

Slide 1 of 24

Title Slide: Considerations for Analysis of ELS:2002 Data

Slide 2 of 24

This module describes the analytic considerations that should be kept in mind when using ELS:2002 data. Specifically, it describes the ELS cohorts; ELS sample migration between rounds; second follow-up statuses; merging linked data files; High school transcripts; subsampling populations; and item nonresponse issues.

Slide 3 of 24

There are two cohorts in ELS, sophomores, from 2002, and seniors, from 2004. The overlap between the sophomore and senior cohorts is shown in this graph. Specifically, this graph shows which sample members from each of these target groups are included in each segment of the First Follow-up cross-sectional sample.

The amount by which the two cohorts overlap is shown in blue in the graph. As you can see, there were 13,253 members of the sample who were in both the 10th and 12th grade cohorts of ELS.

The freshened group of seniors is shown in yellow on the right. There were 171 12th graders added to ELS in 2002 that were not in the original 10th grade cohort.

The red portion of this cross-sectional sample, shown on the left, represents the 10th grade cohort members who were not 12th graders in the first follow-up (F1), because either they: were dropouts, early graduates, home schooled, or in the 11th grade in 2004.

As shown by the black bracket at the top of the graph, the 2002 sophomore cohort extends from the left-hand side of the graph over to the beginning of the yellow freshened cohort. In ELS, this 2002 sophomore cohort is identified by the G10COHRT (spelled G-1-0-C-O-H-R-T) cohort flag, where G10COHRT=1 for all F1 sample members in the 2002 sophomore cohort. The total size of the G10COHRT is 16,081.

The 2004 senior cohort is all those in the ELS First Follow-up who were in the 12th grade in 2004. This group is identified by the G12COHRT (spelled G-1-2-C-O-H-R-T) cohort flag, where G12COHRT=1 for all members of the senior cohort. This includes all those in the blue and yellow areas of the graph, as enclosed by the black bracket on the bottom of the graph. The total size of the G12COHRT is 13,424.

Slide 4 of 24

As discussed in the module titled, 'ELS:2002 Sample Design, Weights, Variance, and Missing Data,' when using cohort flags within ELS data to define your population of interest, analysts must be mindful of sample member status in each round (also referred to as wave or sometimes follow-up) of data collection. Make sure to consult the data file

Considerations for Analysis of ELS:2002 Data

documentation to ensure accurate analysis. The data file documentation can be accessed by clicking on the underlined screen text, 'Using Cohort Flags.'

Slide 5 of 24

When conducting longitudinal analyses with ELS data, it is important to keep in mind that some students within the sample changed schools between the base year and the first follow-up.

This table shows the extent of school migration, or transfers, in ELS.

Because of student migration, the number of schools in which ELS students were enrolled increased from the original 752 schools in the base year to a total of 1,200 schools in the first follow-up. This resulted in an increase in the number of schools within ELS.

The number of students who migrated was relatively small (about 1,620 out of 15,520), so the average number of students enrolled in the base year schools remained close in the first follow-up to what it was in the base year. The number of students per school decreased from about 20.3 students per base year school to about 16.8 per base year school in the first follow-up. Because of the small number of students who transferred, the average number of students in each transfer school was only about 1.35.

Slide 6 of 24

It is important to remember that if you are conducting an analysis that depends upon school level data from the first follow-up, or F1, administrator questionnaire, there will be no data available because the school administrator questionnaire was only administered in base year schools. Data for transfer students will be missing and designated as such using the -8: "Survey component legitimate skip/NA" reserve code.

Additionally, it is important to remember that there are no school IDs provided for transfer schools on the first follow-up file. Analysts will not be able to link to external data for transfer schools using the first follow-up data file. There is a school file, called BYF1TSCH, that has IDs for the transfer schools, which will facilitate linking to external data sources such as the Common Core of Data (or CCD) and the Private School Universe Survey (or PSS). To fully use these school IDs, a restricted-use license will be required. Within the third follow-up, or F3, data file, some school characteristics variables have been included in both the public- and restricted-use data files. The F3 public-use file can be obtained through eDAT.

Slide 7 of 24

Students also change schools when they graduate from high school and enter college. Between the first and the second follow-ups, many students enrolled in postsecondary education.

Considerations for Analysis of ELS:2002 Data

This table shows the total number of colleges in which ELS students were enrolled in 2006, or at the time of the second follow-up. As you can see, about 12,800 of the ELS students were enrolled in either a 4- or 2-year institution, with an average of about 5 ½ students at each of about 2,500 institutions.

This small number of students clustered within each institution may be smaller than what is needed to support some types of analysis, such as hierarchical linear modeling (or HLM). Analysts should use caution when attempting to conduct such analyses.

Slide 8 of 24

There are some implications of sample migration for purposes of multi-level modeling.

Although the number of students per base year school declined from 20.3 to 16.8 students per school, it may remain large enough for some multi-level modeling purposes.

When used for purpose of multilevel modeling, the sample of 2002 high school sophomores available for this modeling can only include those who remain in their base year schools at the first follow-up. Because of the students who transferred, however, this sample will no longer be nationally representative of 2002 high school sophomores.

Lastly, there may not be enough sample members in each postsecondary institution to analyze student experience nested within institution using hierarchical linear modeling (or, HLM).

Slide 9 of 24

In order to obtain the maximum amount of data from respondents within the time available for the second follow-up, complicated routing procedures were used. The ELS respondent population was divided into six groups, creating six potential respondent types for the second follow-up student questionnaire.

These six respondent types were:

- Regular enrollees, who had enrolled immediately after high school and were still enrolled in postsecondary education at the time of the second follow-up data collection in 2006;
- Delayers were those who had not enrolled immediately after high school but were enrolled at the time of the second follow-up;
- Leavers had enrolled immediately but left by the time of the second follow-up;
- Delayers/Leavers had not enrolled immediately and had left by the time of the second follow-up;
- Non-enrollees had not enrolled as of the second follow-up; and
- High schoolers, the smallest group, were those who were still enrolled in high school at the time of the second follow-up.

Considerations for Analysis of ELS:2002 Data

Sample members were routed through the questionnaire based on their answers to gate questions asked at the beginning of the questionnaire. Accordingly, complex skip patterns were used. It is important to remember that, just like other rounds of ELS, sample members didn't necessarily receive every question on the questionnaire. Different from other ELS rounds, however, sample members may have received very different questions, based on their postsecondary enrollment status at the time of the second follow-up.

For more information about the definitions of each of these groups, see the ELS:2002 Base Year to Second Follow-up Data File User's Manual, which can be accessed by clicking on the underlined screen text, "Second Follow-up."

Slide 10 of 24

This table shows the data collected by respondent type, as different data were collected from sample members depending upon their respondent type.

In addition, different questions were asked of these groups across the four main areas of the questionnaire: Respondents who had not yet completed high school were asked about high school completion. Respondents enrolled in postsecondary education were asked a set of questions about their enrollment history and experiences. Respondents who were working were asked about their employment.

Lastly, respondents of all types were asked life event questions, such as whether the respondent had any biological children. You can click the 'pause' button to continue to view the information presented on this slide if needed. When you are ready to continue, click the 'next' button.

Slide 11 of 24

Let's examine one of the many F2 questionnaire variables available for analysis on the data file. This table shows the frequency distribution of the composite variable, F2PS1LVL (which is the level of the first postsecondary institution attended). This is just one of the many variables that are available for analysis for sample members who were enrolled in a postsecondary institution.

The first three categories in the table provide the basic information about who attends which level of postsecondary education first—a 4-year, 2-year, or less than 2-year—institution. From these data, we can begin to estimate the enrollment rate.

In order to calculate this rate, some analysts might simply recode all of the missing data (coded as -3, -4, -8, and -9 in this instance) to {system missing} within their software package and then compute the ratio of the first category for "four or more years" to the sum of the three other valid response codes.

However, doing this would yield an incorrect enrollment rate, because all of the high school graduates who did not enroll in postsecondary education would be missing from the calculation.

Considerations for Analysis of ELS:2002 Data

We know that responses coded as -9 are truly missing, and should either be imputed or changed to system missing as they don't provide a meaningful response value for inclusion in the enrollment rate calculation.

Analysts might also consider recoding the 3,613 Item Legitimate Skips (those responses coded as -3) as system missing; however, this would also yield an incorrect enrolment rate, as not everyone who needs to be counted as part of the denominator of the desired enrollment rate has been identified.

Slide 12 of 24

To properly analyze this variable, analysts must examine it in conjunction with information provided by the respondent type variable. By considering respondent type, we can find out more about the 3,613 legitimate skips. Notice that there are two groups—Non-enrollees and those still in high school in 2006—who total exactly 3,613 cases. Accordingly, both of these groups constitute the “Item Legitimate Skips” that appear in the level of the first institution enrolled variable, F2PS1LVL, because these sample members are still in high school and have not had yet attended a postsecondary institution.

By considering both of these variables together, analysts could recode these -3s to a new value of, say, 4, which would provide a new postsecondary enrollment status variable that will enable the correct number of sample students in the denominator of the enrollment rate sought.

Slide 13 of 24

Some ELS data files may need to be merged for certain analysis, Linking IDs are provided on files to facilitate merging. This slide shows what is required to aggregate course level transcript data available in ELS to student level data.

Suppose an analyst wants to estimate the percentage of 2003-04 high school graduates who complete a mathematics course higher than Algebra II.

Let's also suppose that the analyst has already created the analysis file shown in the upper right of the slide, containing the variables Stu_ID, F1RACE, BYINCOME, BYS34, and BY55 using the ELS ECB.

The analyst would then need to access the course level transcript data file, HSTRNSTU, and build a data file that shows which students have completed a course higher than Algebra II.

There are three basic steps required to create the needed analysis file that would then be merged into the existing analysis file.

First, the analyst would extract all of the mathematics courses taken by each student and create a variable to indicate if the course completed was higher than Algebra II.

Considerations for Analysis of ELS:2002 Data

There are several existing processes by which analysts can code course data from the transcript file, which will be discussed in a moment.

Next, the analyst would process this new file of math courses by STU_ID to find out which students have taken a math course higher than Algebra II. A new variable would then be created (MATH_ABOVE_ALGII) where students that have taken a higher level course than Algebra II would be assigned a value of 1, and all others would be assigned a value of 0.

Finally, the analyst would merge this new variable by STU_ID to the previously prepared analysis file containing the other variables selected to examine which students are most likely to take mathematics courses higher than Algebra II.

Slide 14 of 24

now. The ELS High School Transcript files provide course level information obtained from sample students' transcripts. Transcript data are only provided on the ELS:2002 restricted-use data file. The transcripts are coded, course-by-course, into a separate set of records for each student that are linked to the main student level file by student ID. In order to use this course level data, it is necessary to extract it from the HSTRNSCH link file and merge it to the student-level data file.

High school transcript data were collected from a majority of the ELS sample. Ninety-six percent of base year schools, and 80% of transfer schools provided transcripts. Overall, high school transcript data are available for 91% of the student sample. It is important to note that the collection strategy used obtained transcript data for both dropouts and early graduates.

In the slides that follow we will take a closer look at Classification of Secondary School Course Codes System (or CSSC) course codes and courses. More information regarding CSSC codes can be found by clicking on the underlined screen text, 'CSSC codes.'

Slide 15 of 24

The CSSC is maintained by NCES. The CSSC code of each course is the basic unit of the classification employed in coding courses from the ELS:2002 transcripts by course content. The code is six digits long. The first two digits are the "main category;" the next two digits are the "subcategory;" and the last two digits are the "course;" 270425 an example of a CSSC code in ELS, is Mathematics, Pure Mathematics, Algebra 2. In this CSSC code we see the first two digits (27) represent the main category of Mathematics. The next two digits (04) represent the subcategory of Pure Mathematics. The last two digits (05) represent the course, Algebra 2.

Here is an example of the CSSC codes for some math courses from the High School Transcript Study website. In this table we see the subcategories, course titles, and descriptions of the courses that are grouped under the main category of mathematics courses, or section 27 of the CSSC course codes.

Considerations for Analysis of ELS:2002 Data

The CSSC codes alone are generally not useful for purposes of analysis or even descriptive reporting because they are too narrow. In the next slide we will examine one hierarchy of mathematics courses that will help analysts use transcript data in ELS analyses.

For purposes of analyzing course-taking patterns, these CSSC codes must be grouped into higher-level course classification structures or taxonomies. For example, mathematics courses need to be grouped at a higher level for many analyses, such as if a researcher was interested in algebra course-taking.

Slide 16 of 24

In this SAS code, we see a hierarchy of mathematics courses listed on the right – ranging from ‘no math’ and ‘math tutoring’ to ‘AP Calculus’ and ‘Undefined math.’ By looking to the left, we see that most “math courses” that we typically think of include more than one CSSC code. For example, there are three CSSC codes associated with ‘AP Calculus.’

Running this SAS code will create a composite dividing the entire academic mathematics curriculum into 19 different course areas generally forming a hierarchy of course-taking in the mathematics curriculum.

Such composites are one form of “course taxonomy,” or mapping of CSSC codes into course areas. Other, broader taxonomies map CSSC codes into hierarchical levels of course areas across subject matter areas, or, indeed the entire school curriculum.

Slide 17 of 24

Several other such taxonomies are used or have been used in NCES reports.

One is the secondary school taxonomy (or SST), which is a broad taxonomy that is used in the career technical education reports produced by NCES and in the Condition of Education. It divides the entire school curriculum of courses into four broad areas of academic, vocational, enrichment/other, and special education courses. The SST is a mutually exclusive and strictly hierarchical taxonomy.

Another broad taxonomy is the High School Transcript Study, or HSTS taxonomy, which is used in reports generated from the high school transcripts collected by NAEP. This taxonomy has three broad areas: academic, vocational, and personal use/other. Analysts who elect to use these taxonomies should read the technical documentation closely to understand how they were created. For example, it is important to note that within the SST and HSTS taxonomies computer-related courses, special education, and some vocational courses are in particular treated differently. Additionally, while the SST is strictly hierarchical, some courses appear in two areas in the HSTS.

Slide 18 of 24

A third taxonomy was developed by Burkam and Lee. The Burkam and Lee pipeline taxonomy measures “how far” a student has progressed in each of the four traditional academic areas of mathematics, science, English, and foreign languages. For example the Math pipeline has seven hierarchical levels, which are: no math; non-academic math; low academic math; middle academic math I and middle academic math II; and advanced mathematics I, II, and III. In this mathematics pipeline, middle academic I consists of algebra III, trigonometry, solid geometry, statistics, etc.; and middle academic II consists of precalc and intro to analysis; while middle academic III consists of calculus and AP calculus.

Another subject area pipeline taxonomy is MATHLVLS, which we just examined on the previous slide and suggested as a possible taxonomy that could be used to estimate the percentage of 2003-04 high school graduates who complete a mathematics course higher than Algebra II.

Slide 19 of 24

Finally, there are also several specialized taxonomies currently being used by NCES. The curriculum concentration taxonomy differentiates academic concentrators (students with at least four credits in English, three credits in math, and three credits in science) occupational concentrators (students with at least three credits in occupational specific vocational) and those students who satisfy both requirements.

Additionally, within ELS there are composites within the data file that define both academic and occupational concentration; Science, Technology, Engineering, and Mathematics (or STEM) courses; and core versus non-core courses. Core courses are those that typically satisfy requirements for a diploma (e.g., math, English, social studies, and science), whereas non-core courses do not.

Analysts should explore the detailed documentation available with each of these taxonomies before using them within analyses. Additionally, most analysts will elect to create their own taxonomies to address their specific research questions.

Slide 20 of 24

There are a number of composite variables included in the ELS data file to further facilitate the use of the high school transcript data. Variables F1RTR09 through F2RTR12 are the number of courses taken by sample students in grades 9, 10, 11, and 12. F1RENG_C is the units the student received in English.

The composite variable, F1RSCPIP, describes the science course taking pipeline by noting the highest science course taken. The Science Pipeline indicates how high the student’s course-taking in science extended. The possible levels were: no science course taken; primary physical science; secondary physical science and basic biology; general biology; chemistry 1 or physics 1; chemistry 1 and physics 1; chemistry 2 or physics 2 or advanced biology.

Considerations for Analysis of ELS:2002 Data

Other examples of transcript composites that many data users find helpful are F1RGP, which is the overall standardized GPA from the transcript using normalized grade point average. The variable F1RHTUN provides the total Carnegie Units of courses taken, and F1RAP provides the total credits of AP/IB courses by course area.

In addition, there are some composites that combine data from the College Board on Advanced Placement and college entrance exam test scores with transcript data to provide composites with more complete information than is possible with either source alone. Examples of these are TXAP and TXSAT/TXACT.

Slide 21 of 24

Another consideration relates to appropriately addressing inconsistent data. As an example, a respondent may have a graduate degree but not a bachelor's degree on record. Of the ~1050 F3 respondents who reported earning a graduate degree, roughly half of them did not report any BA information.

In the majority of these cases, the ELS PETS data do not contradict these respondents' self-report of a graduate degree. In addition, the ELS PETS data often include information about BA degrees earned by these respondents. In other words, it appears that in the F3 survey, these respondents may have provided their highest degree, but may have not provided information about any lower, (e.g. bachelor's), degrees, for reasons varying from prior round non-response to skipping of this previous-degree loop during the F3 questionnaire.

Researchers have several options available for handling such data inconsistencies.

The first option is logical imputation. In this case, the researcher might assume that anyone with an F3ATTAINMENT value indicating receipt of a graduate degree has also earned a BA, because bachelor's degrees are required for most forms of graduate education, by definition. During the imputation process for F3 survey-based composites F3HS2BA and F3PS2BA, this is the assumption that was used.

The second option includes using the ELS PETS data instead of the ELS F3 survey data to check for BA completion. As an administrative record, the postsecondary transcript is a reliable indicator of a sample member's completion. This is particularly helpful in cases where the sample member's response is missing. However, users should note that not all PETS variables are in PowerStats, and so it is suggested that restricted-use versions of the PETS files be used if this level of detail is sought. Documentation specific to the ELS PETS files is included with those datasets, and it should be referenced for issues such as missing data.

Third, a combination of the first two options may also be possible depending on the researcher's ability to use Public-use data or if Restricted-use data must be acquired.

Slide 22 of 24

Many analyses focus on specific subpopulations, such as high school graduates, public school students, postsecondary students, or STEM majors.

When analyzing data based on complex samples, each case in the dataset contains important information needed to properly calculate variance estimates. Rather than deleting cases from your dataset that are not part of your subpopulation of interest, or using common subsetting commands such as “if” to limit cases entering your analysis, be sure to use the subpopulation selection command that is part of your software package. In Stata, this command is “subpop”, while in SAS the command is “domain”. For additional information, consult your software package’s technical documentation in conjunction with the resource document that can help you use subsetting commands in SAS and Stata to properly analyze ELS data. This resource document can be accessed by clicking on the underlined screen text ‘subpopulation command.’

Slide 23 of 24

This module has described some of the analytic considerations that should be kept in mind when using ELS:2002 data. Specifically, it described the ELS cohorts; ELS sample migration between rounds; second follow-up statuses; merging linked data files; High school transcripts; and subsampling populations.

Slide 24 of 24

Important resources that have been provided throughout the module are summarized in this slide along with the module’s objectives for your reference.

This concludes the ELS:2002 dataset training. You may now click the exit button to return to the landing page.