Creating Time Series Data Sets: Reconciling the Conflicting Imperatives of Continuity and Change

Steven J. Ingels[1]

American Educational Research Association annual meeting,
Session 67.053, San Diego, California, April 16, 2004

Paper presented as part of the Symposium on the Education Longitudinal Study of 2002, sponsored by the Longitudinal Studies Special Interest Group of AERA.

[1] Steven J. Ingels is Senior Education Research Scientist at RTI International (Research Triangle Institute), 1615 M Street NW, Washington DC 20036; sji@rti.org.

Creating Time Series Data Sets:
Reconciling the Conflicting Imperatives
of Continuity and Change

Steven J. Ingels

Abstract. The paper is divided into four parts. *Part 1* provides an overview of the evolving design of the NCES high school longitudinal studies (NLS-72, HS&B, NELS:88, and ELS:2002). *Part 2* describes the various kinds of intercohort analyses that can be undertaken with the studies.

Part 3 (the bulk of the paper) addresses threats to true change measurement. It does so by examining the tension between the need for continuity to ensure true replication (measurement conditions must be kept constant), and the necessity for updating of survey design, content, and methodology (nothing can be frozen in time and remain relevant). Threats to comparability are catalogued across the dimensions of sample design and definition; test and questionnaire content and format; and methods of data collection and processing.

Incremental improvements that may sharpen cross-sectional estimation may pose significant risk to cross-cohort change measurement. At the same time, some changes in design, content and methodology will be both necessary and desirable. *Part 4* of the paper summarizes recommendations for dealing with the tradeoffs between strict continuity for replication, and change in response to altered circumstances and new methodological opportunities.

Part 1: The evolving design of the NCES longitudinal high school studies.

1.1. Background: the four studies (and some more)

In response to the need for policy-relevant, time-series data on the high school experience and post-high school transitions of nationally representative samples of secondary school students, the National Center for Education Statistics (NCES) of the U.S. Department of Education has carried out longitudinal studies for over thirty years. Four separate studies (and eight cohorts[2]) now comprise the NCES longitudinal high school cohorts series: the National Longitudinal Study of the High School Class of 1972 (NLS-72); the sophomore and senior cohorts of High School and Beyond (HS&B); the National Education Longitudinal Study of 1988 (NELS:88); and the Education Longitudinal Study of 2002 (ELS:2002). Taken together, these studies represent the educational experience of youth from four decades -- the 1970s, 1980s, 1990s, and the first decade of the 21st century. A brief description of these studies follows. (See also Figure 1 below.)

1.1.1 The National Longitudinal Study of the High School Class of 1972: NLS-72. The National Longitudinal Study of the High School Class of 1972 (NLS-72) began in the spring of 1972 with a survey of a national probability sample of 19,001 seniors from 1,061 public, secular private, and church-affiliated high schools. The sample was designed to be representative of the approximately three million high school seniors enrolled in more than 17,000 schools in the spring of 1972. Each sample member was asked to complete a student questionnaire and a 69-minute test battery. School administrators were also asked to supply survey data on each student, as well as information about the schools' programs, resources, and grading systems. Five follow-ups, conducted in 1973, 1974, 1976, 1979, and 1986, were conducted, as well as a postsecondary education transcript study.[3]

1.1.2 High School and Beyond: HS&B. The second in the series of NCES longitudinal high school studies was launched in 1980. HS&B included a cohort of high school seniors comparable to the NLS-72 sample. However, the study also extended the age span and analytical range of NCES longitudinal studies by surveying a sample of high school sophomores. Base year data collection took place in the spring term of the 1979–80 academic year with a two-stage probability sample. More than 1,000 schools served as the first-stage units, and 58,000 students within these schools were the second-stage units. Both cohorts of HS&B participants were resurveyed in 1982, 1984, and 1986; the sophomore group also was surveyed in 1992.[4] In addition, to better understand the school and home contexts for the sample members, data were collected from teachers (a teacher comment form in the base year asked for teacher perceptions of HS&B sample members), principals, and a subsample of parents. High school transcripts were collected for a subsample of sophomore cohort members. As in NLS-72, postsecondary transcripts were collected for both HS&B cohorts; however, the sophomore cohort transcripts cover a much longer time span (to 1993).

With the study design expanded to include a sophomore cohort, HS&B provided critical data on the relationships between early high school experiences and students' subsequent educational achievement in high school. For the first time, national data were available that showed students' academic growth over time and how

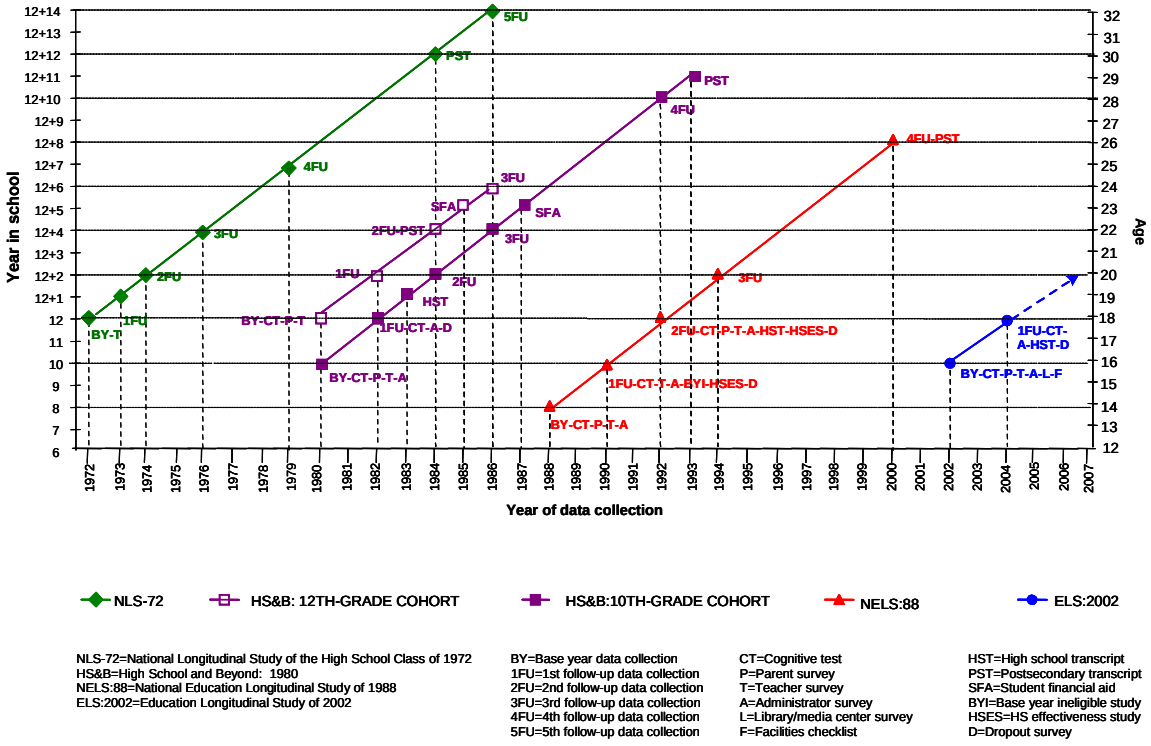
[2] While five cohorts are represented by independently-drawn samples, additional cohorts were formed through sample freshening. Hence the study series comprises eight distinct cohorts: 1972 seniors, 1980 sophomores, 1980 seniors, 1988 eighth graders, 1990 sophomores, 1992 seniors, 2002 sophomores and 2004 seniors.

[3] For detailed information on the NLS-72, see Riccobono, Henderson, Burkheimer, Place, and Levinsohn (1981) and Spencer, Sebring and Campbell (1987). Also see the NCES web site: http://nces.ed.gov/surveys/nls72/. For all the NCES studies, documentation and research reports can generally be found on the NCES website, in the ERIC database at the Educator's Reference Desk (http://www.eduref.org/), or at the International Archive of Education Data (IAED) at the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan [*http://www.icpsr.umich.edu*].)

[4] For a summation of the HS&B sophomore cohort study, see Zahs, Pedlow, Morrissey, Marnell, and Nichols (1995). For further information on HS&B, see the NCES web site: *http://nces.ed.gov/surveys/hsb/*.

family, community, school, and classroom factors promoted or inhibited student learning. Researchers were able to use data from the extensive battery of achievement tests within the longitudinal study to assess growth in knowledge and cognitive skills over time. Moreover, data were then available to analyze the school experiences of students who later dropped out of high school, and eventually, to investigate their later educational and occupational outcomes. These data became a rich resource for policymakers and researchers over the next decade and provided an empirical base to inform the debates of the educational reform movement that began in the early 1980s.

Figure 1. Longitudinal design for the NCES high school cohorts: 1972- 2006



1.1.3 National Education Longitudinal Study of 1988 (NELS:88). Much as NLS-72 captured a high school cohort of the 1970s and HS&B captured high school cohorts of the 1980s, NELS:88 was designed to study high school students of the 1990s—but with a premeasure of their achievement and status, prior to their entry into high school. NELS:88 represents an integrated system of data that tracked students from junior high or middle school through secondary and postsecondary education, labor market experiences, and marriage and family formation.

Data collection for NELS:88 was initiated with the eighth-grade class of 1988 in the spring term of the 1987–88 school year. Some 24,599 students in 1,052 schools participated. Along with a student questionnaire and test battery, NELS:88 included surveys of parents (base year and second follow-up), teachers (base year, first, and second follow-ups), and school administrators (base year, first, and second follow-ups), and high school transcripts were collected. The sample was also surveyed after scheduled high school graduation, in 1994 and 2000, and postsecondary transcripts were collected in 2000-2001. The NELS:88 base year sample was freshened

in 1990 and 1992 to make it nationally representative of a sophomore and a senior cohort in those respective years.[5]

1.1.4 Education Longitudinal Study of 2002 (ELS:2002). Like the HS&B sophomore cohort, ELS:2002 began with a survey of 15,362 participating tenth graders in a nationally-representative sample of 752 high schools. In the 2002 base year, parents, school principals, librarians, and teachers were also surveyed, and students completed tests in reading and mathematics as well as a questionnaire. The first follow-up is currently (2004) underway. The first follow-up returns to the base year sample of schools to measure change two years later, but also follows transfer students, early graduates, and dropouts. The first follow-up sample has been freshened so that it is representative of high school seniors in 2004. High school transcripts will be collected in the autumn of 2004. Further follow-ups are planned for 2006 and thereafter.[6]

1.1.5. Other possible linkages for cross-cohort comparison. While the four studies have been designed to be comparable, this is not to say linkages with other time series data sets are not possible as well. Some of the more important of these possible linkages are discussed below.

The high school cohort transcript studies and the National Assessment of Educational Progress (NAEP) transcript studies can be linked because both code using the Classification of Secondary School Courses (CSSC) and conform to the overall framework for aggregating courses in analysis, the Secondary School Taxonomy (1987, 1998). While the NAEP trend sample complements the achievement trend data derivable from the longitudinal studies, some direct linkages to the regular national NAEP are available as well, since 12th grade mathematics results have been (1992) or will be (2005) equated (that is, the NELS:88 1992 test results have been put on the NAEP scale, as will be the 2004 ELS:2002 results). A similar linkage has been effected with the Program in International Student Assessment (PISA 2000 reading, PISA 2003 math), a series of age cohorts of 15-year-olds.

Certain linkages are also possible with the Department of Labor (Bureau of Labor Statistics) national longitudinal youth cohorts, the NLSY79 and NLSY97. The *National Longitudinal Survey of Youth 1997 (NLSY97)* is a survey of young men and women born in the years 1980-84; respondents were ages 12-17 when first interviewed in 1997, and have been interviewed annually since. In the *National Longitudinal Survey of Youth 1979 (NLSY79)*, subjects were born in the years 1957-64; respondents were ages 14-22 when first interviewed in 1979. The NLSY79 is still underway, now with biennial follow-ups.

The fact that the NCES studies are grade cohorts[7] whereas the NLSY studies are age cohorts poses an obstacle, though not an insuperable one, to linkage and comparison. Because the NLSY studies cover a range of years and collect grade information on respondents (just as the grade-based NCES studies also collect date of birth), certain subsets of the NLSY samples can be converted into grade cohorts (for example, NLSY97 respondents born between 1980 and 1984 who are also enrolled in tenth grade in the 1997-98 school year can be viewed as a reasonable approximation of the sophomore cohort of 1998). In addition to data from youth (and for NLSY97, parent) interviews and a baseline cognitive aptitude measure, the NLSY surveys include considerable linked school and transcript data. While, beyond the standard classification variables, only a small subset of items can be said to be comparable between the studies, there remains a deep complementarity. For example, both studies provide an

[5]For further information about NELS:88, see the base year (1988) through fourth follow-up (2000) data file user's manual: Curtin, Ingels, Heuer and Wu (2002), or see the NCES web site: *http://nces.ed.gov/surveys/nels88/*.
[6]For information about ELS:2002 see the base year data file user's manual (Ingels, Pratt, Rogers, Siegel and Stutts, 2004). Also see *http://www.nces.ed.gov/surveys/els2002/*.
[7] Interestingly, although grade cohorts in their pre-postsecondary genesis, the NCES studies are closer to birth cohorts in a postsecondary educational context. In other words, while they supply a sample of entrants to higher education of similar age (and thus may function as a quasi-birth cohort), they do not supply a sample of, e.g., all beginning postsecondary students, since late entrants into the postsecondary system are not included. This fact mutes some of the differences between the NCES and BLS (NLSY) studies post-high school.

important basis for studying high school dropouts as well as alternative completers such as those who hold the GED. However, the NLSY data provide a longer time frame, and continuous event history-format labor market data, while the NCES studies provide a larger sample and a richer picture of the student's school behaviors and attitudes and of the school context that may have influenced students to drop out or obtain equivalency certification instead of a regular high school diploma. Probable later labor market outcomes for HS&B or NELS:88 students can be modeled from the NLSY data. The use of the two study series in conjunction to effectively build on and extend them both is interestingly illustrated by the work of Kilburn, Hanser and Klerman (1998). NLSY79 examined the relationship between AFQT scores and probability of military enlistment. The U.S. Department of Defense wanted to re-estimate its enlistment models with more current data (1992) as opposed to the original NLSY79 data from the1980 administration of the ASVAB/AFQT. Based on NELS:88 math and reading test results, AFQT scores were estimated for the NELSL:88 1992 senior cohort so that NELS:88 could be used to replicate the earlier enlistment decision studies that were based on NLSY79. While the AFQT was renormed in 1997 using the NLSY97, the same techniques applied to the NELS:88 senior cohort could be applied to ELS:2002 2004 seniors to supply the NLSY with an additional "pseudo-cohort" for military enlistment modeling purposes. Other national probability samples as well may provide comparison points.[8]

1.2 Specific ways in which the NCES high school panel studies have evolved.

1.2.1 The quest for the ideal grade cohort starting point. In NLS-72, senior-year test scores were used as a control variable and predictor as analysts studied the transition from high school into postsecondary education, the labor market, and family formation. In HS&B, with a sophomore cohort, test scores were also an outcome – gain scores for the span from tenth to twelfth grade could be computed for major achievement domains – and related to curriculum exposure and to school processes, while controlling for background and prior achievement. By measuring change within the same schools at two points in time, school effects could be analyzed. Also by starting with sophomores, a population of future dropouts could be identified, and studied in the follow-up rounds.

Despite the success of HS&B, at the time of NELS:88, the decision was made to begin even earlier, at eighth grade. There were several advantages. One was to capture the population of early dropouts, those who leave school prior to tenth grade, in addition to the later dropouts (that is, after tenth grade, as had been studied in HS&B). A second advantage was the opportunity to obtain a premeasure of achievement just prior to the transition to high school.

There were, however, also disadvantages to an eighth grade starting point, quite apart from the large expense, and the loss of statistical efficiency attendant upon the need, for cost reasons, to drastically subsample (the alternative being to conduct surveys in a sample of more than three thousand high schools). The most significant disadvantage of the NELS:88 design was that the eighth grade starting point offered a weak basis for studying high school effectiveness. Clusters of tenth grade students averaged 30 participants in HS&B schools but only 15 for NELS:88 high schools in 1990 and but 11 for seniors two years later (1992). Nor were these clusters of students strictly representative of students in the school. Nor were the high schools themselves nationally representative (only the eighth grade school sample was drawn by probability methods; the high schools were simply those to which the eighth grade cohort had dispersed). A special High School Effectiveness Study attempted to address some of these weaknesses of the NELS:88 design for school effects research, with mixed success.

[8]For example, major national studies of high school seniors, employing test and survey measures, were conducted in 1960 (Project Talent) and 1965 (the Equality of Educational Opportunity Survey, Coleman et al. 1966) (see Schrader and Hilton in Hilton [ed.] 1992 for a discussion of comparability issues); also, the high school graduating classes of 1975-present have been surveyed by Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, a key source of trend data on, in particular, drug use and associated factors. (The study added 8th and 10th grade cohorts in 1991). Items that are strictly comparable across such data sets are, however, uncommon.

It is perhaps then unsurprising that the latest installment in the series, ELS:2002, has returned to the HS&B sophomore cohort design, and takes as its starting point tenth graders within a nationally representative sample of schools. On the other hand, the quest for an ideal terminus or end point may be more problematic still. While the student questionnaire elicits information about expected occupation at age 30, only one of the studies has been taken to age 30. The three completed studies all reflect different end points, and thus also affect comparability (modal ages at last data collection vary as follows: 24 for the HS&B senior cohort, 26 for NELS:88, 28 for the HS&B sophomore cohort, and 32 for NLS-72).

Freshening to maintain relevance for additional grade cohorts. While HS&B contained two distinct cohorts, 1980 sophomores and seniors, NELS:88 began with a single cohort of eighth graders. Two years later most eighth graders were high school sophomores, and four years later, seniors. Nevertheless, without further additions to the sample, the future NELS:88 samples would not have been comparable to HS&B sophomores or to NLS-72 and HS&B seniors. In order to permit intercohort analysis, the NELS:88 eighth-grade cohort sample was freshened. At the time of the 1990 first follow-up, through an application of the half-open interval procedure (see Kish 1965), tenth-graders who were outside the sampling frame in 1988 (either because they were not eighth graders at the time or because they were not in the country) were given a chance of selection into the study. This freshening procedure was repeated in 1992 to achieve a nationally representative sample of seniors.[9] In this way, NELS:88 comprises three distinct but overlapping cohorts: 1988 eighth graders, 1990 tenth graders, and 1992 twelfth graders. Given the success of the NELS:88 freshened-cohort approach, sample freshening has also been instituted in ELS:2002, so that it too will have both a sophomore and a senior cohort – without incurring the enormous expense occasioned in HS&B by drawing two distinct samples for sophomores and seniors.[10]

Oversampling for rare policy-relevant populations. One notable change over time is increasingly ambitious oversampling of subgroups, both at the level of school sector and student race and ethnicity. The study series has several precision requirements that drive the subdomain sample sizes. The requirement with the largest impact on sample size is that the study be able to detect a 15 percent change in proportions across waves with 80 percent power using a two-tailed alpha = 0.05 test. This requires a minimum subgroup respondent sample size of 1,356 cases, and because response rates are seldom much over 90 percent and longitudinal studies face attrition over time, a yet larger number of cases must be pursued. While samples for some subdomains can be met without oversampling, for rarer populations, oversampling is a must.

NLS-72 included but did not oversample Catholic and other private schools. In HS&B, Catholic schools were oversampled to facilitate comparison with public schools. Some non-Catholic private schools were also selected, including a small superstratum of elite private high schools.

In NELS:88, the proportion of private schools was further increased, with stratification by sector following four major groupings. Schools were stratified into public, Catholic, independent schools

[9] By virtue of having a 1992 senior cohort, NELS:88 also was able to effect a link, via test equating, to the 1992 12th-grade NAEP (that is, NELS:88 mathematics results were put on the NAEP scale, as will occur for ELS:2002 seniors in 2004 vis-à-vis the 2005 NAEP math assessment).

[10] In fairness to HS&B, it should be said that the two-cohort approach served a special objective. The study was to provide three ways to look at school effects – following 1980 sophomores to 1982; comparing 1980 sophomores and seniors in the same schools to each other; and comparing 1980 seniors to the 1972 senior cohort of NLS-72. (Coleman et al. 1979, p.177). Comparison of sophomores and seniors in the same school provides a pseudo-longitudinal design on the assumption that the two within-school cohorts are essentially similar – and provides insights into school effects more quickly than waiting for the actual longitudinal design within a single cohort to be effected. A like strategy has been used for test development in the NELS:88 and HS&B field tests. For example, in the 1987 NELS:88 field test, the proposed item pool was tested on eighth, tenth, and twelfth graders, so that items could be chosen for eighth grader that were, based on these simulated longitudinal results, likely to show substantial change in later rounds.

(members of the National Association of Independent schools) and other private schools. In HS&B, 12 percent of schools were private (84 Catholic and 38 other private), while in NELS:88 nearly 23 percent were non-public (104 Catholic and 133 non-Catholic private schools). In ELS:2002 as well 23 percent of the school sample is Catholic (95) or other private (77).

In High School and Beyond, school sampling was used as a device for increasing the number of Hispanic students. HS&B selected a sample of both public and Catholic high schools in which Cuban Americans were disproportionately represented, as well as a high-Hispanic public school sample that targeted Puerto Ricans, the goal being to support separate Hispanic subgroup analyses of students with ancestral origins in Cuba, Puerto Rico, and Mexico.

In NELS:88, racial/ethnic oversampling was done within-schools rather than at the school level. (One problem with a strategy of first selecting schools with high concentrations of a particular group to supplement those who would already fall in the sample is that it tends to skew the weights). In addition to continuing to oversample Hispanics, in NELS:88, Asians were oversampled as well. In ELS:2002, this within-school racial/ethnic oversampling strategy was continued.

High school(and postsecondary) archival data: gathering administrative records. The immense value of school transcripts as objective, reliable measures of crucial aspects of students' educational experiences is widely recognized. Early national transcript data sets include the Study of Academic Prediction and Growth (1969) conducted by ETS (in public schools only) and transcripts of 1975-81 high school graduates, as collected for the Bureau of Labor Statistics' NLSY79 (1979 Youth Cohort of the National Longitudinal Survey of Labor Market Experience).[11] Analysis of these data sets (Adelman 1983) provided critical information for the National Commission on Excellence in Education, in reviewing trends in coursetaking. While the value of transcript collections had been recognized before the first NCES high school transcript study, understanding of the importance of high school transcripts was magnified by the methodological work conducted with the HS&B data by Fetters, Stowe and Owings (1984). In comparing student self-reports in questionnaires to transcript data, Fetters, Stowe and Owings suggested the sometimes substantial degree to which transcripts are superior to student self-reports of exposure to learning situations in level of detail, accuracy and completeness. (For example, using the transcript as the validity standard, the correlation of student grade reports to school records was a reasonably high .77. Correlations range from .63 to .70 for amount of coursework in mathematics and science, with seniors reporting, on average, that they had taken about one semester more of mathematics coursework than could be found on their final transcripts. Correlations were lowest in social studies and English at .39 and .28). In 1987, five years after the HS&B high school transcript study and using the same coding scheme (the CSSC) to ensure comparability, NAEP initiated a series of high school transcript studies, with further transcript collections in 1990, 1994, 1998, and 2000. NELS:88 included a transcript component in 1992.[12] ELS:2002 is scheduled to collect transcripts late in 2004. The result is not just a series of transcript studies that illuminate the other data collected in the various studies, but also a series of transcript studies that can themselves provide a basis for trend comparisons of course taking.[13] In addition to the high school transcript studies, postsecondary education transcript collections began with NLS-72 and the HS&B senior cohort in 1984, and have continued through NELS:88 (2000/2001).

[11] Because the 1969 ETS transcript study and the NLSY79 did not use the coding scheme consistently employed by the NCES transcript studies (or, as did the NLSY79's successor study, the NLSY97, the taxonomy to which the codes are typically aggregated for analysis), the two earlier studies are not readily comparable to the transcript studies conducted in 1982 and thereafter.

[12] For a detailed account of cross-cohort transcript comparison issues, see Ingels and Taylor (1995), and Alt and Bradby (1999).

[13] See, for example, the following transcript-based trend studies, that make use both of the longitudinal cohorts and NAEP: Hoachlander 1992, Tuma 1996; Legum, Caldwell, Davis, Haynes, Hill, Litavecz, Rizzo, Rust and Vo 1998; and Levesque 2003.

Gathering more contextual data. Although NLS-72 had no parent survey, questionnaires were administered to a subsample of parents in HS&B. Fetters, Stowe and Owings (1984) compared student and parent reports on family background, using the parent report as the standard of validity. Validity coefficients for student responses to family background items ranged from .20 to .90. Among relatively low validity coefficients for sophomores were items such as "other language spoken in the home" (.50), family annual income (.50), mother's occupation (.44), and the presence of various household items meant to form an index related to SES (from .21 to .39). Not surprisingly, the design for NELS:88 (and later, ELS:2002) called for conducting a parent survey for all base year student participants, to gather critical information about family background and the home education support system that the parent was best positioned to supply. While one may have doubts about comparing student-reported family income (HS&B) with parent-reported income (NELS:88 or ELS:2002), this is arguably an instance in which issues of data quality properly overrode considerations of maintaining the strictest comparability to a past survey.

In addition to a parent survey, HS&B included a teacher comment form, on which teachers could rate participating students, and a school administrator questionnaire. More comprehensive teacher surveys were included in NELS:88 (and the ELS:2002 base year). In addition to again collecting teacher ratings of the specific student, in NELS:88 information was also gathered to try to get at classroom effects. The student's specific learning environment was investigated by identifying curricular contents, instructional practices and objectives specific to a given classroom in which NELS:88 sample members were enrolled. However, it is not clear that the NELS:88 design (three teacher surveys, but with gaps for grades 9 and 11) or items (particularly on instructional goals) were the best suited for this purpose (see Hoffer and Moore 1996, p.iv and *passim*; and Burstein et al. 1995). The ELS:2002 teacher survey does not gather classroom-level data.

ELS:2002 has two additional components not included in any of the prior studies: a library media center questionnaire, and a facilities checklist (the checklist is completed by field interviewers and provides an assessment of the school's physical plant and environment, including feature related to order and safety). The utility of the new components will be assessed in coming months as ELS:2002 base year and first follow-up data are analyzed.

Technological innovation in data collection and processing. While there is some continuity in means of data capture for in-school group administrations – in-school questionnaire and test sessions continue, as in the past, to be administered in paper-and-pencil formats, with (starting with HS&B) optical scanning of the data – out of school administrations have moved from mail questionnaires to computer-assisted designs, such as computer-assisted-telephone interviewing (CATI) and computer-assisted-personal-interviewing (CAPI). The upcoming second follow-up round (2006) of ELS:2002 is scheduled to employ three data capture modalities: CATI, CAPI, and self-administration by means of a web-based survey. Processes such as coding, too have been influenced by computerization, with, for example, automated lookup tables and real-time on-line coding of occupations and other data elements. While by and large these technological innovations are cost effective means to improve data quality, they also, inevitably, raise comparability questions for time series analysis. Another respect in which computer technology has transformed such studies in recent year – and in a way that creates no problems for intercohort comparability – is through improvements in delivering survey information to researchers in a form they readily can use. The early studies were dependent on the use of magnetic tape media on mainframe computers. In recent years the data sets have been disseminated in the form of electronic codebooks (ECBs) on CD-ROMs. With the ECB, the data user can analyze the survey data using a personal computer.

Part 2: Types and comparison points for intercohort comparison

At the student level, three kinds of cross-cohort comparison are possible: time-lag, fixed time, or longitudinal. Each is explained below.

(1) Cohorts can be compared on an *intergenerational* (or *intercohort*) *time-lag* basis. For example, ELS:2002 2004 results (when restricted to sample members who are seniors) can be regarded as the fourth in a series of repeated cross-sections of twelfth graders. That is to say, ELS:2002 seniors in 2004 can be compared to NELS:88 seniors in 1992, HS&B base year seniors in 1980, and to NLS-72 seniors in 1972. Likewise, ELS:2002 sophomores two years later (2004) and NELS:88 1990 sophomores two years later (1992) can be compared (both as dropouts and as students) to HS&B 1980 sophomores in 1982. Most of the intercohort analyses of the NCES data sets have been of this type. Some examples include the following: comparison of senior cohorts (Fetters, Brown and Owings 1984; Green, Dugoni and Ingels 1995); cross-cohort comparison of dropouts (Kaufman, McMillen and Sweet 1996); and comparison of sophomores (Rasinski, Ingels, Rock and Pollack 1993; Wang, Schiller and Plank 1997), as well as transcript comparisons based on graduating seniors (Levesque 2003).

(2) *Fixed time comparisons* are also possible (though seldom pursued by analysts of these data sets), in which groups within each study are compared to each other at different ages though at the same point in time. Thus NLS-72, HS&B senior cohort and HS&B sophomore cohort sample members could all be compared in 1986, some 14, 8, and 6 years after each respective cohort completed high school. Another common time point is 1992. Thus, one might, for example, compare the 1992 educational expectations of the HS&B sophomore and NELS:88 cohorts to explore how 17-18 year olds differ from 27-28 year olds in this respect. Or one might utilize the 1992 life values responses (questions concerning the importance to the respondent of being successful in work, having lots of money, having strong friendships, and so on) to compare HS&B Fourth Follow-Up sophomore cohort members with NELS:88 Second Follow-Up survey participants. In the HS&B base year, with two cohorts within the same school, it made sense, on the assumption that the senior cohort would fundamentally resemble the sophomore cohort two years later, to model change and school effects based on a fixed time comparison of base year sophomores and seniors, then to confirm conclusions in the first follow-up, based on longitudinal data for the sophomore cohort.

(3) Finally, *longitudinal comparative analysis* of the cohorts can be performed by modeling the history of grade (or age) cohorts. Baltes and Nesselroade (1979) argue that there are two streams of change in transition and development, one individual and the other evolutionary or historical, so that the sequential longitudinal study of multiple cohorts is a necessity. Comparative modeling of the history of multiple cohorts may help to distinguish cohort effects from educational and life-stage effects. Within the limitations of a single longitudinal study, such control for cohort effects is not possible. An abbreviated example of multiwave intercohort analysis, in which a part of the focus is cohort differences in educational expectations as they change between sophomore and senior year of high school (including differences in cohort change viewed separately for Whites and Blacks), is provided by Morgan (1996). Another example of longitudinal intercohort analysis, in which HS&B and NELS:88 sophomores and seniors are employed to study the academic trajectories of immigrant youth, is supplied by Glick and White (2003). With the release of the NELS:88 postsecondary education transcripts, some particularly exciting longitudinal comparative analyses of the educational histories of college-bound HS&B and NELS:88 cohort members will be possible. (For example, one can examine the antecedents of eventual postsecondary educational attainment, using high school transcripts to plot four years' coursetaking and grades, summary transcript measures such as grade point average and class rank, high school test scores and questionnaire data, in conjunction with all postsecondary work undertaken in the ensuing 8.5 year period after graduation – thus effectively comparing the educational histories of the two cohorts across the years from fall term 1988 to the end of calendar 2000, and fall term 1978 to the end of calendar 1990. Such an analysis could build on Adelman 1999).

The three kinds of intercohort comparisons are summarized in Table 1 below:

Table 1: Types of Possible Trend Comparisons (NLS-72, HS&B, NELS:88, ELS:2002)

I. Cross-Sectional Comparisons

A. Cross-Cohort Time-Lag Comparisons

1.1980/1990/2002: high school sophomores at three points in time[14]

2.1982/1992/2004: high school sophomores compared two years later

3.1978-82/1988-1992/2000-2004 High School Coursetaking of three sophomore cohorts (1980, 1990, 1992) based on grades 9 – 12 high school transcripts.

4.1972/1980/1992/2004: seniors at four points in time[15]

5.1972/1982/1992/2004: High School Seniors; Adjustment for non-representativeness of 1982 senior sample[16]

6.1974/1982(1984)/1994/2006: high school seniors two years out of high school

7.1984/1994: High school sophomores four years later

B. *Fixed-Time Comparison*

NLS-72 1986 (fifth follow-up, 14 years out of high school), HS&B 1986 senior cohort (third follow-up, 6 years out of high school), HS&B 1986 sophomore cohort (third follow-up, 4 years out of high school)

HS&B 1992 (fourth follow-up, ten years out of high school) versus NELS:88 1992 (second follow-up, modal grade = high school senior)

II. Longitudinal Intercohort Comparisons

Longitudinal comparative analysis of the five cohorts can be performed by modeling the educational (or life course) history of the cohorts.

[14]Must exclude all NELS:88 students who are non-sophomores and all non-students (dropouts).

[15]Must exclude all NELS:88 second follow-up dropouts (including alternative completers), early graduates, and students because they were not spring term 1992 twelfth graders.

[16]NELS:88 conditions as above (seniors only); HS&B must exclude dropouts and non-seniors and statistically adjust for non-representativeness of senior sample.

Possible Time Points for Comparative Analyses

Institution-level comparisons. Comparisons are not limited to cohorts of individuals; not just the student samples, but also the baseline school samples of ELS:2002, NELS:88, HS&B, and NLS-72 are nationally representative, and considerable data have been collected about school-level characteristics. Comparison points include NLS-72 (1972), HS&B (1980), and ELS:2002 (2002) high schools. (The NELS:88 base year school sample was limited to eighth grades.[17]) Also, the HS&B and ELS:2002 school samples, though nationally-representative only in the base year, can also be compared as panels: 1980 high schools two years later (1982) and 2002 high schools two years later (2004).

Table 2: Nationally-Representative School Samples

	Representative School Sample	Non-Representative School Sample
NLS-72	1972	
HS&B	1980, 1993	1982, 1984[18]
NELS:88	1988	1990, 1992
ELS:2002	2002	2004

Individual-level comparisons. In Table 3 comparison points are highlighted. However, with technical adjustments, comparability can oftentimes be achieved even when age/grade/stage parallelism has not been strictly maintained.[19] In addition, survey rounds that coincide with a grade-representative sample are noted by an asterisk. Thus, for example, HS&B (sophomore cohort) in 1980, NELS:88 in 1990, and ELS:2002 in 2002 are nationally representative samples of sophomores. The NELS:88 sample was freshened to make it representative of the nation's sophomores (1990) and seniors (1992). Sample freshening was not conducted in HS&B; its sophomore cohort does not constitute a valid probability sample of the nation's 1982 seniors. Nevertheless, the 1982 HS&B sophomore cohort and 1992 NELS:88 can be compared, for both examine a nationally representative sample of sophomores two years later--consisting of students (most, but not all of them, seniors), early graduates, and dropouts.[20] HS&B 1982 seniors can also be compared to 1972 NLS-72, 1992 NELS:88, and 2004 ELS:2002 seniors, though not without some sample and statistical adjustments.[21] In addition, transcript-based comparisons are possible (see Tables 4-5 for postsecondary transcripts, Tables 6-10 for high school transcripts).

[17]However, the 1988 NELS:88 eighth-grade school sample might be compared to other data sets, such as the ongoing series of NCES Schools and Staffing Surveys or the annual Common Core of Data for public schools and biennial Private School Survey.

[18]A probability subsample of the 1980-1982 HS&B schools was resurveyed in the 1984 Administrator and Teacher Survey. In an institution-level longitudinal follow-up, these schools were re-surveyed in 1992, as part of the National Longitudinal Study of Schools (NLSS). Unlike HS&B in 1982 and 1984, NLSS freshened the HS&B school sample to make it nationally representative of public and private secondary schools in the United States in 1992.

[19]See, for example, the account by T.L. Hilton and J.M. Pollack on estimating postsecondary enrollment change over time using NLS-72 fourth follow-up (conducted over 7 years after graduation) and HS&B third follow-up (conducted just less than six years after high school graduation) data, in Hilton (ed.) 1992.

[20]There are a number of special definitional issues in comparing NELS:88 and HS&B dropouts. For a detailed discussion of these issues, see *Conducting Trend Analyses: HS&B and NELS:88 Dropouts,* (Ingels and Dowd, 1995).

[21]Specifically, out-of-sequence students (non-seniors) and non-students (such as dropouts and early graduates) must be removed from the HS&B analysis sample, and an adjustment made for the exclusion of students who were seniors in 1982 but were not part of the HS&B base year sampling frame, that is, 1982 seniors who were not 1980 sophomores in

Table 3: Comparison Points

	NLS-72	HS&B-So	HS&B-Sr	NELS:88	ELS:2002
Students					
G8				1988*	
G10		***1980****		***1990****	***2002****
G12	***1972****	***1982***	***1980****	***1992****	***2004****
G12 + 1	1973				
G12 + 2	***1974***	***1984***	***1982***	***1994***	***2006****
G12 + 4	***1976***	***1986***	***1984***		
G12 + 5					
G12 + 6			1986		
G12 + 7	1979				
G12 + 8				2000	
G12 + 10		1992			
G12 + 14	1986				
Dropouts					
G10 - G12		***1982***		***1992***	***2004***
follow-up		***1984***		***1994***	***2006***
		(1986, 1992)		(2000)	(???)
Early Graduates		***1982***		***1992***	***2004***
Parents of Seniors[22]		***1980***		***1992***	
Parents of Sophomores		***1980***		***2004***	

the U.S.A. A simplifying assumption here would be that in results and characteristics, these out-of-sequence 1982 seniors are essentially similar to the HS&B 1980 sophomores who failed to progress in the modal grade sequence.

[22]For a crosswalk between the HS&B and NELS:88 parent questionnaires, see Appendix D of the NELS:88 parent survey user's manual (NCES 94-378); for a comparison of the design and implementation of the parent surveys, see section 4.4 of same. A further crosswalk between ELS:2002 and NELS:88 is to be found in the ELS:2002 base year data file user's manual (Ingels, Pratt, Rogers, Siegel and Stutts 2004). The ELS:2002 parent survey encompassed parents of tenth graders; the NELS:88 parent survey parents of eighth and (primarily) twelfth graders and dropouts; and in HS&B a subsample of parents of both sophomores and seniors were surveyed.

*Denotes nationally representative grade sample.

Table 4
NCES Postsecondary Education Transcript Studies

Year conducted/coverage beyond modal high school senior year	Data Source
1984 (HS + 12 years)	NLS-72 PETS
1986-1987 (enrolled postsecondary students)	NPSAS Student Loan Recipient Postsecondary Transcript Survey
1993 (HS + 11 years)	HS&B Sophomore Cohort PETS
1984 (HS + 4 years)	HS&B Senior Cohort PETS
1994-1995 (1993 bachelor's degree recipients)	Baccalaureate and Beyond (B&B)
2000-2001 (HS + 8 years)	NELS:88 PETS

Table 5
Post-High School Graduation Coverage of NCES Longitudinal High School Cohorts Through Postsecondary Education Transcript Studies

	NLS-72 (1984)	HS&B So. (1993)	HS&B Sr. (1984)	NELS:88 (2000)
(fall)	*1972*	*1982*	*1980*	*1992*
G12+1	*1973*	*1983*	*1981*	*1993*
G12+2	*1974*	*1984*	*1982*	*1994*
G12+3	*1975*	*1985*	*1983*	*1995*
G12+4	*1976*	*1986*	*1984*	*1996*
G12+5	*1977*	*1987*		*1997*
G12+6	*1978*	*1988*		*1998*
G12+7	*1979*	*1989*		*1999*
G12+8	*1980*	*1990*		*2000*
G12+9	1981	1991		
G12+10	1982	1992		
G12+11	1983	1993		
G12+12	1984			

Table 6
High School Transcripts Coded in Conformity to CSSC and/or SST

HS&B	NAEP	NAEP	NELS:88	NAEP	NAEP	NAEP	NLSY97	AHAA	ELS:2002
1982[23]	*1987*[24]	*1990*	*1992*	*1994*	*1998*	*2000*	*2000*[25]	*2002*[26]	*2004*

[23] HS&B 1982 graduates are often compared to NAEP 1990 or other graduating classes, but the HS&B sample, unlike NELS:88, was not freshened, and so it falls somewhat short of representing either seniors or graduating seniors in 1982.

Table 7: NCES National High School Transcript Collections

Year conducted	Data Source
1982	HS&B Sophomore Cohort
1987	NAEP HSTS
1990	NAEP HSTS
1992	NELS:88
1994	NAEP HSTS
1998	NAEP HSTS
2000	NAEP HSTS
2004	ELS:2002
2005 (scheduled)	NAEP HSTS

Table 8: Non-NCES National High School Academic Transcript Collections

Year conducted	Data Source
1969	Study of Academic Prediction
1980-1983	NLSY79
2000-2005	NLSY97
2001-2002	AHAA (Add Health)

Table 9: Target Populations for NCES High School Academic Transcript Samples

Target Populations of NCES Transcript Samples	Data Source
Spring Term 1980 Sophomores 2 Years Later	HS&B
1985-1986 juniors who remained in their 1985-86 schools and graduated in academic year 1986-1987	1987 HSTS
8th grade cohort 4 years later; 10th grade cohort 2 years later; 1992 12th grade cohort; all cohorts based on membership in spring term	NELS:88
Graduating High School Seniors	HSTS: 1990, 1994, 1998, 2000, 2005
Spring Term 2002 Sophomores 2 years later; Spring Term 2004 Seniors	ELS:2002

[24] As in the HS&B case, this sample is close to, but not exactly, a sample of graduating seniors. Strictly speaking, the sample represents those 1985-86 high school juniors who remained in their 1985-86 schools and graduated in academic year 1986-87.

[25] The NLSY97 did not code to the CSSC as did the other 8 transcript studies, but did code to the SST-R (Secondary School Taxonomy, 1998 Revision [Bradby and Hoachlander 1999]). For comparability to NLSY97, HS&B, NAEP and NELS:88 data needs to be aggregated up to the more general level represented by the taxonomy. In addition to using the same taxonomy, in order to further ensure comparability, NLYS97 followed the NELS:88 and NAEP 1998 procedures and file structures as much as possible. Although NLSY97 is an age cohort, information available about grade can be used to select a subset of cases that represent a grade cohort. For example, if we assume that nearly all graduating seniors are between 16 and 20 years of age in the year of their graduation, then the subset of NLSY97 youth graduating in 2000 properly represents the high school class of 2000.

[26] The Adolescent Health and Academic Achievement Study (AHAA) – a substudy of the National Longitudinal Study of Adolescent Health (Add Health) – collected high school transcripts (2001-2002) for its originally 7th-12th grade cohorts, and coded them using the CSSC and standard NCES procedures.

Table 10: Eligibility and Exclusion for NCES High School Academic Transcript Collections

Eligibility/Exclusion in NCES Transcript Studies	**Data Source**
Severely disabled and non-English students excluded	HS&B
No students excluded	HSTS: 1987, 1990, 1994, 1998, 2000, 2005
Severely disabled and non-English-proficient students excluded	NELS:88
No students excluded	ELS:2002

Part 3: Threats to change measurement

3.1: Sample design and definition issues.

3.1.1. *Multiple independent samples*. Repeated cross-sections compound sampling error. In an intercohort comparison, each cohort is derived from an independent sample. Therefore change measurement must contend with the fact that differences in multiple sample means will in part be a function of the sampling errors associated with each independent sample. (In contrast, a longitudinal sample is drawn but once, and intracohort analysis does not compound sampling error.) This factor in time series measurement from sample surveys should be noted, but there is nothing to be done about it.

3.1.2 Eligibility and exclusion. Not all students are able to meaningfully respond to research instruments such as the assessments and questionnaires administered in the four studies. Some students are too limited in their English language proficiency to do so, while others may be precluded from participation by a physical or mental disability. HS&B excluded as ineligible students with such barriers to participation, although an overall exclusion rate has not been documented. In NELS:88, 5.3 percent of the base year eighth-grade sample was excluded for such reasons (this figure is similar to the exclusion rate for eighth grade in the National Assessment of Educational Progress [NAEP] in similar subjects in the same period). However, a sample of the NELS:88 ineligible students was followed over time, and some students whose status subsequently changed were incorporated into the 1990 and 1992 rounds. In ELS:2002, no students were classified as ineligible as such, though some were exempted from completing the questionnaire or test, and others were tested under circumstances in which they were provided with special accommodations. Because of the accommodations and increased efforts to work with school personnel to include any student who could at least complete the questionnaire, the overall rate of instrument-exempted sophomores in ELS:2002 is quite low, just under 1 percent. Contextual information was collected for these individuals. Of course, the success of ELS:2002 in this regard may lead to some loss of comparability of results.

There are two interesting questions here, first, how excluding some students affects overall estimates for a cohort; second, how different eligibility and inclusion rules across cohorts may affect change measurement. For both of these questions, it is relevant to get some sense of the degree to which these missing students are atypical. If information is "missing at random" it does not necessarily bias adjusted estimates, but if the excluded students differ from other students, and if different eligibility rules are used over time, then the impact on both cross-sectional and cross-cohort estimation may be serious. The two tables below, based on the NELS:88 experience, provide at least some sense of the characteristics of ineligible students.

Table 11 shows the impact on the base year (1988) to first follow-up (1990) cohort dropout rate, when the ineligibles are excluded from the calculation, and when the sample is expanded to include them. Table 12 compares the excluded group to the expanded sample (that is, both eligible and ineligible eighth-graders) in terms of their 1992 high school enrollment status, and grade (that is, whether they are in the modal grade progression of their cohort, or have fallen behind. Both tables suggest that there may be differences between the included and excluded samples that may affect estimates.

Table 11: Bias Estimates for Eighth Grade Cohort Dropout Rate, 1988-1990. (Percentage of Spring Term 1988 Eighth Graders Not in School Spring Term 1990)

	ELIGIBLE SAMPLE		EXPANDED SAMPLE		BIAS
Total.	6.0 %	(0.48)	6.8 %	(0.40)	**-.8**
Race/Ethnicity.					
Asian	3.1	(1.05)	4.0	(1.02)	**-.9**
Hispanic	9.2	(1.01)	9.6	(0.84)	**-.4**
Black	10.0	(1.94)	10.2	(1.51)	**-.2**
White	4.9	(0.53)	5.2	(0.44)	**-.3**
Sex.					
Male	6.3	(0.69)	7.2	(0.55)	**-.9**
Female	5.8	(0.59)	6.5	(0.51)	**-.7**
1988 Eighth Grade Public School Students.	6.8	(0.55)	7.6	(0.45)	**-.8**

Note: Standard errors appear parenthetically after each estimate. Two small subgroups do not appear under race/ethnicity. One such group, race unknown, comprised about 2 percent of the unweighted expanded sample. The second group, American Indians, comprised just over 1 percent of the sample.

Source: National Education Longitudinal Study of 1988 (NELS:88) First Follow-Up, National Center for Education Statistics, public use file and expanded cohort file; Ingels (1996), NCES 96-723.

Could the fact that a larger proportion of the student population was included in ELS:2002 (95 percent of the potential cohort in NELS:88 as contrasted to 99 percent in ELS:2002) affect cross-cohort estimates of change? Given that we know that the excluded students in NELS:88 tended to be quite different from the included, the answer to this must be affirmative. At the same time, there are ways to make the samples at least somewhat more comparable. Thus while for optimal cross-sectional estimation, all the ELS:2002 cases might be used, for comparison of achievement results across cohorts, perhaps the ELS:2002 cases that reflect testing accommodations should be dropped. In the same way, adjustments are commonly made to render the HS&B and NELS:88 transcript studies comparable to the NAEP high school transcripts. Specifically, only the subset of the HS&B or NELS:88 senior cohort that in fact graduated is included, while graduates on the NAEP file with special education diplomas are excluded from analysis (see Alt and Bradby 1999, Levesque 2003).

Table 12: 1992 School Enrollment Status of the Base Year (1988) Ineligible Sample: Summary

	Expanded Sample	BYIs
ENROLLMENT		
In School, spring 1992	83.3	62.4
Dropout, spring 1992	11.6	30.0
Alt. Completer, spring 1992 (dropped out but receiving alternative instruction)	5.2	7.6
IN SEQUENCE		
Yes	80.0	57.6
No	20.0	42.4

SOURCE: NELS:88 Second Follow-Up Survey (1992), National Center for Education Statistics, U.S. Department of Education; Ingels (1996), NCES 96-723.

3.1.3 Sampling, oversampling, and sample efficiency.

Differences in sampling rates, sample sizes, and design effects across the studies also affect precision of estimation and power of generalization. Asian students, for example, have been oversampled in NELS:88 and in ELS:2002, but not in HS&B, where their numbers were quite small. Also, although Catholic schools were oversampled in three of the four studies, HS&B had few (only 38) private non-Catholic schools, and NLS-72 few non-public schools. The base year (1980) participating sample in HS&B numbered 30,030 sophomores. In contrast, 15,362 sophomores participated in the base year of ELS:2002. Cluster sizes within school were much larger for HS&B (on average, 30 sophomores per school) than for ELS:2002 (just over 20 sophomores per school; larger cluster sizes are better for school effects research, but carry a penalty in greater sample inefficiency). Mean design effect (a measure of sample efficiency that can be related to effective sample size[27]) also is quite variable across the studies: for example, for tenth grade, 2.9 for HS&B, 3.9 for NELS:88 (reflecting high subsampling after the 8th-grade base year), with the most favorable design effect, 2.4, for ELS:2002.

[27] Effective sample size can be quite different from the nominal sample size; effective sample size is more meaningful than raw sample size in terms of statistical analysis -- for example, the sampling variance of a mean standard score is equal to the reciprocal of the effective sample size, not the reciprocal of the raw sample size. Effective sample size may be defined as the ratio of the raw sample size divided by the design effect.

3.2 Instrumentation: Content and format issues

3.2.1 Content and format issues: the test batteries.

Test content in the four study series is summarized in Table 13 below. There has been some constriction in content over the years.

Table 13: Assessment Subjects in the Longitudinal High School Cohorts, 1972-2004

Study and Year Conducted	Test Subjects
NLS-72 1972	Vocabulary, reading, math, inductive reasoning, memory and perception
HS&B 1980 Senior Cohort, 1980	Vocabulary, reading, math, picture number, mosaic comparison, visualization in 3 dimensions,
HS&B 1980 Sophomore Cohort, 1980	Vocabulary, reading, math, science, writing, civics
HS&B 1980 Sophomore Cohort, 1982	Vocabulary, reading, math, science, writing, civics
NELS:88 1990	Reading, math, science, social studies
NELS:88 1992	Reading, math, science, social studies
ELS:2002 2002	Reading, math
ELS:2002 2004	Math

Before directly focusing on the tests used in the various high school longitudinal cohorts, a brief digression on the National Assessment of Educational Progress (NAEP) may provide some context concerning how another prominent time-series assessment program has wrestled with similar issues. For one thing, NAEP maintains a long-term trend sample in addition to the national NAEP. The long-term trend assessment re-uses the same historical pool of items. This means that the content framework for the long-term assessment has diverged from the main assessment, because only the latter is aligned to changes in the curriculum over time (NAGB 2002). Nonetheless, a point can be reached when an historical item pool ceases to be relevant, a situation that apparently has been reached for the science assessments, though the mathematics and reading long-term trend assessments are continuing at this time.

A particular object lesson in trend measurement is supplied by NAEP's 1986 experience of its "reading anomaly." Forsyth et al. (1996) provide a summary:

> Subsequent investigations (Beaton & Zwick, 1990) showed that the anomaly could be traced to seemingly inconsequential changes in booklet configuration, administration procedures, item context, and post-stratification procedures—each one designed to provide better information—yet which made it impossible to compare results across assessments.

Beaton and Zwick offered the moral, "When measuring change, do not change the measure." But this is of course easier said than done, and like so many essential truths, is in need of some qualification (which it is one of the purposes of this paper to provide). Forsyth et al. (1996) conclude:

> Hundreds of seemingly small facets of the configuration could be tweaked to produce modest improvements in estimation within one time point, but these changes could have a greater impact on the overall results than actual differences over a two-year period in what students know and can do.

Based on a core of common items (supporting an IRT-based equating) or an equipercentile procedure, some, though incomplete, linkages have been created for the NLS-72 to ELS:2002 test batteries (see Table 14 below). The 1972-1980-1982 seniors were put on the same (i.e., the NLS-72) scale (in reading and mathematics) for the "High School Excellence Study" (Rock, Ekstrom, Goertz, Hilton and Pollack 1985). The HS&B sophomore math scores were put on the NELS:88 sophomore scale (Rasinski, Ingels, Rock and Pollack 1993), and ELS:2002 sophomore results in both math and reading have been put on the NELS:88 scale. Further equating will take place as the data from the ELS:2002 first follow-up are prepared for release.

Table 14: NCES Linked Test Scores for the Longitudinal High School Cohorts, 1980-2004

Base Test	Linked Tests
HS&B 1980 math (G10)	NELS:88 1990 math, ELS:2002 2002 math
NELS:88 1990 reading (G10)	ELS:2002 2002 reading
NELS:88 1992 math (G12)	ELS:2002 2004 math
NELS:88 1992 math (G12)	NAEP 1992 math
ELS:2002 2002 reading (G10)	PISA 2000 reading
ELS:2002 2002 math (G10)	PISA 2003 Math
ELS:2002 2004 math (G12)	NAEP 2005 math

The tests have evolved over time, in ways that may somewhat affect comparability. In NLS-72, seniors marked answers on an answer sheet (separate from the test booklet) while in 1980 and 1982 (HS&B) and NELS:88, answers were marked in the test booklet. (In ELS:2002, answers were marked in the base year test booklet, but in the first follow-up, on a separate answer sheet.) The HS&B format of inclusion of answers as an integral part of the test booklet is thought to have given a modest advantage to HS&B test takers (see Rock, Hilton, Pollack, Ekstrom, and Goertz, 1985, for further details). Other differences between the NLS-72 and the HS&B/NELS:88 tests include improved mapping in the latter tests and the procedure of blackening an oval versus blackening a box (Hilton, 1992, cites a study by Earles, Guiliano, Ree and Valentine, that indicates such format differences are significant for speeded tests, accounting for about one half a standard deviation in difference of result. Since the changes to be measured in high school achievement are comparatively small, the potential of artifact to overwhelm the measurement of true change is indeed a concern.[28]

Nevertheless, the biggest differences in test administration concerns efforts in NELS:88 and ELS:2002 to make the tests more adaptive, that is, to tailor them to the student's ability level. This was done in the NELS:88 follow-ups by assigning test form on the basis of the prior round ability estimate (theta). Adaptiveness was achieved in ELS:2002 by having a two-stage base-year test (the first stage a routing test, the second stage tailored to the test takers ability as measured on the routing test). In the ELS:2002 first follow-up, the prior round ability estimate is being used for form assignment. A more adaptive assessments improves estimation in the later studies, but it also makes for some interestingly misleading intercohort comparisons. For example, the correlation between SES and achievement seems to

[28] Consider that in NELS:88, gain in mathematics between eighth grade and tenth grade was just under half a standard deviation in base year units while gains in reading over the two years were about a third of a standard deviation (Scott, Rock, Pollack, and Ingels 1995); the difference between the math achievement of sophomores in 1980 and in 1990 was just less than a quarter of a standard deviation (Rasinski, Ingels, Rock and Pollack 1993). In this context, the NAEP experience, as documented by Beaton and Zwick (1990) -- in which the impact of (*inter alia*) item context and format differences across assessments over time was apparently larger than the trend effects that were being measured – is readily understandable.

have been increasing in the United States in recent years (e.g., based on comparison of HS&B results to NELS:88), although this higher correlation is an artifact of the higher reliabilities for the tests. The tests became more adaptive in 1990; the effect of the adaptiveness can be seen in Table 15 below.

Table 15: Reliabilities for mathematics and reading tests, 1972-2002[29]

Study	*Reading*	*Math*
1972 (NLS-72)	.79	.86
1980 (HS&B)	.79	.85
1988 (NELS:88)	.80	.89
1990 (NELS:88)*	.86	.93
1992 (NELS:88)*	.85	.94
2002 ELS:2002*	.86	.92

*Semi-adaptive test. The 1990 and 1992 assessments were based on multiple forms of varying difficulty, with form assignment dependent on the prior-round ability estimate (*theta*). The 2002 assessment was two stage, the first stage a routing test, results of which determined the level of difficulty of the second-stage form to be assigned.

3.2.2 Content and format issues: the questionnaires.

3.2.2.1 Classification variables: conceptual and linguistic change.

While one might expect the most basic or standard of the classification variables to remain reasonably stable – and this has indeed been the case for some (such as sex or gender) – many of the variables have been subject to change. Change is of two kinds. In one case, the underlying concepts remain intact, but the label has changed. In other instances, the conceptualization itself has changed with time.

An example of a change, not in underlying concept but in its label and in nuance, is provided by descriptions of students with disabilities. In NLS-72, descriptors such as "a cripple" were used, whereas in HS&B, "orthopedic handicap" was referred to. Indeed, terminology further shifted from one of "handicapped" students in HS&B to students with "disabilities" in NELS:88 and ELS:2002. Likewise there were changes in race labels (for example, from Negro in NLS-72, to Black in HS&B, to Black or African-American in later studies).[30]

[29] Sources: Rock, Hilton, Pollack, Ekstrom, and Goertz 1985; Rock and Pollack 1995; Ingels, Pratt, Rogers, Siegel and Stutts 2004.

[30] This is not to say, in these cases, that there is no change in underlying social meaning when, for example, a race label changes – see Smith 1992 for a discussion of some of the implications for social meaning of changed racial labels – but it is not a change that invalidates use of the new label in conjunction with the old, that is, it still can be mapped to the same group over time.

Certainly survey and census data serve as a reminder of the degree to which race is an historically and culturally conditioned concept, ever mutable, equivocal, and contested. Martin, DeMaio and Campanelli (1990), reflecting on racial classifications used in the U.S. Census Bureau between 1850 and 1990, note that although many people tend to think of race as a stable and enduring characteristic, "no single set of racial categories has been used in more than two censuses, and most were only used once." The 2000 Census was to be no exception.

In HS&B and NELS:88, students were asked to mark one race only. In light of the 2000 decennial census and revised race-reporting guidelines issued by the Office of Management and Budget (OMB), a new race category was added, and, more important, students were allowed to mark all applicable races, thus generating a further category, multiracial.

The new race category is Native Hawaiian or Other Pacific Islander. For purposes of cross-cohort comparisons, cases identified in ELS:2002 as "Native Hawaiian or Other Pacific Islander" can be combined with the category "Asian" to achieve comparability with HS&B and NELS:88. However, for students who considered themselves to be multiracial and marked more than one race, there is no ready means to map them back into a one-race scheme. With 5 race categories, and values based on a single race reported, none reported, the 10 possible combinations of 2 races, 10 possible combinations of 3 races, the 5 possible combinations of 4 races, and the possibility of a combination of all 5 races, there are 32 separate race categories. When race is crossed by ethnicity (race by Hispanic or not Hispanic), there are 64 possible race-Hispanic ethnicity combinations.

Ideally, race would have been asked two ways in the ELS:2002 base year, both allowing multiple races, and forcing a choice of one. Competition for space on the questionnaire did not permit additional race questions. (It is also a matter of uncertainty whether respondents would react well to being forced to take two such conceptually different views of the nature of racial identification. A major impetus for the multiracial category was the resistance of many multiracial individuals [a resistance that was certainly encountered in NELS:88] to be forced into the choice of a single race). Under the circumstances, the change does to some degree disrupt trend analysis along one of the most important of its dimensions, racial classification. It is impossible to know, for example, whether a student who marked White and Black in ELS:2002 would have marked White, or have marked Black, or have refused to give a response, in NELS:88, in which only one race was allowed. There are over 700 non-Hispanic multiracial sophomores recorded in the ELS:2002 base year data set, but the distorting effect on cross-cohort estimation is likely to be greatest for small population subgroups with many claimants to multiple race, such as the American Indian category

While in the main, key classification variables have been constructed *to the extent possible* in the same way across studies, there are some deviations, which raise the issue of whether differences are desirable, as a means of updating or maintaining the meaning of the construct – or whether the introduction of differences threatens cross-cohort replication and comparability. The critically important socioeconomic status variable (SES) will serve as an illustration. Continuities and differences in SES constituents and construction in the three prior studies are summarized below in Table 16. Table 17 summarizes the elements comprising the SES measure in ELS:2002.

Table 16. Socioeconomic composite, the National Education Longitudinal Study of 1988 as compared to the National Longitudinal Study of the High School Class of 1972 and the High School and Beyond longitudinal study

NLS-72, HS&B (student reported)	NELS:88 (parent reported)	NELS:88 student survey substitutions
Father's occupation	Father's occupation	Father's occupation
	Mother's occupation	Mother's occupation
Father's education	Father's education	Father's education
Mother's education	Mother's education	Mother's education
Family income	Family income	Household items
Household items	—	

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002).

Table 17. Socioeconomic composite, the Education Longitudinal Study of 2002

Preferred source (parent reported)	Student report substitution if missing from parent	Imputed if still missing
Father's occupation	Father's occupation	Father's occupation
Mother's occupation	Mother's occupation	Mother's occupation
Father's education	Father's education	Father's education
Mother's education	Mother's education	Mother's education
Family income	—	Family income

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002).

ELS:2002 largely follows the NELS:88 model above in that in both studies the composite is based on five equally weighted, standardized components: father's education, mother's education, family income, father's occupation, and mother's occupation. Parent data are used to construct this variable. Student data are substituted where parent data are missing. However, for parent education and occupation, where both parent and student reports are missing, ELS:2002 education and occupation values are imputed. Unlike NLS-72 and HS&B, family income was not asked of students in NELS:88 or ELS:2002. While in NELS:88 a student-provided household item index, which served as an income proxy, was substituted when income data were missing, a different procedure was followed in ELS:2002. When parent data on income were missing, income was statistically imputed.

Some differences in the constituents of the SES composite reflect changing social circumstances. For example, many fewer mothers worked in 1972. The importance of gathering information about maternal occupation increased with the passage of time and the increasing labor market participation of American females.

While in general time series studies need to be conservative about item changes, there are times when items must change in order to continue to be comparable and to have some chance of measuring change. The household items index was originally one of the five components of SES. In NELS:88, however, it was replaced by maternal occupation. The household items were re-asked, however, so that the item scale could be used as a proxy for income (when parent-reported income was missing). Apart from the changing role of the household items scale, a further issue is the need to change the specific items in order to continue to consistently measure the same construct. The item list, then, provides an

excellent example of why, at times, change may be necessary, and keeping things the same the impediment to true change measurement.

For NLS-72, owning a color television discriminated between people of various income levels (see Table 18 below). By the time of HS&B, 8 years later, this was no longer so. By 2002 (arguably even earlier, since by the end of the 1980s home computers had begun to supplant typewriters), HS&B items such as ownership of a typewriter had ceased to function as good proxies for family income, while other items, such as access to the Internet or having a digital video disc player did.[31] Of the ten 1972 items, only three remained on the index for the ELS:2002 base year: receiving a daily newspaper, a regularly received magazine, and an electric dishwasher. The items that in the 2001 ELS field test were most highly correlated with income were, in order beginning with the highest: electric dishwasher, internet access, fax machine, computer, and daily newspaper. While items differ across the index over time, in each case the items are those that are needed to provide a measure that has a reasonable correlation with income.

[31] The household items were asked in ELS:2002, but the index was not used in the creation of SES, since missing income data were imputed.

Table 18: Household Item Index for SES

NLS-72	HS&B	NELS:88	ELS-FT
Newspaper	Newspaper	Newspaper	Newspaper
Dictionary	> 50 books	Dictionary	Dictionary*
Encyclopedia	Encyclopedia	Encyclopedia	Encyclopedia*
Magazine	--	Magazine	Magazine
Record Player	Place to study	Place to study	Place to study*
Tape Recorder	Room of own	Room of own	Room of own
Color TV	Calculator	Calculator	Calculator*
Typewriter	Typewriter	Typewriter	Internet access
Electric dishwasher	Electric dishwasher	Elec. dishwasher	Elec. dishwasher
Two cars	Two cars	> 50 books	>50 books
		An atlas	Fax machine
		Clothes dryer	Clothes dryer
		Washing machine	Washing machine*
		Microwave oven	Microwave oven*
		Computer	Computer
		VCR	VCR*
			DVD player
			Digital camera*

*Denotes item dropped from the ELS:2002 main study questionnaire, owing to a low point biserial correlation with income in the field test. For the final version of the ELS:2002 household item index, based on 2001 field test data, the Pearson correlation of the 10-item scale to parent-reported income was .458.

Another aspect of SES also represents an area where change over time may have to be contended with and may require updating of features of the survey is in the area of occupations and their relative prestige. Occupations can come into or go out of existence. The occupational structure of a nation may change. And the prestige attributed to various occupations may change as well. To accommodate the factor of prestige change, two sets of prestige scores were drawn upon in NELS:88: the 1961 Duncan

socioeconomic indicator measure that had been employed in NLS-72 and HS&B, as well as a 1989 revision by Nakao and Treas (1992). The same strategy has been employed in ELS:2002. While use of new prestige scores can successfully update the studies over time, changes in broad occupation rubrics, to reflect a changing occupational structure and the growth of new technologies, may pose the more unsettling choice of either continuing with an increasingly obsolete scheme, or adopting a new one better aligned to the times.

A further issue that arises in connection with SES is how to treat the definition of the relevant family unit itself. Should same sex couples in a parenting role, too rare to worry about in 1972 but becoming far more frequent thirty years later, count the same as a mother and father in the construction of SES? We would argue that in this instance, maintaining comparability over time argues for relaxing the definition of "mother and father" to include same sex dual parentship[32]

But the final issue of interest to our inquiry in connection with SES has to do with the use, mandated by new agency statistical standards, of imputation of missing values for key variables. The consequences of imputation for ELS:2002 are two. First, there is no longer a need for an only moderately correlated proxy for income – imputed income should replace the household index in the construction of SES. Second, since all the constituents of SES are subject to imputation, it is now possible to create an SES composite with no missing data, and this has been done in ELS:2002. For the HS&B sophomores, SES was missing for around 9 percent of the participants, and for NELS:88 (in 1990) just under 10 percent. The availability of imputed variables (including both key classification variables and achievement test scores) also poses a novel question for analysts interested in intercohort comparisons. Since imputed values are flagged, it is the analyst's choice whether or not to employ them.

If theory is any guide, the imputed key variables should decidedly have the effect of improving cross-sectional estimation. On the other hand, since imputation was not used in the prior studies, it is also possible that use of ELS:2002 imputed values might decrease comparability of results across studies. To explore the issue of the magnitude of the effect of imputation on analysis, NCES is sponsoring methodological analyses in which (a) the SES composite will be computed both ways, that is, with the household items index as in NELS:88 and with imputed income as in ELS:2002; and (b) various cross-cohort estimates will be run both ways, that is, with imputed and unimputed. In this way, at least the magnitude of the effect of these changes can be quantified.

3.2.2.2 *Response and format effects.* Embedding methodological studies in field tests is essential to improving data quality. Yet implementing the results is yet another source of difference as response formats are changed. Field test experiments that help measure and categorize errors contribute to increased understanding of the strengths and weaknesses of a study, and enrich its documentation. Within a time series, field test experiments that are acted upon, that is, become the basis for changed forms or procedures, raise difficult issues of the tradeoffs between better data quality and strict

[32] While we have emphasized changes in the sociolinguistic meaning or labeling of concepts central to the standard classification variables, manifestations of schooling itself may change in ways that affect cross-cohort comparability. One example of this is the increase in numbers of year-round schools, which must be accommodated in sampling, and in their implication for the appropriate timing points for measuring status and growth through the assessment battery. A second example is the recent trend toward increased numbers of home-schooled students. A number of ELS:2002 high school sophomores were in a home school setting two years later. While the numbers may be too small to supply a robust analysis sample, and also unrepresentative (no probability sample of the nation's home-schooled seniors was drawn), such students, to be meaningfully surveyed, may require a different questionnaire, or different question wordings even where the content can match that of the in-school cohort members. In addition, the presence of these home-schooled students in the study may affect comparability, in that they represent a group of students who in most cases would have remained in high school and been resurveyed in the school setting (or as dropouts) in the prior studies.

comparability. Small gains in the former at the price of a great deal of the latter normally should not be pursued, but it is often difficult to weigh specific gains or losses since the effect of changes is cumulative and ultimately requires a model of their total impact.

As a follow-up to Sudman and Bradburn (1982), who argue against "mark all that apply" formats because of the difficulty in interpreting the data and the likelihood of more missing data, an experiment was embedded in the NELS:88 second follow-up field test to assess the effect of "mark all that apply" question instructions as contrasted to explicit "yes" or "no" response options. Results are reported in Rasinski, Mingay and Bradburn (1994). Significantly fewer response options were selected with the mark-all-that-apply instructions, which may be interpreted (though not unequivocally) as indicating such formats are more likely to lead to incomplete data.

In past questionnaires, both kinds of formats had been used. While the field test experiment does not resolve the consistency issue – whether or not to change historically "mark all that apply" formats – it does permit one to measure the cost in data quality of doing so, and to enter the appropriate caveats about the meaning and limitations of the data.

3.2.2.3 *Order and context effects.* The order of presentation of individual items or modules of items may affect response (see Schuman and Presser 1981; Sudman, Bradburn and Schwarz 1996) and, thereby, comparability if consistency of item order across cohorts is not maintained. Consistent order of questionnaire modules and items has not always been achieved across the four high school cohort studies, though any effect on estimates is unknown. This issue could be explored in future field tests. Context effects are another issue that has received little attention in the study series, and that might properly be the subject of future field test experiments. Does one maintain cross-cohort comparability simply by repeating a question, or must one strive to repeat the entire context in which the question was embedded, and which gives it special nuances of meaning? The existing literature (see, for example, Tourangeau, Bradburn, Rasinski and D'Andrade 1989) suggests that such effects are not likely to be a problem for most factual items, though more of a concern for attitudinal and Likert-style agree-disagree response scales. While one should, conservatively, strive to preserve both question order and context to the extent possible in the repetition of items across both waves and cohorts, it would also be useful to experimentally investigate contextual effects for some typical items from the study series.

3.2.2.4 *Changing times and privacy concerns: what can be asked, with whose consent?* Some changes in content over time, that have affected the trend comparisons that can be made with the study series, reflect a changing climate of consent arrangements and judgments about what questions are proper to ask adolescents, particularly when their parents have not given express written consent. Both HS&B and NELS:88 collected a considerable amount of information about tobacco, alcohol, and illicit drug use among high school students, and these data have been extensively analyzed over the years. They also collected some information about sexual attitudes (for example: would you consider having a child if you weren't married?; Which of the following is your most important source of information about methods of birth control?) and whether students had gotten in trouble with the law. Finally, HS&B and NELS:88 inquired into student religious affiliations and degree of religiosity. Given the interest in comparing Catholic and public schools, it was deemed important to be able to identify Catholics in non-Catholic schools and non-Catholics in Catholic schools. However, none of these topics have been included in the ELS:2002 student questionnaires.

A good example of recent legislation that has influenced the questions that some surveys will pose (and specifically, trend items that were asked or dropped on ELS:2002) is the Protection of Pupil Rights Amendment (PPRA). PPRA is a federal law that affords certain rights to parents of minor students with regard to surveys that ask questions of a personal nature. This law requires that explicit written consent be

obtained from parents before minor students are asked to participate in a U.S. Department of Education survey that requires into any of the following areas:

1. Political affiliations;
2. Mental and psychological problems potentially embarrassing to the student and his/her family;
3. Sexual behavior and attitudes;
4. Illegal, anti-social, self-incriminating and demeaning behavior;
5. Critical appraisals of other individuals with whom respondents have close family relationships;
6. Legally recognized privileged or analogous relationships, such as those of lawyers, physicians, and ministers;
7. Religious practices, affiliations, or beliefs of the student or student's parent; or
8. Income (other than that required by law to determine eligibility for participation in a program or for receiving financial assistance under such program.)

Previously, posing these questions to voluntary survey participants, required only implicit parental consent. Because of the large expense to the Government of obtaining explicit parental consent, and the likelihood, even granting the extra funds needed to mount such efforts, of lowered response rates and introduction of biases, questions falling into the eight areas listed above have been dropped from ELS:2002, in particular questions from the prior surveys about tobacco, alcohol, and drug use, problems with the law, religiosity and religious affiliation, and sexual attitudes. To the accommodations to a changing reality that may threaten portions of a time series data collection, then, must be added the matter of more stringent consent requirements.[33]

3.2.2.5 *The advancing field*. Educational research does not stand still. New findings, revised models of the educational process, and new statistical methodologies for data analysis (that may impose new requirements on the data collected), all may argue for certain changes in content, each time a new longitudinal cohort is launched. At the same time, commonality of content is required for replication and comparison. To be sure, there is a kind of implicit rotation through of topics and items, in which some become obsolete and are dropped, creating space for new questions – but there is also need to maintain some common core of items over time. The current policy – to actively involve members of the research community, representatives of which participate in a Technical Review Panel at the time of instrument design – seems the most prudent way to address the issue of which topics should be added, and which dropped, although even with the best of advice, it remains a difficult area within a cross-cohort design. An especially challenging aspect of honoring advances is seen in the fact that many times groundbreaking research takes place on a very small scale, or in a qualitative dimension, with uncertainty about how to translate it reliably to a large-scale quantitative setting.

3.2.2.6 *The changing policy agenda*. Policy questions change as well. Sometimes new policy questions arise owing to changes in the nature of education itself (a good example of this is the extensive introduction of informational technology, particularly computers, into the learning environment –

[33] Changes over time in the stringency of consent procedures pose a dilemma. Certainly consent protections are of paramount importance. At the same time, there is uncertainty about the practical meaning of consent under different scenarios. The experience of most survey researchers is that when explicit consent forms are not returned, this seldom truly means that the person contacted indeed has an objection to the survey. Explicit consent procedures tend to create an ambiguous residual category of cases that have neither granted nor denied permission, and that are difficult, even after great effort and expenditure, to wholly resolve. Another set of issues is brought up by Singer (2003). Singer reports a methodological experiment that indicated that there were substantial numbers of respondents who were willing to participate in a survey but not if an explicit consent form was required, while at the same time, many people signed explicit consent forms without a real understanding of their content. Since the willing participants who were not willing signers of the form were not missing at random, Singer concluded that explicit consent increases the likelihood of statistical biases.

educational technology was not a major issue for NLS-72 yet, perforce, has substantial coverage in the ELS:2002 questionnaires). In other cases, changes in policy interests reflect the evolution of policy debates and shifts in the political landscape. New longitudinal studies within a sequential cohorts design need to be able to guess correctly about whether policy changes represent ephemeral fashions, or serious long-term shifts that must be accommodated if the study series is to maintain its relevance.

3.3 Data collection and processing: methodology issues

3.3.1 Response rates, nonresponse, and sample attrition.

Differences in response rates, and in nonresponse adjustment procedures, may make for differences that affect comparability. As Martin (1983) points out, there is also an issue of the nature of the nonrespondent population (e.g., historical shifts in nonresponse from individuals who were unavailable, to individuals who refuse). Even though in general the magnitude of survey nonresponse has increased in recent years, at least it may be said that the NCES longitudinal high school cohort studies have generally had outstanding response rates from individual participants, nor have there been notable declines over time, nor has there been a notable shift in the nature or reasons for nonresponse. However, there is less cooperation from schools, primarily because such studies must compete with greatly increased mandated testing. The problems with school cooperation have had two implications. First, school response rates have dropped (e.g., fewer schools agreed to participate in ELS:2002 than in HS&B or NELS:88). Second, in order to gain school cooperation, the burden of the study has had to be reduced. Thus ELS:2002 collects substantially less test and questionnaire data than did its predecessors.

3.3.2 Mode effects

Survey responses can be influenced by the mode of questionnaire administration. The most basic contrast is between self-administration, and interviewer administration by telephone or in person. Interviewer administrations allow for more probing on the part of the interviewer, as well as more questioning for clarification on the part of the interviewee. For sensitive behaviors, particularly topics subject to positive or negative social desirability biases,[34] self-administration is generally thought to be the best method of data collection. The interviewer effect as a source of bias is well-documented – it is stronger in face-to-face interviews than telephone, but in either, socially undesirable behaviors may be under-reported and socially desirable behaviors over-reported compared to self-administration (Tourangeau, Rips, and Rasinski 2000).

A second source of mode effect is technology (computerized versus not). Compared to interviews that are not computer-assisted, any computerized modality – whether CATI or CAPI (or Web self-administration) – affects data quality, and in a specific direction, that is, it improves it. Prior round information or sampling frame data can be preloaded. This affords the opportunity to provide the respondent with memory-stimulating prompts, or to disambiguate seeming contradictions. CATI and CAPI also minimize skip pattern error by automating the interview, and permit more complex branching questions to be asked. They make the interview more conversational by helping it to flow within a logical context. And, CATI and CAPI provide for immediate inter-item consistency checks and range checks, as well as for error resolution. Any non-computerized self-administration will typically have more missing data than a computerized (or interviewer) administration. But in a trend series, the good may be an enemy of the better, since, in cross-cohort comparisons, an apparent change over time may instead be an artifact of improved data quality.

[34] These can include such NELS:88 questions as use of drugs, or, on the parent questionnaire, how much the child is left alone – an item on which, in NELS:88, parents provided a far lower estimate than did their children.

A third source of mode effect is the medium of apprehension, that is whether the interview depends on aural or visual comprehension. It may make a difference whether the interview is read and seen, or heard. Interviewer accent may have an effect in face-to-face and telephone surveys. While interviewer administrations typically involve interviewer speech, this does not mean that visual cues may not be used. For face-to-face interviews, sometimes showcards are employed. Calendars (to anchor temporally-bound items) and lists (such as lists that are long [such as the detailed precoded occupational categories used in the NCES high school cohort studies) or require rank ordering) may be mailed to a respondent prior to a CATI interview. In the non-CATI telephone interviews that took place in HS&B and NELS:88, respondents normally had the hardcopy questionnaire in hand while being interviewed over the phone. In the CATI interviews (1994 and 2000 for NELS:88), the interview experience was purely auditory with no visual aids.

Mode effects can be an issue within rounds and across rounds in a longitudinal study; but likewise mode effects can be an issue across cohorts. For example, if we look at the data capture methods for the follow-up round two years out of high school, we see the following:

Table 19: Data Collection Modalities, HS + 2 years

Cohort:	**NLS-72**	**HS&B Sr.**	**HS&B So.**	**NELS:88**	**ELS:2002**
Year:	*1974*	*1982*	*1984*	*1994*	*2006*
Modes:	SAQ	SAQ	SAQ	CATI	Web
	PAPI	PAPI	PAPI	SAQ	CATI
				PAPI	CAPI

In other words, for the studies in general, potential mode effect problems are pervasive, in that they portend whether one is doing a one-point-in-time cross-sectional analysis, an across-rounds or intracohort analysis, or an intercohort analysis. The NLS-72 and both cohorts of HS&B used the primary methodology of a self-administered questionnaire (SAQ) that was returned in the mail. However, secondarily, an interviewer-administered paper-and-pencil interview (PAPI) was also employed. For example, in the 1982 follow-up of seniors, some 75 percent of the cohort returned the self-administered mail questionnaire and an additional 19 percent completed the questionnaire through either in-person or telephone interviews, while 6 percent were nonrespondents. To minimize mode effects, respondents who completed the questionnaire by telephone were required to have a copy of the questionnaire in front of them while doing so. (This measure no doubt helped stabilize content and format, though it does not address the problem of increased likelihood of social desirability biases in responses that are associated with interviewer administration.) In NELS:88, the primary modality of data collection was computer-assisted telephone interviewing (CATI). However, for sample members without telephone service or in other special circumstances, a hard copy version of the interview was administered in a face-to-face interview or was self-administered. In effect, there were two questionnaires for NELS:88, since the complex skip logic of the CATI interview could not be contained within the confines of a paper instrument. The plan for ELS:2002 in 2006 is to have but a single electronic instrument, which will be administered in three modalities: through self-administration on the Web, or through interviewer administration over the telephone (CATI) or as a computer-assisted personal interview (CAPI).

Given the pervasiveness of the mode effect issue for the longitudinal high school cohorts – within rounds, across rounds, across cohorts – it is surprising that little research has been devoted to this issue, and perhaps especially surprising in the context of the thorough treatment other domains of measurement error have received in NCES studies (see, for example, Salvucci, Walter, Conley, Fink and Saba 1997 for

a recent summary). Mode of administration is recorded in the data set so that analysts always know whether a given instrument was administered in one modality or another, but no methodological conclusions can be drawn from the data set, since individuals were not randomly assigned to a given mode of administration. True mode effect experiments would be of value in future field tests.

3.3.3 House effects

One of the features in which the United States government's statistical system differs from those of Canada and many European and Asian nations is in its use of private contractors. Much of the data collection and processing, and even design, test score scaling, and analysis, is out-sourced to the private sector. To be sure, the Bureau of the Census does some studies for other federal agencies (in the case of NCES, the Schools and Staffing Survey). But the bulk of the work is done by private for-profit (such as Westat, Abt Associates, Mathematica Policy Research) or private not-for-profit (such as Research Triangle Institute [RTI International] , National Opinion Research Center [NORC at the University of Chicago], American Institutes for Research [AIR]) organizations outside the government.

There is some evidence of effects attributable to which particular survey house is conducting a given data collection. Smith (1982), for example, finds differences, many of them based on differences in field interviewer training, between the General Social Survey and the American National Study. Cohen and Potter (1988) examined a national study in which data collection was split between two of the major survey contractors. Even given a commonly defined context of detailed specifications from a single client for administering an identical questionnaire, statistically significant variations in estimates were seen. Seemingly very subtle (though pervasive) variations in survey house organizational structure, field supervision, staffing, mix of other projects, or interviewer training, can produce somewhat different measurements. It is unknown to what extent there may have been house effects in the case of the sequential longitudinal cohorts of high school students examined in this paper. However, major changes in data collector have occurred (for example, for NLS-72 at the final round, from RTI to NORC; for NELS:88 in its final round, from NORC to RTI). It remains an interesting and unanswered question whether house effects have occurred in such cases.

3.3.4 Evolving statistical standards and techniques

Statistical techniques, and agency standards, also may change. A significant example of standards change is provided by ELS:2002. Missing data for key estimates are now imputed. In the prior studies, they were not. The analyst is faced with a choice between using an improved cross-sectional estimate, or using instead a theoretically poorer estimate (i.e., unimputed) that was constructed in a manner more consistent with the practice in earlier rounds. NCES-sponsored work in progress will compare ELS:2002 imputed and unimputed estimates and attempt to measure the impact of imputation on bivariate statistics within the time series. The analysis will also measure the effects of the different way (imputation) of dealing with missing income (and other) data in the construction of SES. An example of changing statistical techniques is provided by methods for nonresponse adjustment of weights. In NLS-72 and HS&B, weighting cells were constructed based upon the known characteristics of the sample units. In ELS:2002, rather than a weighting cell approach, a more recent technique, to wit, propensity modeling, was used. In NELS:88 a mix of the two approaches is encountered (propensity at the school level, weighting cell at the student level). However, results of nonresponse adjustment tend to be highly correlated regardless of method. These differences, therefore, should not lead to greatly different estimates (Kalton and Cervantes 2003; Rizzo, Kalton and Brick 1996). Nevertheless, the larger point is that any difference in weighting technique or other statistical method must be evaluated for its impact on cross-cohort comparability, and that many small differences may collectively have a larger impact.

3.3.5 Impact of confidentiality/disclosure editing

Protecting the confidentiality of data is of paramount importance, particularly data that contain information about specific individuals or institutions. Standards in this area have evolved over time. Disclosure risk analyses were not performed for the earlier studies, but were begun in earnest with the base year of NELS:88. The base year of ELS:2002 took the process some steps yet further, with the use of such techniques as "data swapping" in addition to data coarsening and perturbation based on a disclosure risk analysis. In data swapping, some variables for a sample case that has been paired with another case will be exchanged. By so doing, even if a tentative identification of an individual is made, because every case in the file has some undisclosed probability of having been swapped, uncertainty remains about the accuracy and interpretation of the match.

There is a double question here of the impact of such confidentiality edits to the data. The first question pertains to the implications of applying confidentiality edits to some but not all of the data sets in the series, and the second question is how much these edits change estimates from their original form. While perturbation techniques such as swapping do result in changes in estimates generated from the data, before-and-after weighted distributions and correlations for swapped ELS:2002 variables strongly suggest that after applying the disclosure limitation techniques, the analytic utility of the data files had not been compromised. While the issue of the total impact of many small changes in combination is a further question for the study series, there is good reason to be sanguine that any adverse impact of confidentiality editing can be kept to an absolute minimum.

Part 4: Summary and Conclusions

Accurate trend measurement faces several challenges. Sampling error tends to be more of a problem for intercohort comparisons than for intracohort, since there is sampling error each time an independent sample is drawn. Differences in two sample means estimated from independent samples will be a function not only of the real differences in means, but also the sampling errors associated with both measurements. Hence small (but not therefore necessarily unimportant) differences may be harder to detect.

In estimating trends based on results from three or more[35] sample surveys, a number of nonsampling errors also may arise. Differences in instrument format, question order, content stability (items added or dropped), and wording, as well as data collection mode or methodology and results, are potential sources of nonsampling error. While the requirements of change measurement dictate that the same measures be repeated in the same way, there are also strong disincentives to holding measures and methodologies constant. The goals, the subject, and the technology of education measurement do not remain static. The educational policy agenda changes over time; the manner and matter of education changes as curriculum content and instructional methods are revised; improvements arise--in survey methodologies, data capture technologies, and in measurement techniques--that promise large benefits if implemented. Language and social concepts change as well.

A fundamental question is what can be done to ensure that maximum intercohort comparability can be achieved. As Martin (1985) points out, trend studies should maximize the chances that they can be replicated by both employing explicit standardized definitions and procedures, and by thoroughly documenting all procedures. Fortunately for the NCES longitudinal high school cohorts, a good deal of attention has been paid to documentation, from the very start over thirty years ago. Something that may have been lacking in the beginning, but that has become increasingly important to both standardizing procedures and to ensuring that that the standards are high, is the successive iteration of explicit statistical standards by the agency. The latest edition of the Standards (Seastrom 2003) brings them to a new level of usefulness and sophistication. While the explicit codification of statistical standards at this level comes thirty years too late for NLS-72, such documents, reflective of the most serious deliberation about the state of the art in survey and statistical work, can provide profoundly valuable guidelines for identifying the changes that are essential to the continuing integrity of study series such as these.

Methodological work in field tests has been productive, but the results of field test experiments often pose the standard dilemma of whether to make a change that may improve one-point-in-time cross-sectional estimation or longitudinal results, but at a cost to trend analysis. (Indeed, since trend analysis is only one use of the high school cohort studies, decisions about changes are all the harder to make.) Certainly further methodological work might be directed to several areas, such as mode of administration, context and order effects, that have as yet unmeasured possible implications for the studies. It would be valuable for instrument developers to know in the future, for example, to what extent the core of items to be repeated across the cohort studies are subject to context effects or other distortions.

Finally, where major changes appear to be necessary in a time series data collection, it is highly desirable to follow the same path that is often used in assessment programs – building a bridge study so that the impact of the change can be evaluated.

[35] A minimum of three data points normally will be required for an analysis to detect trends.

REFERENCES

Adelman, Clifford. (1983). *Devaluation, Diffusion and the College Connection: A Study of High School Transcripts, 1964-1981*. Report to the National Commission on Excellence in Education. Washington, DC: American Enterprise Institute for Public Policy Research.

Adelman, Clifford. (1999). *Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment*. Washington, DC: U.S. Department of Education.

Adelman, Clifford. (2004). *Principal Indicators of Student Academic Histories in Postsecondary Education, 1972-2000*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Available: http://www.ed.gov/rschstat/research/pubs/prinindicat/index.html

Alt, Martha Naomi, and Bradby, Denise. (1999). *Procedures Guide for Transcript Studies*. NCES Working Paper Series, No. 1999-05. Washington, D.C.: National Center for Education Statistics.

Baltes, Paul B., and Nesselroade, John R. (1979). In: J.R. Nesselroade and P.B. Baltes, *Longitudinal Research in the Study of Behavior and Development*. New York: Academic Press.

Beaton, Albert E. and Zwick, Rebecca. (1990). *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. Princeton, NJ: ETS, NAEP Report 17-TR-21.

Bradby, Denise, and Hoachlander, Gary. (1999). *1998 Revision of the Secondary School Taxonomy*. NCES Working Paper Series, No. 1999-06. Washington, D.C.: National Center for Education Statistics.

Burstein, Leigh, McDonnell, Lorraine M., Van Winkle, Jeanette, Ormseth, Tor, Mirocha, Jim, and Guiton, Gretchen. (1995). *Validating National Curriculum Indicators*. Santa Monica, CA: RAND.

Cohen, Steven B. and D.E.B. Potter. (1988). *Data Collection Organization Effects in the National Medical Expenditure Survey*. Washington, DC: National Center for Health Services Research and Health Care Technology Assessment.

Coleman, James S., Bartot, Virginia, Lewin-Epstein, Noah, and Olson, Lorayn. (1979). *Policy Issues and Research Design*. Report to National Center for Education Statistics under Contract 300-78-0208 (High School and Beyond). Chicago: National Opinion Research Center.

Curtin, Thomas R., Ingels, Steven J., Wu, Shiying, and Heuer, Ruth. (2002). *NELS:88 Base Year to Fourth Follow-Up Data File User's Manual* (NCES 2002–323). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Duncan, Otis D. (1961). A Socioeconomic Index for All Occupations. In A.J. Reiss (Ed.), *Occupations and Social Status* (109–138). New York: Free Press.

Fetters, William B.; Brown, George H.; and Owings, Jeffrey A. (1984). *High School Seniors: A Comparative Study of the Classes of 1972 and 1980*. Washington, D.C.: National Center for Education Statistics.

Fetters, William B; Stowe, Peter; and Owings, Jeffrey A. (1984). *Quality of Responses of High School Students to Questionnaire Items***.** Washington, D.C.: National Center for Education Statistics.

Forsyth, Robert, Hambleton, Ronald, Linn, Robert, Mislevy, Robert, and Yen, Wendy. (1996). *Report by the Design Feasibility Team to the National Assessment Governing Board on Redesign of the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board. Available: http://nagb.org/pubs/appj.html.

Glenn, Norval D. (1977). *Cohort Analysis*. (07-005). Beverly Hills, CA: Sage.

Glick, Jennifer E., and White, Michael J. (2003). "The Academic Trajectories of Immigrant Youths: Analysis Within and Across Cohorts." *Demography*, *40*(4): 759-783.

Groves, Robert M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Hilton, Thomas L., ed. (1992). *Using National Data Bases in Educational Research.* Erlbaum: Hillsdale, N.J., Hove and London.

Hoachlander, E. Gareth. (1992). *Participation in Secondary Vocational Education, 1982-1987*. (NCES 91-667). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Hoffer, Thomas B., and Moore, Whitney. (1996). *High School Seniors' Instructional Experiences in Science and Mathematics*. (NCES 95-278). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Ingels, Steven J., and Baldridge, John. (1995). *Conducting Trend Analyses of NLS-72, HS&B and NELS:88 Seniors*. NCES Working Paper Series, No.95-05. Washington, D.C.: National Center for Education Statistics.

Ingels, Steven J., and Dowd, Kathryn L. (1995). *Conducting Trend Analyses: HS&B and NELS:88 Dropouts.* NCES Working Paper Series, No. 95-07. Washington, D.C.: National Center for Education Statistics.

Ingels, Steven J., and Taylor, John R. (1995). *Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data*. NCES Working Paper Series, No. 95-06. Washington, D.C.: National Center for Education Statistics.

Ingels, Steven J. (1996). *Sample Exclusion in NELS:88—Characteristics of Base Year Ineligible Students: Changes in Eligibility Status After Four Years* (NCES 96–723). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Ingels, Steven J., Pratt, Daniel J., Rogers, James E., Siegel, Peter H., and Stutts, Ellen. (2004). *ELS:2002 Base Year Data File User's Manual*. (NCES 2004–405). U.S. Department of

Education, National Center for Education Statistics. Washington, DC: NCES; available NCES website.

Kalton, Graham, and Cervantes, I.F. (2003). "Weighting Methods." *Journal of Official Statistics 19*(2).

Kaufman, Philip, McMillen, Marilyn M., and Sweet, David. (1996). *A Comparison of High School Dropout Rates in 1982 and 1992*. (NCES 96-893). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Kilburn, M. Rebecca, Hanser, Lawrence M., and Klerman, Jacob A. (1998). *Estimating AFQT Scores for National Education Longitudinal Study of 1988 (NELS:88) Respondents*. (MR-818-OSD/a). Santa Monica, CA: RAND.

Kish, Leslie. (1965). *Survey Sampling*. New York: Wiley.

Koretz, Daniel M., and Berends, Mark. (2001). *Changes in High Schol Grading Standards in Mathematics, 1982-1992*. (MR-1445-CB). Santa Monica, CA: RAND.

Levesque, Karen. (2003). *Trends in High School Vocational/Technical Coursetaking: 1982-1998*. (NCES 2003-025). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Legum, Stanley, Caldwell, Nancy, Davis, Bryan, Haynes, Jacqueline, Hill, Telford J., Litavecz, Stephen, Rizzo, Lou, Rust, Keith, and Vo Ngoan. (1998). *The 1994 High School Transcript Study Tabulations: Comparative Data on Credits Earned and Demographics for 1994, 1990, 1987, and 1982 High School Graduates, Revised*. (NCES 98-532). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Martin, Elizabeth. (1983). "Surveys as Social Indicators: Problems in Monitoring Trends." IN Rossi, P.H., Wright, J.D., and Anderson, A.B. *Handbook of Survey Research*. Orlando, FL: Academic Press.

Martin, Elizabeth, DeMaio, Teresa J. and Campanelli, Pamela C. (1990). "Context Effects for Census Measures." *Public Opinion Quarterly*, 54(4):551-566.

Morgan, Stephen L. (1996). "Trends in Black-White Differences in Educational Expectations: 1980–1992." *Sociology of Education, 69*(4): 308–319.

NAGB. (2002). *National Assessment Governing Board: Long-term Trend Policy Statement*. Washington, DC: Author.

Nakao, Keiko, and Treas, Judith. (1992). *The 1989 Socioeconomic Index of Occupations: Construction from the 1989 Occupational Prestige Scores*. General Social Survey Methodological Report No. 74. Chicago: National Opinion Research Center, University of Chicago.

Rasinski, Kenneth A., Ingels, Steven J., Rock, Donald A., and Pollack, Judith. (1993). *America's High School Sophomores: A Ten Year Comparison*. (NCES 93-087). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Rasinski, Kenneth A., Mingay, David, and Bradburn, Norman M. (1994). "Do Respondents Really Mark All That Apply?" *Public Opinion Quarterly*, 58:400-408.

Riccobono, John A., Henderson, Louise B., Burkheimer, Graham J., Place, Carol, and Levinsohn, Jay R. (1981). *National Longitudinal Study: Base Year (1972) through Fourth Follow-Up (1979) Data File User's Manual*. U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. (Note: may be downloaded from the International Archive of Education Data (IAED) at the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan *http://www.icpsr.umich.edu*].)

Rizzo, Lou, Kalton, Graham, and Brick, J. Michael. (1996). "A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse." *Survey Methodology*, *22*.

Rock, Donald A., Hilton, Thomas L., Pollack, Judith M., Ekstrom, Ruth B., Goertz, Margaret E. (1985). *Psychometric Analysis of the NLS-72 and the High School and Beyond Test Batteries.* Washington, D.C.: National Center for Education Statistics.

Rock, Donald A.; Ekstrom, Ruth B.; Goertz, Margaret E.; Hilton, Thomas L.; Pollack, Judith M. (1985.) *Factors Associated With Decline of Test Scores of High School Seniors, 1972 to 1980. A Study of Excellence in High School Education: Educational Policies, School Quality, and Student Outcomes*. (NCES 85-217). Washington, D.C.: National Center for Education Statistics.

Rock, Donald A., and Pollack, Judith M. (1995). *Psychometric Report for the NELS:88 Base Year Through Second Follow-Up* (NCES 95–382). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Salvucci, Sameena, Walter, Elizabeth, Conley, Valerie, Fink, Steven, and Saba, Mehrdad. (1997). *Measurement Error Studies at the National Center for Education Statistics*. (NCES 97-464). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Schuman, Howard, and Presser, Stanley. (1981). *Questions and Answers in Attitude Surveys*. NY: Academic Press

Seastrom, Marilyn. (2003). *NCES Statistical Standards*. (NCES 2003-601). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003601

Scott, Leslie A., Rock, Donald, Pollack, Judith, and Ingels, Steven J. (1995). *Two Years Later: Cognitive Gains and School Transitions of NELS:88 Eighth Graders*. (NCES 95-436). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Singer, Eleanor. (2003). "Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits." *Journal of Official Statistics*, *19*(3): 273-285.

Smith, Tom W. (1982). "House Effects and the Reproducibility of Survey Measurements: A Comparison of the 1980 General Social Survey and the 1980 American National Election Study." *Public Opinion Quarterly*, *46*: 54-68.

Smith, Tom W. (1992). "Changing Racial Labels: From Colored to Negro to Black to African American." *Public Opinion Quarterly*, *56*: 496-514.

Spencer, Bruce D., Sebring, Penny A, and Campbell, Barbara. (1987*). The National Longitudinal Study of the High School Class of 1972 (NLS-72) Fifth Follow-Up (1986) Sample Design Report*. (NCES 88-403). Washington, DC: U.S. Government Printing Office. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=88403

Sudman, Seymour, and Bradburn, Norman M. (1982). *Asking Questions*. San Francisco: Jossey-Bass.

Sudman, Seymour, Bradburn, Norman M., and Schwarz, Norbert. (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Tourangeau, Roger, Bradburn, Norman M., Rasinski, Kenneth, and D'Andrade, R. (1989). "Carryover Effects in Attitude Surveys." *Public Opinion Quarterly*.

Tourangeau, Roger, Rips, Lance J., and Rasinski, Kenneth. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Tuma, John. (1996). *Trends in Participation in Secondary Vocational Education: 1982-1992*. (NCES 96-004). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Wang, H., Schiller, K.S., and Plank, S. (1997). "A Comparison of 1980 and 1990 Sophomore Mathematics Achievement." IN James S. Coleman et al., *Redesigning American Education*. Boulder, CO: Westview.