate degrees were asked the number of courses and credit hours they earned in their second teaching assignment and 29 teachers with undergraduate degrees and 19 with graduate degrees were asked about the number of credits hours. Of those, 37 percent of the teachers with undergraduate degrees gave responses which were directly confirmed, while 53 percent gave overestimates. The same pattern occurred among teachers with graduate degrees. An even higher error rate was found among those reporting the number of undergraduate credits earned, with only 14 percent being directly confirmed, and 45 percent shown to have overestimated the number of courses.

Teachers were asked to report whether the courses were taken using the semester system, quarter system, or both. Ninety-three percent of teachers who reported that all courses were within the semester system were correct. However, teachers who reported all courses were within the quarter system were about equally likely to be either correct (53 percent) or incorrect (47 percent). Teachers who reported they had taken courses within both semester systems <u>and</u> quarter systems showed the lowest rate of confirmation (44 percent). It is important to note, however, that 32 percent of the teachers in this category had only partial data available, and in this case the availability of all transcripts might have confirmed that the teachers were correct. The high level of error in teachers' self-reports, especially those attending quarter systems or both systems, might be explained by teachers failing to consider all schools that were attended.

Those teachers who taught at least one course in science and mathematics were asked to state the exact total number of courses taken in one of seven disciplines. The proportion of teachers who were able to correctly state the exact number of courses ranged from 30 percent (in mathematics) to 71 percent (in physics). However, in many cases, teachers were able to correctly state the exact number because they had taken no courses at all within the discipline. If these zeroes were excluded, the proportion giving correct answers was much lower, ranging from 8 percent (in other natural sciences) to 44 percent (in chemistry). The general tendency was for teachers to overstate the number of courses they had taken. The largest difference occurred in mathematics, with teachers' self-reports showing an average of 6.5 undergraduate courses, while the transcripts showed an average of 5.7. Teachers' reports on the graduate courses they had taken showed a similar pattern. However, they often were able to state the exact number of courses they had taken because graduate education is typically more specialized and they had taken no courses in the discipline. If the zeroes were excluded, teachers were actually less accurate in reporting on graduate courses than in reporting on undergraduate courses.

Summary

Several different types of teacher errors were detected when examining the SASS questionnaires and comparing them with the transcripts, including: 1) item nonresponse; 2) errors of omission; 3) bias on the part of the respondent; 4) telescoping, that is, reporting events as having happened more recently than was actually the case; and 5) differences in judgment. In this study, item nonresponse refers primarily to teachers' ability or willingness to provide detailed course data. Errors of omission included such matters as respondents failing to list all college they had attended. Chaney speculates that errors in identifying term types were also likely errors of omission, in which respondents only described the term type for one or two colleges attended, and failed to allow for courses taken at other colleges. Errors that appeared to show bias on the part of respondents included the general pattern for teachers to overstate their preparation in their second teaching assignment and in mathematics and science as compared with the records on their transcripts. Telescoping was not found to be a major factor in this study. The errors in reporting the year a degree was earned were roughly evenly split between reports that were too recent and those that were too early.

The final category, "differences in judgment," reflects the fact that, in a transcript study, while it is possible to take many actions to help assure uniformity in how teachers' records are compared, such as including special advance training of the transcript coders, providing a dictionary of courses with appropriate codes, having supervisors monitor the coding, and running computer checks of the analysis file for consistency across all transcripts, the respondents are on their own. Chaney concludes that "it should not be surprising that different respondents will answer a questionnaire differently, or that their judgments will sometimes disagree with those of a trained coder" (1994, p. 20).

Chaney (1994) notes that there are inherent problems with using transcripts as the source by which teacher reports are compared. For example, transcripts may be illegible, use unknown abbreviations, contain errors or inconsistencies, be inconsistent from one institution to another, and vary in terms of the way failures, withdrawals, remedial/not-for-credit courses, incompletes, etc. are treated. Thus, collecting data by either teacher self-reports or transcript data has advantages and disadvantages.

These findings suggest the type of data collected, the resources available, and considerations such as the amount of burden to be placed on the survey respondent will determine the kind of data to collect. Chaney recommends the following general guidelines for record check studies of teacher certification and training. Teacher self-reports may be more accurate when relatively large categories are used, identifying that they had not taken any courses in a field, and describing their backgrounds in teacher education and their primary teaching assignment. Transcript data may be more accurate when detailed or complicated information is required, such as second teaching assignment and semester versus quarter systems.

CHAPTER 7
Cognitive Studies

This chapter presents a summary of results from some of the exploratory cognitive research conducted by NCES. The scope of these qualitative studies is much smaller in scale than the reinterview, "multiple indicators," and record check studies discussed earlier in this report. Nonetheless, NCES has made efforts to probe respondents to gain in-depth information about newly developed questions and questions about concepts which are difficult to measure. Cognitive research methods are particularly appropriate for trying to understand how respondents interpret instructions and questions, recall information, and respond to word and question order (Nolin and Chandler, 1996, p. 2). Another issue arises from NCES use of structured interviews and self-administered questionnaires which follow a prescribed sequence of actions (a protocol) to ensure that all respondents are exposed to the same measurement process. While a protocol maximizes the probability that all respondents are asked exactly the same questions, it may make the interviewer-respondent interaction awkward since protocols do not allow the interviewer and respondent to follow normal rules of conversation (Brick, Tubbs et al., 1997, p. 3-1).

Cognitive Research Methodology

Current methods for conducting cognitive laboratory research are discussed in chapter 1 of this report, but the main points of that discussion will be summarized here. Cognitive research draws on three different literatures: research in cognitive psychology on memory and judgment, research in social psychology on influences against accurate reporting, and evidence from survey methodology research regarding response errors in surveys. Researchers generally agree on five stages of action relevant to survey measurement error: 1) encoding of information; 2) comprehension; 3) retrieval; 4) judgment of appropriate answer; and 5) communication. Beyond acceptance of these five stages, cognitive research takes different paths.

Forsyth and Lessler (1991) concluded that no guidelines were available for choosing one cognitive research method over another, due at least in part to a lack of theoretical and empirical work exploring how methodological details can affect cognitive laboratory results. Nonetheless, they offer a summary of four general sets of methods that have been implemented. These four methods are expert evaluation, expanded interviews, targeted methods, and group methods.

Expert evaluation methods involve no interaction with respondents. For example, experts may watch or listen to tapes of interviewer/respondent interactions and code behaviors, or they may analyze the survey form. *Expanded interview methods* refer to interviews where the survey questions are accompanied by probes about how the respondents perceive the survey items and how they decide to answer them. Probe questions are used to focus respondents' attention on particular aspects of the questions or on the whole question-answering process.

Targeted methods use survey items as stimulus material for other tasks. *Group methods* bring several people together to discuss topics of interest or to complete experimental versions of a questionnaire in a controlled setting. One of the reasons group formats are important is the social factors that distinguish group tasks from other laboratory tasks. Focus groups are probably the best known format of group interviews.

In summary, all of these methods provide more information about the question-answering process than can be obtained through simply asking the survey questions and recording the answers. (For more details on these methods, see table 1 and the pages that follow it.) The methods differ according to their timing and the amount of control the researcher has over what is observed. The task timing may be either concurrent, immediately after the respondent answers the questions, delayed, or unrelated. Either the respondent decides what information will be observed, as in concurrent think-aloud interviews, or response data are independently processed by the researcher as in behavior coding. All cognitive laboratory methods are basically qualitative studies even though some of the methods do collect quantitative information.

The methods NCES has used include expert evaluation, expanded interviews, and targeted methods. For example, NCES analyzed interviewer/respondent behavior through interviews recorded during the NHES:93 reinterview and used concurrent think-aloud techniques to learn how respondents understood the questions and instructions in the Teacher Listing Form.

The sample size for the cognitive studies was typically small. While NCES wanted to gain as much information as possible during these studies, it was concerned about respondent burden. BPS 90/92 included 62 respondents in its cognitive research study. NHES completed between 25 and 45 interviews for each of its components, and SASS included between 19 and 100 respondents per component depending on the type of cognitive research undertaken.

Cognitive Research Results

A review of the following research is included in this chapter

- **Beginning Postsecondary Students (BPS) Longitudinal Survey**
  BPS 90/92 Field Test — Techniques such as followup probing questions and memory cueing methods used to evaluate field test design

- **National Household Education Survey (NHES)**
  NHES:93 — Behavioral coding methodology used to evaluate a small sample of recorded interviews from the School Readiness and School Safety and Discipline components

- **Schools and Staffing Survey (SASS)**

Teacher List Validation Study and Teacher Listing Form Study	Concurrent think-aloud techniques used to evaluate quality of the Teacher Listing Record
Public School 1991-92 Field Test Questionnaire	Concurrent think-aloud techniques, paraphrasing, and retrospective interviewing used to gain in-depth information on new items

This is not an exhaustive review of the cognitive research NCES has conducted, although it should be indicative of its qualitative research efforts. For example, NCES conducted cognitive interviews and focus groups to test the initial and revised versions of the NHES:91 questionnaires. NCES also taped a small number of the extended interviews during the actual NHES:91 data collection to do a systematic content analysis as a way of evaluation respondent comprehension. Cognitive interviews and focus groups were also held to test the NHES:93 questionnaires. (See Nolin and Chandler, 1996.) Cognitive research conducted on SASS include the Jenkins (1993) and Jenkins and Ciochetto (1993) work with the Student Records Questionnaire, the Jenkins (1997) work with the Public and Private School Teacher Questionnaire, and cognitive interviews conducted with private school administrators and with private school teachers (DeMaio, 1990a and b).

Beginning Postsecondary Students (BPS) Longitudinal Survey

The Beginning Postsecondary Students (BPS) Longitudinal Survey provides national data concerning issues in access, choice, enrollment, persistence, progress, curriculum, and attainment in postsecondary education, graduate/professional school, and rates of return to society. It was initiated with a cohort of individuals beginning their postsecondary studies in the 1988-89 school year, regardless of when they completed high school, making information about "nontraditional" students available at the national level for the first time. Nontraditional students delayed continuing their education because of military service, family responsibilities, or other reasons (Burkheimer et al., 1992, pp. I-1 to I-2, II-1).

The field test design included reliability and validation reinterviews to identify potential sources of response error and to develop survey methods for reducing or eliminating errors resulting from difficulties understanding the questions or respondent recall. The reliability reinterviews paralleled other NCES reinterview studies and were discussed in chapter 4. The validation reinterviews, however, included questions based on cognitive laboratory research applied to telephone interviewing. Thus, despite the name of the study, the results of the validation reinterview are discussed here.

Methodology and Design

The validation reinterview included items from three major sections of the survey: educational experiences, education financing, and work experiences. These three areas were selected, in part, because they included particularly difficult comprehension and recall tasks. Research (Lessler et al., 1989; Cannell et al., 1989) has indicated that difficulties in comprehending questions, recalling requested information, and selecting appropriate response categories could have large effects on response and estimation accuracy. Thus, the validation reinterview was designed to gather information about

- How respondents interpreted the questions
- How accurately respondents recalled the requested information
- Whether respondents had difficulty selecting descriptive response categories

The validation reinterview included cognitive research techniques such as followup probing questions and memory cueing methods. Forsyth and Lessler refer to followup probing as a variation of the think-aloud method.

> *Interviews for specific surveys are conducted under similar instructions; however, a researcher who makes followup probes has identified a set of focal issues based on analyses of the question-response task. For example, a researcher may be interested in how respondents interpret technical terms, how they make choices among provided response alternatives, or what their approaches are to memory retrieval when questions cover long recall periods. If general think-aloud responses do not address the pre-identified issues, then an interviewer can ask specific probe questions that do* (1991, pp. 398-99).

Forsyth and Lessler describe memory cue tasks as a general type of expanded interview. They are used to "assess recall errors due to a respondent's failure to remember events during an interview. Memory cue tasks have been used to assess the potential for reducing recall error by providing cues during the survey interview" (1991, p. 399).

Seventy-five BPS respondents were randomly selected to participate in the study. The interviews were conducted during the last 4-1/2 weeks of CATI operations. Sixty-two respondents (a response rate of 82.7 percent) completed the validation reinterview (Burkheimer et al., 1992, p. VI-12).

Summary of Results

Questions on educational experiences were specifically included to investigate student education goals, and to learn the calendar and credit systems with which the respondents were most familiar. The BPS item asking about student classification was intended to cover alternative year-based systems by offering a classification *or* year in each response category (Burkheimer et al., 1992, p. VI-13).

Education experiences

- B7f How were you classified by (fill in school name) during this term (fill in dates)? (Read choices first time through; subsequently, read as necessary.)

  (1) First-year or freshman
  (2) Second year or sophomore
  (3) Third year or junior
  (4) Senior
  (5) Special student (nonmatriculated)
  (6) Other (specify) ________________________

Responses to the validation followup questions suggested that response alternatives should be more explicit. As a result, Burkheimer et al. (1992, p. VI-13) suggested replacing the "or" structure with explicit specifications such as "Freshman (first year student)."

Questions on educational financing were included in the validation reinterview study because of concerns about the accuracy of respondents' reports. The followup questions were designed to obtain information on respondents' perceptions of response accuracy and to determine whether accuracy could be improved through the use of memory cueing methods. Respondents were first asked to estimate the amount spent on tuition and fees, and the amount spent on educational expenses for each term of enrollment. When asked to rate the accuracy of their answers, they reported relatively low confidence in their estimates. For example, after the respondent gave an estimate of tuition and fees, the interviewer cued the respondent by asking if the institution charges student activity fees, laboratory fees, athletic fees, supplies fees, etc. Interviewers then asked if respondents wished to revise their estimate of tuition and fees. Eleven percent of respondents revised their estimates of C1, tuition and fees, and 13 percent revised their estimates of C2, educational expenses. The authors concluded, "the magnitudes of the revisions remained relatively small and did not exceed two standard errors. In summary, the memory cue method was not particularly effective in enhancing response accuracy" (Burkheimer et al., 1992, p. VI-18).

Education financing

- C1 For (fill in name of first, second, third, etc., school/college in which enrolled during or after February 1989), let's talk about the term from (starting and ending dates of first enrollment for credit beginning with the first term that includes or follows February 1989). About how much were the tuition and fees before any financial aid or waiver? $________
- C2 How much were the expenses for books and supplies, room and board, and other related additional expenses? $________

Burkheimer et al. (1994) also selected a set of educational experience items (B7i through B7o) which asked respondents for information about education goals. These items were selected because results from a pretest suggested that some respondents interpreted the questions as addressing relatively long-term goals (for example, master's or Ph.D. degree) instead of the short-term goals (for example, bachelor's degree) that researchers expected. Field test validation reinterview results confirmed that this situation is a potential source of measurement error. For example, 15 of the respondents who reported working toward a degree or special award (42.9 percent) indicated that coursework at their institution was insufficient. About two-thirds of those respondents were "working toward" a master's or Ph.D. degree. Thus, Burkheimer et al. (1994, pp. VI-13 and VI-14) suggested revising question wording to make the time frame more explicit.

NCES selected one question in the work experiences section of the survey because it was unsure respondents would understand the response categories. If respondents reported a time period in which they were simultaneously employed and enrolled in school, then they were asked to select a response category that characterized their role at the time.

- D21 (If employed since February 1989 but prior to end of last enrollment period, ask this question; otherwise go to D22) During the time when you were both working and enrolled in school/college, how did you view your primary role in postsecondary education? (Read all choices.)

  1 = Student who works to help pay expenses while in school/college.
  2 = Student who works to earn extra spending money while in school/college.
  3 = An employee who attends school/college to gain skills necessary for job advancement.
  4 = An employee who attends school to expand new career possibilities.
  5 = An employee who attends school to expand personal knowledge/skills.
  6 = Other (specify).______________________________

Only five respondents (12 percent) reported that the response categories were not completely adequate. Open-ended reports suggested that the source of difficulty was not in the student/ employee descriptions, but in the "purpose" descriptors. Specifically, interpreting the difference between working to pay educational expenses and working to earn extra spending money. It was difficult for respondents to distinguish between these two purposes for working while enrolled in school. While the question was successful in its original intent to distinguish between students and employees, minor rewording was suggested to reduce potential measurement error (Burkheimer et al., 1992, p. VI-18).

The validation reinterview also included items asking for detailed information about jobs. Burkheimer et al. (1992) used probe questions to ask respondents to rate confidence in their responses to the starting and ending dates. Eighty-three percent of the respondents judged their responses "very accurate," the remaining 17 percent judged their responses "somewhat accurate." These results support the expectation that respondents have relatively little difficulty recalling information about non-recurring events.

National Household Education Survey (NHES)

The National Household Education Survey (NHES) collects education data from U.S. households using random digit dialing (RDD) and computer-assisted telephone interviewing (CATI) procedures. This design gives NCES added flexibility to evaluate interviews through recording. NCES evaluated a small sample of recorded interviews from NHES:93 using behavioral coding methodology adapted from Oksenberg, Cannell and Kalton (1991). The behavioral coding methodology was applied to tape recordings of 25 School Readiness (SR) and 45 School Safety & Discipline (SS&D) interviews. Of the 45 interviews recorded from SS&D, 25 were parent interviews and 20 were youth interviews (Brick, Tubbs et al., 1997, p. 3-9).

Methodology and Design

Behavioral coding methodology was developed to pre-test and evaluate structured questionnaires. It is based on assessment of interviewer and respondent behavior patterns, providing systematic data on deviations from protocols and the extent to which the respondent provides data as expected. Interviewer behavior was coded using five categories

- Read the question exactly as worded
- Read the question with a minor wording change
- Read the question with a major wording change
- Clarified the question for the respondent
- Displayed some affect

Minor changes in question wording included insertion or omission of particular words that did not, in the opinion of the coders, alter the meaning of the question. Major changes in question wording were those that could change the meaning of the question, as, for example, by omitting whole parts of the question. The last category, "displayed some affect," was added to the Oksenberg, Cannell, and Kalton (1991) scheme for the NHES:93 study "to try to pick up whether particular questions, especially ones that cover sensitive material, were difficult for the interviewer to administer in a neutral manner" (Brick, Tubbs et al., 1997, p. 3-3).

Respondent behavior was coded using six categories

- Gave a "correct" response
- Interrupted the interviewer before completing the question
- Clarified the question
- Qualified the answer with respect to accuracy
- Did not provide an adequate answer
- Expressed sensitivity to the question

Note that a "correct" response means the respondent used one of the precoded answer categories for a question. "Correct" in this instance is not an indication of the validity of the response.

Since multiple interactions between the interviewer and the respondent could occur for each questionnaire item, all appropriate categories for each questionnaire item were recorded. "For example, the interviewer may have made minor changes [Minor] to the question wording and also provided clarification [Clarify] for the question. Similarly, the respondent may have asked the interviewer for clarification [Clarify] about the question, but ultimately provided the correct [Correct] response to the question" (Brick, Tubbs et al., 1997, p. 3-5).

Summary of Results

Table 60 shows the frequency and percentage of each code for both interviewer and respondent, by interview. For interviewers, the behavior categories, "read the question exactly as worded," and "read the question with a minor wording change," were coded most often (85 to 89 percent), followed by "clarified the question for the respondent" (9 to 11 percent). Respondents provided codable responses for the majority of the questions on all three forms. Ten percent of the respondent codes for the SS&D parent interviews were "qualified the answer with respect to accuracy."

Table 60. -- NHES:93 codes for interviewer and respondent behavior, by form

Behavior code	School Readiness (N=25)		School Safety & Discipline: Parent (N=25)		School Safety & Discipline: Youth (N=20)	
	Number	Percent	Number	Percent	Number	Percent
Interviewer						
Exact	2,654	70.0	2,095	62.3	1,231	65.6
Minor	658	17.4	772	23.0	440	23.4
Major	21	0.6	51	1.5	7	0.4
Clarify	371	9.8	382	11.4	171	9.1
Affect	88	2.3	63	1.9	28	1.5
All codes	3,792	100.0	3,363	100.0	1,877	100.0
Respondent						
Correct	2,738	86.9	2,402	79.7	1,411	87.5
Interrupt	66	2.1	90	3.0	6	0.4
Clarify	126	4.0	118	3.9	41	2.5
Qualify	140	4.4	301	10.0	93	5.8
Not adequate	70	2.2	93	3.1	62	3.8
Sensitive	12	0.4	8	0.3	0	0.0
All codes	3,152	100.0	3,012	100.0	1,613	100.0

Note: Percentages do not always equal 100 due to rounding.

SOURCE: Derived from table 3-8, Brick, Tubbs et al., (1997), *Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)*, Working Paper 97-02, p. 3-10.

Only one question from SR received the "major" rating more than once. This question, R93, accounted for 4 of the 21 behavioral codes.

- R93 How about on Saturday and Sunday? How many hours does (child) watch television or video tapes at home on...
  a. Saturday?
  b. Sunday?

Coder comments indicated the interviewer skipped the introductory sentence for this item all four times. Brick, Tubbs et al. note that "this may reflect the fact that the introduction is redundant with the answer categories"; however, they also point out that "this question is also embedded within a sequence of items where the interviewer needed to clarify the question(s) and where the respondent frequently qualified the answer" (1997, p. 3-13).

In the 25 School Safety & Discipline parent interviews, interviewers "read the question with a major wording change" 51 times. While this category was only a small percentage of all codes (1.5 percent), there were six questions that received this rating from 3 to 5 times. These questions were

- P2 Is (child)
  White
  Black

American Indian or Alaskan Native
Asian or Pacific Islander, or
Another Race?
What is that? __________

- P9 Does (child's) father live in the household or does (he/she) have a stepfather or foster father who lives in the household? (father, stepfather, foster father, no)
- P9a What is his name?
- PY29 (Have you heard/Do you know) of money or other things being taken directly from students or teachers by force or threat of force at school or on the way to or from school this school year? (yes/no)
- PY34 (Have you heard/Do you know) of any incidents of bullying during this school year? For example, do some students pick on others a lot or can they make other students do things like give them money? (yes/no)
- PY94 [Do you/Do your parents/Does your (mother/stepmother/foster mother/father/stepfather/foster father)/Does (adult respondent)] think it is all right for [(child)/you] to drink alcoholic beverages, for example, beer, wine coolers, or liquor? A small amount on special family occasion or for religious purposes does not count. (yes/no)

Coders' reasons for the "major" ratings included pausing inappropriately, omitting either all or part of an example, and omitting either all or part of a qualifying phrase (Brick, Tubbs et al., 1997, pp. 3-14 to 3-15).

In the 20 SS&D youth interviews, interviewers "read the question with a major wording change" only seven times. When interviewers were classified as exhibiting behavioral codes other than "exact" or "minor," it was usually the same questions which elicited the behavior from the parent interviews (Brick, Tubbs et al., 1997, p. 3-17).

Brick, Tubbs et al. concluded that "the results of this analysis indicate that the majority of questions in the three questionnaires were read as written by the interviewer (or with only minor revision) and the respondents provided a 'codeable' response" (1997, p. 3-17). However, interviewers did make a high number of minor changes to the introductory items for each section of the questionnaire. These changes could reflect interviewers' need to adapt the introduction to what the respondent has said, or to interviewers not understanding the importance of reading items exactly (1997, p. 3-13).

Another issue was the use of Likert scales [e.g., strongly agree, agree, disagree, strongly disagree (PY21a-6); very important, somewhat important, not too important, not at all important (PY22-23)] in telephone interviews. Questions interviewers had to clarify commonly included a Likert scale.

Schools and Staffing Survey (SASS)

The Schools and Staffing Survey (SASS) is an integrated set of surveys designed to obtain national level data on students, teachers, and administrators. SASS has conducted several cognitive research studies using a variety of the specific methods described in chapter 1. Among the methods SASS has implemented are concurrent think-aloud interviews, followup probes, retrospective think-alouds and probe questions, and paraphrasing. Summaries for three of these studies are included in this chapter.

The 1993 Teacher List Validation Study (TLVS) was not the first cognitive research SASS conducted, however it was broader in scope than anything else that had been undertaken. When coupled with its followup companion, cognitive research on the Teacher Listing Form (TLF), it is indicative of the types of cognitive research methods NCES uses to gather more detailed information about the operational procedures it uses for conducting large scale surveys. These techniques are also used to identify potential sources of measurement error, including measurement error resulting from poorly worded questions and measurement error resulting from the operational approach for measuring a particular concept. Finally, a review of *Results of Cognitive Research on the Public School 1991-92 Field Test Questionnaire* (Jenkins, Ciochetto and Davis, 1992) provides examples of the kinds of information that have been learned from the in-depth or extended interviews that SASS routinely conducts when developing new questions for any its questionnaires or when items have been identified as "potentially problematic."

How to Count Teachers, Part 1: Teacher List Validation Study (TLVS)

The Schools and Staffing Survey (SASS) uses the Teacher Listing Record (TLR) to obtain a list of the teachers in each school. The U.S. Bureau of the Census then selects a sample of teachers from each school for the Teacher Survey portion of SASS.

Methodology and Design

The Teacher List Validation Study (TLVS) was conducted to evaluate the quality of the TLR. The goals were to attempt to find out the types of teachers/nonteachers that the schools included or excluded in their counts, and to find out reasons why the schools excluded certain teachers and included persons that should not have been included. Because of the way the samples were selected, statistical testing on the results would have been inappropriate. Although some counts were greater than others, there was no statistical evidence to confirm the results.

The study had two components. The first component, Reinterview and Reconciliation of the TLR, compared the teacher count computed by a school using the TLR to the teacher count computed by the LEA for that school, or, for private schools, to the teacher count reported in the 1991-92 Private Schools Survey. The objectives of the first component were to determine (1) if the schools filled out the TLR according to the instructions; (2) if the schools listed eligible in-scope teachers; and (3) if the school districts (Local Education Agencies, or LEAs), could provide more accurate listings of teachers than the public schools. The primary objective of the second component, TLR versus the School Questionnaire, was to determine if the TLR or the school questionnaire produced a more accurate count of teachers in the school. Both components shared a general objective of finding out if certain types of teachers/nonteachers were systematically missed, or if they were included incorrectly.

Component 1: Reinterview and Reconciliation of the TLR. The sample was chosen after approximately 85 percent of the TLRs had been received. The sample consisted of 290 public schools, 254 LEAs, and 300 private schools. The 200 schools (100 public and 100 private) with the largest percent differences between the counts were selected for the TLVS. The reinterviews took place from mid-February through the end of March, 1993.

For the public schools, 50 reinterviews were conducted in person and 50 by telephone. The list was ranked according to highest difference count; every other school was reinterviewed by telephone. For the private schools, 50 were selected for in-person visits and 50 were selected for telephone interviews. The cases were ranked by largest to smallest schools (not by difference count); the top 50 were interviewed personally, while the rest were interviewed by telephone.

The in-person visits to schools began with a reinterview with the respondent who originally completed the TLR. The school respondents were instructed to complete another TLR and to "think aloud" as they filled out the form. The interviewer then compared the reinterview TLR to the original TLR and reconciled any differences. After reconciling the original and reinterview TLRs, the interviewer attempted to determine reasons for discrepancies between the school and the LEA.

The school respondents contacted by telephone did not complete another TLR. Instead, the telephone interviewer attempted to determine the reasons for differences between the TLR completed by the school and the TLR completed by the LEA. (Since the private school telephone cases were selected based on size—the largest school in the telephone sample had 11 teachers—burden was not a concern.)

Component 2: TLR versus the School Questionnaire. The sample was selected in the same way as, but did not overlap with, the component 1 sample. The component 2 sample consisted of 290 public schools and 300 private schools. Once again, 100 public and 100 private schools were selected. The samples were chosen based on the timing of school questionnaire receipt. The 100 public and 100 private schools selected were the ones with the largest difference rates that were among the first 90 percent of returned surveys.

To begin the second component of the TLVS, another TLR was mailed to each selected school. After the TLR close-out, a field test version of the 1993-94 school questionnaire was mailed. No personal visits were conducted as part of the second component of the TLVS; reconciliation was conducted over the phone. To ensure that the respondent would have the necessary information available, copies of both the completed TLR and school questionnaire were mailed back to the respondents along with a letter describing the study and alerting them that someone would be calling to discuss the differences in teacher counts on the two forms. Reinterviews were conducted from the beginning of May through mid-June, 1993 (Royce, 1994, pp. 1-8).

Summary of Results

The findings from the first component of the TLVS suggested that public schools provided more accurate lists of teachers than their corresponding LEA. While both public schools and LEAs often omitted part-time and specialized subject matter teachers from the TLR, LEAs often incorrectly listed guidance counselors, while public schools most often erroneously listed librarians and speech therapists as teachers. Although the private schools incorrectly included nonteachers and incorrectly excluded teachers, the instances were few in each teacher/non-teacher group. Private schools had a much lower incidence rate of excluding teachers and including nonteachers than public schools or LEAs.

The second component of the TLVS indicated that both the public and private schools were more accurate listing teachers using the TLR than the school questionnaire. Public and private schools often omitted part-time teachers when reporting their teacher count using the school questionnaire. Finally, the types of nonteachers most often included in error on the questionnaire by the pubic and private schools were librarians and pre-kindergarten teachers.

Unfortunately, the reinterview and reconciliation did not gather adequate reasons for why schools excluded certain teachers and incorrectly included other persons. Those respondents who provided reasons usually said they had simply forgotten about "that person" or had not thought a person should/shouldn't be included (Royce, 1994, p. 9).

Recommendations

Based on the TLVS, the Bureau of the Census recommended NCES

- Continue to administer the TLR to public schools, instead of LEAs, to obtain the most accurate teacher counts
- Base teacher counts on the TLR, rather than the school questionnaire
- Clarify TLR instructions regarding who is to be included as a teacher

The Bureau of the Census also recommended that more in-depth information gathering techniques be applied to the TLR. Specifically, they recommended cognitive interviews to attempt to uncover with greater understanding "why" respondents made the kinds of errors they did when filling out the form.

How to Count Teachers, Part 2: SASS Teacher Listing Form Study

As mentioned in the previous section, the Schools and Staffing Survey (SASS) uses the Teacher Listing Record (TLR), also referred to as the Teacher Listing Form (TLF), to obtain a list of teachers in each school. The Bureau of the Census then selects a sample of teachers from each school for the Teacher Survey portion of SASS. The accuracy of this list of teachers is crucial to the final estimates of teacher characteristics provided on the Teacher Survey. The exclusion of teachers and/or the inclusion of nonteachers contributes to the magnitude of the measurement error for the survey. The exclusion of teachers, for example, means that a segment of the population was not covered by the sample, leading to bias in the estimates. SASS conducted the TLVS to determine the accuracy and completeness of the lists of teachers provided by the schools. This cognitive study attempted to answer the question: Why do respondents make errors when completing the TLR and the School Survey?

Methodology and Design

Cognitive interviews were conducted with one person, most often the school principal, from 19 schools each (9 public and 10 private) in York County, Pennsylvania; Frederick County, Maryland; and Spotsylvania and Richmond Counties, Virginia. A combination of cognitive techniques were used during the interviews, including the concurrent think-aloud technique, the use of paraphrasing, and retrospective interviewing. The length of the interviews varied between 30 to 90 minutes depending on the size of the school. The interviews were tape-recorded with the respondent's permission (Jenkins and Von Thurn, 1996, pp. 2-3).

Summary of Results

Some of the misreporting problems were a result of respondents not reading or not understanding the importance of the instruction to call the Bureau of the Census if the grade range printed on the cover differed from the grade range of the school. When NCES has defined the school differently than it is defined in a particular state, returning the form without notifying the Bureau of the Census can lead to misreporting. For example, in the Midwest, NCES may define a small school as two schools based on grade range when the state defines it as one school.

The cognitive interviews revealed that it would be easier for respondents to report the total counts of teachers for the requested categories *after* preparing the list. Operationally, respondents need to determine who to list before they can determine how these people should be reported in the categories requested. Also the include/exclude list did not specify a definition for all of the categories.

Another interesting result the cognitive interviews brought to light was the difficulty respondents had in reporting the subject a teacher is most qualified to teach. Respondents either reported all of the subjects teachers were equally qualified to teach or they had great difficulty making the decision.

Finally, it was determined that the questionnaire did not necessarily ask respondents for information in a way that was familiar to them. For example, the question asking if the school is only prekindergarten, kindergarten, or postsecondary asked for the opposite of what respondents expected.

- Is the institution on the cover page a school that has ONLY prekindergarten, kindergarten, and/or postsecondary students? [yes/no]

Most schools teach elementary or secondary students. The emphasis of the question on cases that are the exception rather than the rule and the emphasis on the word "only" was a potential source of error. If respondents overlooked the word "only" and concentrated on "prekindergarten" or "kindergarten" they may have interpreted the question as asking if the school had prekindergarten or kindergarten. Many respondents also required clarification for the word "postsecondary." A suggested replacement phrase was "beyond high school."

Of all the instructions, respondents had the most difficulty understanding the instructions in the fifth paragraph under "How are columns (f)-(o) completed?"

- If an elementary teacher teaches a departmental class, e.g., music, art, reading, math, or science, mark "x" in column (h) ("Other"), under "Elementary." Mark "General elementary" only for elementary teachers who teach in self-contained classes (i.e., teach the same class of students all day or most of the day).

One problem is that "mark 'x' in column (h) ('Other')" can be lost by respondents who assume the second line contains only examples. It appeared that some respondents dropped their eyes to the capitalized mark in "Mark 'General elementary.'" Another problem is that elementary schools are here instructed to enter departmental classes under column headings different than those secondary schools use to enter similar information.

Many of the respondents overlooked the item that asked the month and day the school closes

- Enter the month (April, May, or June) and day this school will close at the end of the 1993-94 school year.

because it was squeezed in between the instructions for listing the teachers and the area for actually listing the teachers. It is also conceptually unrelated to either of these things. In addition, respondents had difficulty interpreting the question because they were uncertain if the question referred to the day the school closed for the students or the day the school closed for the teachers. These were often not the same day, and the question did not specify. Another difficulty surfaced when one school kindergarten ended a few days earlier than grades 1 through 6.

Finally, the cognitive interviews revealed several problems with the item which identified teachers as "new" to the profession. Again, the respondents' own situation was likely to come to mind first. They were not thinking in terms of a teacher's total years in the teaching profession "at all schools" when they began to fill out the TLF.

- Enter an "N" if the teacher's total years in the teaching profession at all schools and school districts (not just their current school/district) is less than 3 years, not counting this school year.

The instructions for this item do clearly state that the teacher's total teaching experience be taken into account, and most respondents caught their mistake. However, there were other problems associated with this item. For example, some respondents placed an (X) in the column instead of an (N). The instructions for other items up until this point required the respondent to mark (X). This was cited as an example of top-down processing. Other reasons for ambiguity in this item included determining if the question was asking for state or national figures, deciding if college teaching should be considered in the teacher's total teaching experience, and how to report teachers with interrupted service.

Recommendations

Jenkins and Von Thurn (1996) derived several recommendations from the cognitive interviews. Specifically, these included

- Reorganize the TLF so that important conceptual information is not interrupted by unrelated information
- Note more prominently on the cover page what to do if the school's grade range differs from the one preprinted on the cover
- Reword the screener questions
- Reverse the list and column reporting
- Re-design the include/exclude list
- Give respondents a decision rule for choosing the subject matter the teacher is most qualified to teach
- Re-design the questionnaire so that its questions more closely mirror likely respondent situations

SASS Public School 1991-92 Field Test Questionnaire

One of the reasons NCES conducts field tests is to get feedback about newly developed items. The Public School 1991-92 Field Test Questionnaire included several newly developed items and questions that attempted to measure difficult concepts. Cognitive research was conducted to gain in-depth information about these items.

Methodology and Design

Jenkins, Ciochetto, and Davis (1992) conducted cognitive interviews with principals of 20 public schools, four in each of five mid-western states: Oklahoma, North Dakota, South Dakota, Nebraska, and Iowa. They used the concurrent think-aloud technique, a procedure in which respondents are asked to read aloud and to verbalize their thoughts as they complete a self-administered questionnaire. The respondents' observations were recorded.

Summary of Results

The results of the cognitive research on the Public School 1991-92 Field Test Questionnaire were given in two sections. The first section focused on respondents' understanding of the school for which they were supposed to report. The second section reviewed the questionnaire item-by-item.

Respondent's understanding of the school for which they are to report. The Teacher List Validation Study and the followup cognitive research revealed the inadequacy of the important instruction on the cover page about what to do if the grade range for the school differed from the one preprinted on the teacher listing form. Therefore, determining if respondents understood which school they were supposed to report on was seen as an important goal of this research. Confusion most often arose when the school was closely associated with another school. For example, in a small school system where typically two or three schools made up the entire school district, these schools were often under the same roof or in buildings clustered together in a group. Although each of these schools usually had its own principal, each principal felt capable of reporting the information requested for all of the schools. The "K-12" answer categories and references, regardless of the grade range of the school for which the questionnaire was intended, justified the principals' assumption that the questionnaire was meant for the entire school system and not just for their individual school. Jenkins, Ciochetto, and Davis concluded that this misunderstanding "led to the systematic overreporting of student and teacher counts on the form...[and] resulted in a large number of inter-item inconsistencies" (1992, p. 5).

Item-by-item review. The item-by-item review provides very detailed information about the difficulties respondents had with each of the instructions and questions included on the survey. The following are examples of the kinds of results which were obtained during the interviews.

Respondents reported item 2 was difficult to read and understand because of its complicated structure.

- 2. How many students (in head counts) were enrolled in THIS SCHOOL (the school named on the questionnaire label) in grades K-12 or comparable ungraded levels -

Include only students enrolled in the school named on the questionnaire label. Do NOT include prekindergarten or postsecondary students.

a. On or about October 1 of THIS SCHOOL YEAR?
b. On or about October 1 of LAST SCHOOL YEAR?

The question is interrupted by two parenthetical phrases and an instruction. Part b of the question was difficult for the respondents to provide because the information was often archived and the respondent had to request someone retrieve it. As a result, a number of respondents approximated a response.

Interviewers reported that respondents slowed down considerably when reading item 7 and read the instructions looking for clarification.

- 7. How long is the school day for most students in this school?

Report BOTH hours and minutes, e.g., "6" hours and "0" minutes, "5" hours and "45" minutes, etc. If the length of the day varies by grade level, record the longest day.

Respondents were unclear if the question was asking how long the school day was from beginning to end, which would include time for recesses, between class periods, and for lunch, or how long the day was in instructional time only.

The placement of item 16 was identified as a problem. Many respondents, especially those in elementary schools, thought the item was somehow related to the preceding series of questions on Limited English Proficiency (LEP). The question was placed at the bottom of the page without a transitional statement to alert the respondent that the topic had changed. Respondents in middle or high schools did not understand the purpose of the question since they had been reporting for these grades all along.

- 16. Does this school have any students in any of grades 7 through 12?

The intent of item 33 was ambiguous to respondents.

- 33. Are this school's standardized student test results released to the public at least annually? (Yes or No)

Respondents were unclear about the meaning of the phrase “released to the public.” Interpretations of the phrase varied from “released to the media” exclusively, to sending individual test results home to parents, or giving a verbal report of the summary results at a schoolboard meeting (Jenkins, Ciochetto and Davis, 1992, p. 31).

Respondents had difficulty understanding the questionnaire concept of full- and part-time. Results from the cognitive interviews revealed that respondents did not think in terms of two-way classifications. Jenkins, Ciochetto, and Davis (1992, p. 31) gave several examples of questions respondents posed about how to classify employees in different situations. One of these questions was how should they classify an employee who worked part-time for the school, but full-time for the school district? Another question was how to classify employees who worked at jobs that by definition could never be considered full-time jobs, such as bus driver. Some respondents ignored the full-time versus part-time classification and recorded all of the employees at the school as full-time, leading to systematic overreporting of the number of employees in full-time positions.

NCES used these results to redesign the questionnaire before it was mailed as part of the full-scale study. The 20 respondents who were willing to participate in the cognitive interviews contributed substantially to this redesign.

CHAPTER 8
Summary

To some extent, the original intent of this report was to draw conclusions across multiple surveys, but the data did not lend themselves to cross-survey comparisons. The idea was to group items that were included in the measurement error studies into categories based on the type of item (i.e., factual, attitudinal, date, retrospective, etc.) and to compare results of the measurement error studies for these items. Among the reasons that did not make this feasible were 1) the items were not easily collapsed into comparable categories, and 2) the statistics and methodologies used to examine measurement error were inconsistent from survey to survey and from study to study. However, listed below are topical summaries of our review of a number of NCES measurement error studies.

Success of item revisions. There were instances when revised items were included as part of the reinterview for the following cycle of the survey to test the success of the revision. For example, the 1987-88 SASS School Administrator Survey included an item asking administrators "Which of the following college degrees have you earned? (Mark all that apply)," followed by a list of degrees. When the item was repeated in the reinterview, reports of bachelor's degrees had a GDR of 20.3 and an IOI of 98.5 and reports of master's degrees had a GDR of 9.9 and an IOI of 49.4. NCES decided to revise the question format so that in the 1990-91 cycle administrators were asked about bachelor's and master's degrees in two yes or no questions (e.g., Do you have a Bachelor's degree? yes/no). The results from the reinterview for bachelor's degree were a GDR of 1.3 (there were too few cases to estimate IOI) and for master's degree a GDR of 1.7 and an IOI of 11.3. Similar items were also included in the 1987-88 and 1990-91 SASS Teacher Survey, with virtually the same results.

A second example comes from the NHES surveys. In the NHES:91 Early Childhood Education (ECE) component reinterview, two items that concerned researchers involved children's involvement in a Head Start program: "Is (child's) program at this daycare center a Head Start Program?" (GDR 12.9, IOI 41.3) and "Is the program at the (first/next) (nursery school/prekindergarten) a Head Start Program?" (GDR 13.5, IOI 43.9). It was decided that these questions were too indirect: children may have been incorrectly identified as Head Start children, possibly because parents were unsure what constituted a Head Start program, possibly because children were enrolled in programs that also enrolled Head Start children. The NHES:93 School Readiness (SR) component contained revised versions of the questions that were more direct and seemingly more successful: the two that were included in the reinterview were "Is (child) now attending or enrolled in Head Start?" (GDR 4.9, IOI 31.3) and "[Prior to starting (kindergarten/first grade), did/has] (child) ever (attend/attended) Head Start?" (GDR 3.6, IOI 19.7).

Parent versus child reliability. Comparisons were made across groups of respondents for the NHES:93 School Safety and Discipline (SS&D) component. The reliability of responses from the students was similar to the reliability rates of their parents, suggesting that youth can respond effectively to telephone surveys like NHES. Comparison of child to parent responses in High School and Beyond (HS&B) and the National Education Longitudinal Study of 1988 (NELS:88) studies also showed that children were fairly reliable reporters for the items measured.

Mode effects. The comparison of mode effects for the SASS School Survey showed that reinterviews conducted by mail in the original interview as well as in the reinterview have less variance than telephone reinterview respondents (Bushery, Royce, and Kasprzyk, 1992). SASS has made trying to obtain reinterviews using the same mode (preferably mail) part of its reinterview study design. In general, NCES has been following the guideline of using the same mode for the reinterview as was used in the original interview, even before Bushery et al. reported on it. The NPSAS:92-93 Field Test changed modes between original interview and reinterview. Subsequently, the results of this study may be of limited benefit, as the reinterview data collection agent has warned.

Difficulties of response process. A number of the reinterview studies, including RCG:91, B&B:93/94 Field Test, BPS:90/92 & 90/94 Field Tests, and NSOPF-93 Faculty Field Test, contained items that required the respondent to recall events from the past (e.g., employment, degrees earned, etc.). Brick, Cahalan et al. (1994) found that items requesting recent or current information generally have lower response variance than items requesting retrospective or future information. Increased response variance also occurs in items asking for respondents to remember events occurring on a specific date in the past. This type of question is particularly troublesome since it not only requires the respondent to recall retrospective data, but also requires that the respondent link past events with specific dates. RCG:91 asked respondents if they were looking for work during the week of April 22, 1991 (GDR 58.8). In the NHES:91 ECE component, parents were asked in what month and year their child started kindergarten (GDR 27.4, IOI 48.0). RGC:91, the B&B Field Test reinterview, the BPS Field Test reinterview, and the HS&B validity evaluation study all included items whose higher rates of inconsistency researchers concluded might be explained by poor respondent recall. For example, B&B respondents were asked to provide their employment history, albeit in a slightly different manner, in the field test and the field test reinterview. Analysis showed that some respondents failed to report in the reinterview brief periods of employment, or unemployment, following graduation that they had reported during the original interview. These differences could be due to respondent recall. Likewise, in BPS, the more recent data collected on financial aid had higher correlations than previous year data which respondents had to recall. Results from the HS&B validity evaluation study also showed that the quality of retrospective information declined over time.

Items requesting dollar amounts. Income, expenditures, financial aid, grants, etc. are popular items on NCES surveys. However, these items tend to have high response variance. For instance, on the SASS Library 93 Reinterview study, expenditures on subscription acquisitions had a GDR L-Fold of 44.4 and L-Fold index of 49.2. Similarly, reported total income on the BPS:90/94 Field Test showed a low correlation (r=0.40).

Obtaining reliable data concerning dollar figures, especially income, has been consistently problematic.

Attitudinal or nonfactual items. Attitudinal or nonfactual items generally have lower reliability than factual items (SASS, BPS field test, HS&B validity). For example, SASS reinterviews have shown that the items asking for the opinions, perceptions, and future expectations of teachers and school administrators are, almost without exception, subject to high response variability. In the BPS field test, students were asked to rate their satisfaction with various services, programs, and features of their school. Only about 65 percent of the students gave identical ratings in the two interviews.

Items with many response categories. These items are subject to high response variability. SASS reinterview studies demonstrated that moderate reductions in variability can be achieved by combining responses to 4-point scales into two categories. Similar results were shown in the NSOPF-93 field test reinterview study. Another example was already mentioned under *Success of item revisions.* SASS revised a "mark all that apply" question format to a direct question format, producing more reliable data.

Adjacent questions using different Likert scales. Cognitive research on the NHES:93 SS&D component has shown that, at least in telephone interviews, adjacent questions using different Likert scales cause respondents difficulty.

Composite variables. Generally, composite variables are more consistent than individual items when measuring complex concepts. Two composite variables were studied as part of the NHES:93 reinterview: 1) the percentage of hours children spent watching television during a week and 2) a developmental scale based on a set of items on the SR survey. Both of these variables appeared more reliable than the individual items (Brick, Rizzo and Wernimont, 1994). Other NCES measurement error studies (RCG:91 and the HS&B multiple indicator study) attempting to compare results from individual items to composite variables found similar results. For example, the HS&B study included a comparison of responses by sets of twins. These data showed composites had much higher correlation coefficients than those for individual items. In NHES:95, however, an overall participation composite variable had a GDR higher than most of the individual items.

Future Methodological Considerations. NCES continues to conduct research in the area of measurement error for its surveys. This report can serve as one of the references used by managers as they continue to address a number of questions. Is there a "best" approach to examining measurement error for a particular survey? Depending upon the approach chosen, is there an acceptable (minimum) sample size? In the case of reinterview studies, is reconciliation the best approach to estimating bias? Are there some measurement error methodologies that NCES should continue to invest in, and some NCES should not? Since different surveys use different methodologies to examine measurement error, is one method better, and should all programs move to that method?

References

Abraham, S., Suter, N., Spencer, B., Johnson, R., Zahs, D., Myers, S. and Zimbler, L. 1994. *1992-93 National Study of Postsecondary Faculty Field Test Report.* (NCES 93-390). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Abt Associates, Inc. 1993. *The National Postsecondary Student Aid Study, Field Test Report.* Chicago, IL. Contractor report to the National Center for Education Statistics.

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.

Alwin, D. and Jackson, D. 1980. "Measurement Models for Response Errors in Surveys: Issues and Applications," *Sociological Methodology 1980*: 68-113.

Andrews, F. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modelling Approach," *Public Opinion Quarterly,* Summer, 48, 2: 409-422.

Bailar, B. and Dalenius, T. 1969. "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model," *Sankhya B* 31: 341-360.

Bailey, L., Moore, T. and Bailar, B. 1978. "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample," *Journal of the American Statistical Association,* 73, 1: 16-23.

Biemer, P. and Fecso, R. 1995. "Evaluating and Controlling Measurement Error in Business Surveys," in B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds.), *Business Survey Methods*. 257-282. New York: John Wiley & Sons, Inc.

Biemer, P. and Forsman, G. 1992. "On the Quality of Reinterview Data with Application to the Current Population Survey," *Journal of the American Statistical Association,* 87, 420: 915-923.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N. and Sudman, S. 1991. *Measurement Errors in Surveys*. New York: John Wiley & Sons, Inc.

Biemer, P. and Stokes, S. 1991. "Approaches to the Modeling of Measurement Error," in P. Biemer, R. Groves, L. Lyberg, N. Mathiewetz, and S. Sudman (eds.), *Measurement Error in Surveys*. 487-516. New York: John Wiley & Sons, Inc.

Bohrnstedt, G. 1983. "Measurement," in P. Rossi, R. Wright, and A. Anderson (eds.), *Handbook of Survey Research.* 70-122. New York: Academic Press.

Brick, J., Cahalan, M., Gray, L., Severynse, J. and Stowe, P. 1994. *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates.* (NCES 95-640). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Celebuski, C., Nolin, M., Squadere, T., Ha, P., Wernimont, J., West, J., Chandler, K., Hausken, E. and Owings, J. 1992. *National Household Education Survey of 1991: Preprimary and Primary Data Files User's Manual.* (NCES 92-057). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Celebuski, C., Nolin, M., Squadere, T., Ha, P., Wernimont, J., West, J., Chandler, K., Hausken, E. and Owings, J. 1991. *NHES:91 1991 National Household Education Survey Methodology Report.* Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Nolin, M., Ha, P., Levinsohn, M. and Chandler, K. 1994a. *National Household Education Survey of 1993: School Safety and Discipline Data File User's Manual.* (NCES 94-218). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Nolin, M., Ha, P., Levinsohn, M. and Chandler, K. 1994b. *National Household Education Survey of 1993: School Readiness Data File User's Manual.* (NCES 94-193). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Kim, K., Nolin, M. and Collins, M. 1996. *Estimation of Response Bias in the NHES:95 Adult Education Survey.* (Working Paper 96-13). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, M., Rizzo, L. and Wernimont, J. 1997. *Reinterview Results for the School Safety & Discipline and School Readiness Components.* (NCES 97-339). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, M., Tubbs, E., Collins, M., Nolin, M., Cantor, D., Levin, K. and Carnes, Y. 1997. *Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey.* (Working Paper 97-02). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, M., Wernimont, J. and Montes, M. 1995. *The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component.* (Working Paper 96-14). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Burkheimer, Jr., G., Forsyth, B., Wheeless, S., Mowbray, K., Boehnlein, L., Knight, S., Veith, K. and Knepper, P. 1992. *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report. BPS:90/92.* (NCES 92-160). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Burkheimer, Jr., G., Forsyth, B., Whitmore, R., Wine, J., Blackwell, K., Veith, K., Borman, G. and Knepper, P. 1994. *Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS 90/92) Final Public Technical Report.* (NCES 94-369). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Bushery, J., Royce, D. and Kasprzyk, D. 1992. "The Schools and Staffing Survey: How Reinterview Measures Data Quality," *American Statistical Association Proceedings of the Section on Survey Research Methods*: 458-463.

Chaney, B. 1994. *The Accuracy of Teachers' Self-Reports on Their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey.* (Working Paper 94-04). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Davis, C. and Sonnenberg, B. 1994. *Programs and Plans.* (NCES 94-133). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

DeMaio, T. 1990a. "Cognitive Interviews with Private School Administrators." Bureau of the Census Memorandum, June 6.

DeMaio, T. 1990b. "Cognitive Interviews with Private School Teachers." Bureau of the Census Memorandum, April 26.

Faupel, E., Bobbit, S. and Friedrichs, K. 1992. *1988-89 Teacher Follow-up Survey (TFS) Data File User's Manual.* (NCES 92-058). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Feindt, P. 1996. *Reinterview Report: Response Variance in the 1993 Library Survey.* Bureau of the Census Memorandum from Preston Jay Waite to Sherry L. Courtland, January 17.

Fellegi, I. 1964. "Response Variance and Its Estimation," *Journal of the American Statistical Association,* 59: 1016-1041.

Fetters, W., Stowe, P. and Owings, J. 1984. *High School and Beyond, A National Longitudinal Study for the 1980s, Quality of Responses of High School Students to Questionnaire Items.* (NCES 84-216). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Forsman, G. 1987. "Analys av Aterintervjndata (Analysis of Reinterview Data)," *Urval* 19. Statistics Sweden.

Forsman, G. and Schreiner, I. 1991. "The Design and Analysis of Reinterview: An Overview," in P. Biemer, R. Groves, L. Lyberg, N. Mathiewetz, and S. Sudman (eds.), *Measurement Error in Surveys*. 279-302. New York: John Wiley & Sons, Inc.

Forsyth, B. and Lessler, J. 1991. "Cognitive Laboratory Methods: A Taxonomy," in P. Biemer, R. Groves, L. Lyberg, N. Mathiewetz, and S. Sudman (eds.), *Measurement Error in Surveys*. 393-418. New York: John Wiley & Sons, Inc.

Frey, J. 1989. *Survey Research by Telephone*. Vol. 150 of the SAGE Library of Social Research. New York: SAGE Publications, Inc.

Green, P., Meyers, S., Giese, P., Law, J., Speizer, H., Tardino, V. and Knepper, P. 1996. *Baccalaureate and Beyond Longitudinal Study: 1993/94 First Follow-up Methodology Report.* (NCES 96-149). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Green, P., Speizer, H., Campbell, B. and Knepper, P. 1994. *Baccalaureate and Beyond Longitudinal Study First Followup Field Test Report, 1993.* (NCES 94-371). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Groves, R. 1989. *Survey Errors and Survey Costs.* New York: John Wiley & Sons, Inc.

Gruber, K., Rohr, C. and Fondelier, S. 1996. *1993-94 Schools and Staffing Survey: Data File User's Manual, Volume I: Survey Documentation.* (NCES 96-142). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Gruber, K., Rohr, C. and Fondelier, S. 1994a. *1990-91 Schools and Staffing Survey: Data File User's Manual, Volume I: Survey Documentation.* (NCES 93-144-I). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Gruber, K., Rohr, C. and Fondelier, S. 1994b. *1990-91 Schools and Staffing Survey: Data File User's Manual, Volume II: Survey Documentation.* (NCES 93-144-II). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Gulliksen, H. 1950. "Intrinsic Validity." *American Psychologist,* 5: 511-517.

Hansen, M., Hurwitz, W. and Pritzker, L. 1964. "The Estimation and Interpretation of Gross Differences and the Simple Response Variance," in C. Rao (ed.), *Contributions to Statistics*. Calcutta: Statistical Publishing Society.

Hays, W. 1963. *Statistics*. New York: Holt, Rinehart and Winston.

Ingels, S., Scott, L., Rock, D., Pollack, J., Rasinski, K. and Wu, S. 1994. *National Education Longitudinal Study of 1988: First Follow-up Final Technical Report: Base Year to First Follow-up.* (NCES 94-632). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Ingersoll, R., Han, M. and Williams, J. 1994. *School Library Media Centers in the United States: 1990-91.* (NCES 94-326). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Jabine, T. 1994. *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys.* (NCES 94-340). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Jenkins, C. 1997. *Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year.* (Working Paper 97-10). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Jenkins, C. 1993. "Cognitive Interview Summaries on the Phase I Multiplicity Questions Being Examined for the Student Records Questionnaire." Bureau of the Census Memorandum, November 4, 1992.

Jenkins, C. and Ciochetto, S. 1993. "Results of Cognitive Research on the Multiplicity Question from the 1991 Schools and Staffing Survey Student Records Questionnaire (SASS-36A, B)." Bureau of the Census Memorandum, February 10.

Jenkins, C., Ciochetto, S. and Davis, W. 1992. "Results of Cognitive Research on the Public School 1991-92 Field Test Questionnaire for the Schools and Staffing Survey (SASS-3A)." Bureau of the Census memorandum, June 15.

Jenkins, C. and Von Thurn, D. 1996. *Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey.* (Working Paper 96-05). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Jenkins, C. and Wetzel, A. 1995. *The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation.* (Working Paper 95-10). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Jones, C., Sebring, P., Crawford, I., Spencer, B., Butz, M. and MacArthur, H. 1986. *High School and Beyond 1980 Senior Cohort Second Follow-up (1984).* (CS 85-216). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Kaufman, P., Rasinski, K., Lee, R. and West, J. 1991. *Quality of the Responses of Eighth-Grade Students in NELS:88.* (NCES 91-487). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Kendall, M. and Buckland, W. 1971. *A Dictionary of Statistical Terms*. Edinburgh: Oliver & Boyd.

Kendall, M. and Stuart, A. 1979. *The Advanced Theory of Statistics*, Vol. 2. New York: Macmillan Publishing Company, Inc.

Koch, G. 1973. "An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means," *Journal of the American Statistical Association,* 68: 906-913.

Landis, J. and Koch, G. 1976. "A Review of Statistical Methods in the Analysis of Data Arising from Observer Reliability Studies," *Statistica Neerlandica,* 29: 101-123, 151-161.

Lessler, J. 1984. *Measurement Errors in Surveys*, in C.F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena,* vol. 2: 405-440. New York: Russel Sage Foundation.

Lessler, J. and Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: John Wiley and Sons, Inc.

Lessler, J., Kalsbeek, W. and Folsom, R. 1981. "A Taxonomy of Survey Errors, Final Report," Research Triangle Institute project 255U-1791-03F, Research Triangle Park, NC: Research Triangle Institute.

Lessler, J., Tourangeau, R. and Salter, W. 1989. "Questionnaire Design in the Cognitive Research Laboratory," *Vital and Health Statistics*, 6, 3.

Loft, J., Riccobono, J., Whitmore, R., Fitzgerald, R., Berkner, L. and Malizio, A. 1995. *Methodology Report for the National Postsecondary Student Aid Study, 1992-93.* (NCES 95-211). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Lord, F. and Novick, M. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Mahalanobis, P. 1946. "Recent Experiments in Statistical Sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society,* 100, 109: 325-378.

Messick, S. 1989. "Validity," in R. Linn (ed.) *Educational Measurement,* Third Edition. 13-103. New York: American Council on Education and Macmillan Publishing Company.

Monaco, D., Salvucci, S., Zhang, F., Li, B., Hu, M. and Gruber, K. (forthcoming). *An Analysis of Response Rates in the 1993-94 Schools and Staffing Survey.* Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

National Center for Education Statistics. 1987. *1987 Survey of 1985-1986 College Graduates: Data Base Documentation.* Washington, D.C.: author.

National Center for Education Statistics. 1982. *High School and Beyond: Twins and Siblings' File Users' Manual.* Washington, D.C.: author.

Neter, J. 1969. "Measurement Errors in Anticipated Consumer Expenditures," in N. Johnson and H. Smith, Jr. (eds.), *New Developments in Survey Sampling: A Symposium on the Foundations of Survey Sampling Held at the University of North Carolina, Chapel Hill, North Carolina.* 482-505. New York: Wiley Interscience.

Nolin, M. and Chandler, K. 1996. *Use of Cognitive Laboratories and Recorded Interviews in the National Household Education Survey.* (NCES 96-332). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Oksenberg, L. and Cannell, C. 1977. "Some Factors Underlying the Validity of Self-Report," *Bulletin of the International Statistical Institute*: 325-46l.

Oksenberg, L., Cannell, C. and Kalton, G. 1991. "New Strategies for Pretesting Survey Questions," *Journal of Official Statistics,* 7: 349-365.

O'Muircheartaigh, C. 1986. "Correlates of Reinterview Response Inconsistency in the Current Population Survey," *Proceedings of the Second Annual Research Conference.* U.S. Bureau of the Census 1986: 208-234.

O'Muircheartaigh, C. and Marckward, A. 1980. "An Assessment of the Reliability of World Fertility Study Data," *Proceedings of the World Fertility Survey Conference,* 3: 305-379. The Hague: International Statistical Institute.

Pap, A. 1958. *Semantics and Necessary Truth.* New Haven, CT: Yale University Press.

Pap, A. 1953. “Reduction—Sentences and Open Concepts.” *Methodos,* 5: 3-30.

Pratt, D., Burkheimer, G., Forsyth, B., Wine, J., Veith, K., Beaulieu, J. and Knepper, P. 1994. *Beginning Postsecondary Students Longitudinal Study Second Follow-up Field Test Report; BPS 90/94.* (NCES 94-370). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Pratt, D., Whitmore, R., Wine, J., Blackwell, K., Forsyth, B., Smith, T., Becker, E., Veith, K. and Bobbitt, L. 1996. *Beginning Postsecondary Students Longitudinal Study Second Follow-up (BPS:90/94) Final Technical Report.* (NCES 96-153). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Research Triangle Institute. 1996. *National Postsecondary Student Aid Study: 1996 Field Test Methodology Report.* (Working Paper 96-17). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Royce, D. 1995. *The Results of the 1993 Teacher List Validation Study (TLVS).* (Working Paper 95-09). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Royce, D. 1994. *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report.* (Working Paper 94-03). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Royce, D. 1990. *1989 Teacher Followup Survey (TFS) Reinterview Response Variance Report*. Bureau of the Census Memorandum from Preston J. Waite to Sherry L. Courtland, August 16.

Russell, S., Hancock, M. and Williamson, C. 1990. *1988 National Survey of Postsecondary Faculty Methodology Report.* Arlington, VA: SRI International. Contractor report to the National Center for Education Statistics.

Sarndal, C.-E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Sebring, P., Campbell, B., Glusberg, B., Spencer, B., Singleton, M. and Carroll, C. 1987. *High School and Beyond 1980 Senior Cohort Third Follow-up (1986) Volume I, Data File User's Manual.* Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Selfa, L., Suter, N., Myers, S., Johnson, R., Zahs, D., Kuhr, B., Abraham, S. and Zimbler, L. 1997. *1993 National Study of Postsecondary Faculty Methodology Report*. (NCES 97-467). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Spencer, B., Frankel, M., Ingels, S., Rasinski, K., Tourangeau, R. and Owings, J. 1990. *National Education Longitudinal Study of 1988: Base Year Sample Design Report.* (NCES 90-463). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Sudman, S. and Bradburn, N. 1982. *Asking Questions: A Practical Guide to Questionnaire Design.* San Francisco: Jossey-Bass Publishers.

Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods," in T. Jabine, E. Loftus, M. Straf, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, D.C.: National Academy of Science.

Tourangeau, R., McWilliams, H., Jones, C., Frankel, M. and O'Brien, F. 1983. *High School and Beyond: First Follow-Up (1982) Sample Design Report.* Contractor report to the National Center for Education Statistics.

U.S. Department of Commerce, Bureau of the Census. 1985. *Evaluating Censuses of Population and Housing* (ISP-TR-5). Washington, D.C: author.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. 1995. *Statistical Evaluation Practices.* Washington, D.C.: author.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. 1993a. *National Education Longitudinal Study of 1988: NELS:88 Questionnaires, Base Year through Second Follow-up. Vol. 1 Base Year (1988).* Washington, D.C.: author.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. 1993b. *Quality of Responses in the 1987 National Postsecondary Student Aid Study.* Washington, D.C.: author.

Westat, Inc. 1994. *1991 Survey of Recent College Graduates: Methodology Report.* Washington, D.C.: Contractor report to the National Center for Education Statistics.

Westat, Inc. 1992. *Methodology Report for the 1990 National Postsecondary Student Aid Study, Contractor Report.* (NCES 92-080). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Westat, Inc. 1990. *Methodology Report for the National Postsecondary Student Aid Study, 1987, Contractor Report.* (NCES 90-309). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Whitener, S., Kaufman, S., Rohr, C., Bynum, L. and King, K. 1994a. *1991-92 Teacher Followup Survey Data File User's Manual Public Use Version.* (NCES 94-331). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Whitener, S., Rohr, C., Bynum, L.T., Kaufman, S. and King K. 1994b. *1991-92 Teacher Followup Survey Data File User's Manual Restricted Use Version.* (NCES 94-478). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Zahs, D., Pedlow, S., Morrissey, M., Marnell, P., Nichols, B., and Carroll, C. 1995. *High School and Beyond Fourth Follow-up Methodology Report.* (NCES 95-426). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Appendix

A Summary of a Reinterview Study on Mode Effects for the 1990-91 Schools and Staffing Survey (SASS) School Survey

Comparison of Reinterview Mode Effects for the 1900-91 SASS School Survey

To determine the impact mode change might have on data quality, these SASS reinterviews were generally conducted using the same data collection mode as in the original interview: a mail reinterview for mail respondents and a telephone reinterview for telephone follow-up cases. Reinterviews were conducted for about 465 mail-mail and 270 telephone-telephone respondents. All attempts were made to reinterview the respondent who had answered the original school survey.

The reinterview included 45 questions which could be grouped into the categories:

- Student population/teacher population - 5
- Type of school/community - 4
- Grades and classes - 25
- Teacher vacancies/teacher programs - 11

The results are shown in tables A-1 through A-4. Column headings with "MM" and "MT/TT" refer to how the interviews were conducted. MM means both the original interview and the reinterview were by mail. MT means the original interview was by mail and the reinterview by telephone. TT means both interviews were by telephone. Since no significant differences were found between the MT and TT cases, these two categories were combined.

Student Population/Teacher Population

The reinterview included five questions to assess the quality of data collected on the student population estimate (Questions 1a and 1b) and the teacher population estimate (Questions 8, 9a, and 9b) of the school. All five of the questions were open-ended and the responses later divided into categories.

Table A-1 shows the response variances (L-fold index and GDR) for these items. Results for the student population items indicate that the response variances were all less than 20 percent. However, among respondents where the reinterview was completed by telephone, MT/TT cases, the L-fold index and GDR indicate results more than twice those found when the original and the reinterview were completed by mail, MM cases.

Total results for the three teacher population estimate questions show the L-fold index in the high range. When comparing the different modes, however, the L-fold index of the MM items were all in the moderate range, while the L-fold index of the MT/TT cases were all in the high range—above 50. The response variance for the MM cases was significantly lower than that for the MT/TT cases for all five questions. Significant differences were also found in the agreement rate for both student questions: MM cases were higher than MT/TT cases (from table L, Royce, 1994, not shown).

Table A-1. -- Response variance and gross difference rates by student and teacher population

	L-fold Index			Gross Difference Rate (%)		
Question	**Total**	**MM**[1]	**MT/TT**[2]	**Total**	**MM**[1]	**MT/TT**[2]
1a: Number of Students This Year	12.3	7.4	18.2	8.5	4.9	12.9
1b: Number of Students Last Year	12.1	7.3	16.8	8.4	4.9	12.2
8: New K - 12 Teachers	51.5	42.8	61.1	44.3	36.7	52.6
9a: K - 12 Teachers that Left	53.3	43.8	64.1	44.5	36.8	53.3
9b: No Longer Teaching	54.9	48.1	67.5	39.4	35.5	45.8

[1] MM refers to cases that were originally interviewed by mail *and* reinterviewed by mail.
[2] MT/TT refers to cases that were originally interviewed by mail and reinterviewed by telephone (MT) and cases that were originally interviewed by telephone and reinterviewed by telephone (TT).

SOURCE: Derived from table J, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper No 94-03), p. 12.

Type of School/Community

This group of items examined the level and type of school (Questions 2 and 3), the community in which the school is located (Question 4), and the number of days in the school year (Question 5). Three of the questions were multiple choice and one was open-ended. The MT/TT L-fold index and GDR results were more than twice those found for the MM cases for all three of the multiple choice items; this difference was statistically significant. The last item, the number of days in the school year, examined the exact agreement between the interview and reinterview. The results indicate that the MM cases were significantly more reliable than the MT/TT cases.

Table A-2. -- Response variance and gross difference rates by type of school and community

	L-fold Index			Gross Difference Rate (%)		
Question	**Total**	**MM**[1]	**MT/TT**[2]	**Total**	**MM**[1]	**MT/TT**[2]
2: What is the level of this school? (4 categories)	12.5	8.2	17.3	8.6	5.5	12.1
3: What type of school is this? (4 categories)	26.7	16.3	35.4	6.4	3.4	9.6
4: Which of these best describes the community in which the school is located? (10 categories)	37.6	24.0	51.9	30.4	19.0	42.7

[1] MM refers to cases that were originally interviewed by mail and reinterviewed by mail.
[2] MT/TT refers to cases that were originally interviewed by mail and reinterviewed by telephone (MT) and cases that were originally interviewed by telephone and reinterviewed by telephone (TT).

SOURCE: Derived from table M, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper No 94-03), p. 14.

Grades and Classes

This group of items (see table A-3) included eight questions about programs the school offers (Questions 6a through 6h) and 17 questions about the grade levels of instruction at the school (Questions 7-1 through 7-17). The first eight were yes/no questions, while the grade levels of instruction were "mark all that apply." All parts of Question 6, with the exception of 6g (diagnostic and prescriptive services), exhibited moderate response variance. Question 7 had low response variance for each of the choices, with the exception of "Ungraded" (high), "Nursery" (moderate), and "Postsecondary" (not enough information to compute a reliable index). Among the 23 out of 25 items for which a reliable index could be computed (excluding "Nursery" and "Postsecondary"), the response variance was significantly lower when both the original and the reinterview were completed by mail than when the reinterview was completed over the telephone.

Table A-3. -- Response variance and gross difference rates by grades and classes

	L-fold Index			Gross Difference Rate (%)		
Question	Total	MM[1]	MT/TT[2]	Total	MM[1]	MT/TT[2]
6a: English as a second language	30.1	24.2	36.5	13.7	10.9	16.8
6b: Bilingual education	45.1	31.5	55.9	12.1	6.9	17.8
6c: Remedial reading	48.0	36.4	59.0	16.9	12.1	22.0
6d: Remedial mathematics	47.5	37.7	58.1	22.3	17.7	27.2
6e: Programs for handicapped students	28.1	25.3	31.2	10.4	7.8	13.3
6f: Programs for the gifted and talented	35.4	28.8	41.9	15.5	11.8	19.4
6g: Diagnostic and prescriptive services	59.7	54.0	65.2	20.0	16.2	24.1
6h: Extended day or before- or after-school day-care programs	24.7	19.7	29.7	8.8	6.7	11.2
7-1: Ungraded	57.9	44.9	73.5	8.0	6.5	9.6
7-2: Nursery	29.0	22.9	-	2.3	2.2	2.4
7-3: Prekindergarten	19.9	12.1	28.7	5.2	3.3	7.3
7-4: Kindergarten	9.9	5.7	14.6	5.0	2.8	7.3
7-5: First	10.9	5.7	16.5	5.4	2.8	8.2
7-6: Second	10.2	4.8	16.0	5.1	2.4	8.0
7-7: Third	11.3	5.7	17.4	5.6	2.8	8.7
7-8: Fourth	11.5	6.1	17.4	5.8	3.0	8.7
7-9: Fifth	10.6	5.2	16.5	5.3	2.6	8.2
7-10: Sixth	10.9	4.8	17.6	5.4	2.4	8.7
7-11: Seventh	9.4	3.6	15.7	1.5	1.7	7.5
7-12: Eighth	10.0	4.0	16.6	4.9	2.0	8.0
7-13: Ninth	7.4	4.1	10.9	3.6	2.0	5.4
7-14: Tenth	6.8	4.3	9.5	3.3	2.0	4.7
7-15: Eleventh	6.1	2.8	9.5	2.9	1.3	4.7
7-16: Twelfth	4.7	2.3	7.1	2.3	1.1	3.5
7-17: Postsecondary	-	-	-	2.3	2.2	2.4

[1] MM refers to cases that were originally interviewed by mail and reinterviewed by mail.
[2] MT/TT refers to cases that were originally interviewed by mail and reinterviewed by telephone (MT) and cases that were originally interviewed by telephone and reinterviewed by telephone (TT).

SOURCE: Derived from table P, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper No 94-03), p. 17.

Teaching Vacancies/Teacher Programs

This group of items examined teaching vacancies in the school for the year (Questions 10a, 10b, and 10c), an evaluation program for teachers (Question 11), and programs to help beginning teachers (Question 12). Questions 10a, 10b, 11, and 12 were yes/no questions, while Question 10c was a "mark all that apply" question. Two of the four items had high response variance (Questions 10a and b) and two had moderate response variance (Questions 11 and 12) over all cases. Of these, all but Question 11 had significantly lower response variance in the MM cases than the MT/TT cases. For Question 10c, none of the categories met the minimum requirements to compute a reliable index.

Table A-4. -- Response variance and gross difference rates by teaching vacancies and teaching programs

		L-fold Index			Gross Difference Rate (%)		
Question		Total	MM[1]	MT/TT[2]	Total	MM[1]	MT/TT[2]
10a:	Any vacancies	55.1	40.1	69.8	19.1	13.0	25.7
10b:	Vacancies that could not be filled	52.6	41.2	65.6	10.0	7.4	13.4
10c-1:	Canceled planned course offerings	-	-	-	5.4	0.0	14.3
10c-2:	Expanded some class sizes	-	-	-	2.7	0.0	7.1
10c-3:	Added sections to other teachers' normal teaching loads	-	-	-	8.1	4.3	14.3
10c-4:	Assigned a teacher of another subject or grade level to teach those classes	-	-	-	27.0	21.7	35.7
10c-5:	Used long-term and/or short-term substitutes	-	-	-	16.2	4.3	35.7
10c-6:	Used part-time or itinerant teachers	-	-	-	8.1	4.3	14.3
10c-7:	Hired a less qualified teacher	-	-	-	27.0	21.7	35.7
11:	Evaluation program	45.4	-	57.3	4.3	1.3	7.5
12:	Program for beginning teachers	49.5	34.6	65.2	23.9	16.2	32.2

[1] MM refers to cases that were originally interviewed by mail and reinterviewed by mail.
[2] MT/TT refers to cases that were originally interviewed by mail and reinterviewed by telephone (MT) and cases that were originally interviewed by telephone and reinterviewed by telephone (TT).

SOURCE: Derived from table Q, Royce, (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report* (Working Paper No 94-03), p. 19.

Bushery, Royce, and Kasprzyk (1992) suggest four possible reasons why the reinterviews completed by mail (MM cases) show lower response variances than the reinterviews completed by telephone (MT/TT cases). First, only respondents who answered the original survey by mail were eligible for the mail reinterview. These respondents were more likely to be more cooperative and answer the questions more carefully in both interviews. Second, respondents interviewed by mail may take more time than those interviewed by telephone to look up the answers to questions from records or may go through a more careful, lengthy thought process to provide the needed facts. Respondents interviewed by telephone may not feel free to take the time to look up records while the interviewer is waiting on the phone. Third, mail respondents may leave more difficult or uncertain questions blank. Telephone interviewers may manage to obtain answers to a difficult question, but they may be unreliable answers. Fourth, mail respondents may photocopy the original questionnaire after completing it and refer to their original answers when completing the mail reinterview. However, the authors feel this last explanation would have only accounted for a small part of the mail-mail versus telephone-telephone differences, and so they consider some combination of the first three explanations is the most plausible. Finally, the authors suggest that mail respondents, by definition, are more cooperative and motivated than those followed up by telephone. Mail interviewing, moreover, probably promotes more careful responses and more use of records, compared to those interviewed by telephone.

United States
Department of Education
Washington, DC 20208–5651

Official Business
Penalty for Private Use, $300