Slide 1 of 14

Title Slide: Getting Started with the ELS:2002 Data

Slide 2 of 14

This module introduces users to the data collected across the Education Longitudinal Study of 2002, or ELS:2002. The module describes the contents of the data files and their variables. It describes the process of obtaining ELS data and the resources that are available to learn more about the study, the data, and the data file.

Information presented in this module will be helpful in understanding some of the more detailed information presented in subsequent ELS modules. For this reason, users who are planning to proceed through the subsequent ELS modules and use ELS data for analytic purposes are strongly encouraged to complete this module first.

Throughout this module, underlined blue screen text indicates a link to additional resources.

Slide 3 of 14

There are many different types of data files within ELS:2002. All ELS:2002 public-use data, from the base year through the third follow-up, are available for download from the eDAT. Let's take a look at the eDAT, to get a better sense of how the public-use data are organized.

The largest and most commonly used ELS data files are the student-level data files and the school-level data files, both of which contain questionnaire, composite, and weight variables. Student-level data file questionnaire variables come from the student, parent, teacher, and school administrator questionnaires, while the school-level data file questionnaire variables come from the school principal questionnaire. Here we see that within the student file there are 6,571 variables available for analysis.

Let's expand the Student File option within the eDAT to discuss some of these files in more detail.

Perhaps the most important of all the student files is the ID and Universe Variables file. It contains 15 variables that provide analysts details regarding sample member status in each data collection round (also referred to as wave or sometimes follow-up), the student and school ID variables needed for merging files, and the stratum and PSU variables, which will be discussed in the module titled, 'ELS:2002 Sample Design, Weights, Variance, and Missing Data.' Universe variables will help researchers understand the number of students within the sample that are available for analysis and defining subpopulations.

Notice that four of the variables seen here are automatically tagged with a padlock. The padlock indicates that these variables are automatically included when you select data from this file because they may be critical to any analysis. Analysts can select variables within the eDAT one at a time by checking boxes in the Tag column, or by clicking Tag All from the top left of the eDAT window.

The next group of files that are critical to analysts using ELS data are those listed at the bottom of the Student File list. They are called Weights and Composites for BY, F1, F2, and F3. Let's take a look at Weights and Composites for BY, which includes 169 variables. Notice this file contains the base year weights as well as all of the NCES created composite variables that will facilitate analysis of the data.

In addition to these files, you can see there are many additional files and variables available for analysis in ELS Variables. The External Source Financial Aid Data contain information from the FAFSA and NSLDS. Variables describing the postsecondary institutions attended by ELS sample members are provided in the F2 and F3 Student-Institution Files provided within eDAT.

Slide 4 of 14

Some ELS files and variables are not available within eDAT public-use data files. For example, restricted-use ELS data contain information from high school transcript files and school catalog files that enable investigations of high school course-taking. They contain information on courses students took and the grades they earned. One composite variable, regarding student's GPA, can be found within the public-use student file in eDAT.

Additional restricted-use data include a Barron's competitiveness index file that provides information about institutional selectivity of institutions attended by ELS:2002 sample members; a GED file that contains information about GED tests taken and completed by ELS:2002 sample members; and a geocode file that provides census block and tract information linked to the residential addresses of sample members. Restricted-use variables will either be noted as (restricted) as seen in the case of the weight variable BYEXPWT, or contain a reserve code of -5 for each case within the public-use files provided in eDAT. In some cases, such as within the high school transcript data, NCES creates composite variables for inclusion in the public-use weights and composites files. Analysts should always examine the contents of the composite variables provided by NCES before requesting restricted-use files.

Slide 5 of 14

There is an ELS Variable List spreadsheet that provides analysts with a comprehensive listing of all of the ELS variables available for analysis and the data files within which they are located. This list should be reviewed to determine which ELS data files (either restricted-use or public-use) are needed to meet your research needs. This list is found on the available data page.

Slide 6 of 14

Analysts who require restricted-use data to address their research question will access that data from the restricted-use CD/DVD using the Electronic Code Book, or ECB. Here is an example of the ECB Description Window for the variable BYHOMLNG. The variable list to the left shows two selected variables from the student file – BYHOMLNG and F1HOMLNG. The variable BYHOMLNG has been selected, which populated the working taglist window on the right. The working taglist window displays the variable description for BYHOMLNG, including the variable label, notes regarding changes to the variable, the source of the variable and its location in ELS:2002 data files, a listing of the value labels, derivation and notes, and the SAS code used to construct the composite variable.

Here we see that BYHOMLNG is the sample student's native language-composite. We also see that BYHOMLNG was previously named HOMELANG on the base year (or BY) ECB. We learn that this composite variable was created for confidentiality reasons, and that it groups native languages from student questionnaire items BYS67 and BYS68 into valid values that include: English, Spanish, Other European Language, West/South Asian Language, Pacific Asian/Southeast Asian Language, and Other Language.

The SAS code provided details the exact computation (or derivation) of the composite variable.

As seen here, the ECB is used to select variables from all available ELS restricted-use data for inclusion in a user-created analysis file. In this example, the variable BYHOMLNG is the only variable selected for inclusion in the analysis file. Users with restricted-use data can use the ECB to select the variables relevant to their specific analysis (such as ID and universe variables, weight and composite variables, and other variables of interest) to generate the SAS, SPSS, and Stata syntax files that can be used to produce corresponding system files for analysis.

Slide 7 of 14

There are many different composite variables in ELS:2002 that can be used for analysis. ELS:2002 composites differ in the way that they are constructed.

Some ELS:2002 composite variables combine data from two or more data sources, such as questionnaire and transcript data. An example is F2HSSTAT, which is the high school completion status of sample members as of the second follow-up. Some ELS:2002 composite variables combine data from multiple variables according to some construct, as the socio-economic status, or SES variables do. An example of this type of composite is F1SES2, which is the SES computed for each sample member at the first follow-up. Some composite variables contain logically or statistically imputed values, such as F1SEX and the math achievement variable, F1TXMI1R. Some merge data from linked files to the student level (as many second follow-up, or F2, composites do) to

save researchers from having to merge files themselves. Some composite variables recode questionnaire responses, like the college major variable, F2MAJOR4 does, making it easier for the researcher to carry out analysis of common interests. Lastly, some ELS:2002 composite variables include data from the survey control system that can be used to understand sample members' statuses in different rounds of data collection. F1UNIV1 and F2UNIV1 are examples of such composite variables.

No matter what type of composite variable you may use, the source data for all composite variables is included in the ELS:2002 data files. The code that was used to produce the composite variables, if available, is included in the ECB, or eDAT Description Window. Additionally, details regarding composite variables can be found in the ELS:2002 Data File Documentation. Remember that all composite variables can be located in the Weights and Composites files within the Student File in eDAT.

Slide 8 of 14

The choice of which ELS:2002 data file to use should be based on your research question and what type of variables you would need to address your question of interest. For example, investigations involving mathematics course-taking in high school would require the high school transcripts, which are only available in the restricted-use files. Analysts interested in data from postsecondary transcripts will also need to use the restricted-use files when they are issued in Spring 2015.

Slide 9 of 14

As discussed in the common module titled, 'Acquiring Micro-level NCES Datasets,' restricted-use data are only available to researchers who apply for and are granted a restricted-use license.

In general, the restricted-use file contains more data and a wider range of data values than are included in the public-use files. For most users, the public-use files provide all the data they will need for most analyses, though some users may find that only the restricted files have the specific data they need. It is recommended that researchers who are uncertain of which data file to use first examine the public-use data file and the ELS Variable List spreadsheet to determine whether their specific analytical objectives can be met using public-use data. Also, each data file user's manual includes a table listing all the variables that have been altered in some way for the public-use file. Lastly, it is important to note that information regarding the ELS expanded sample, base year ineligible sample members, and high school and postsecondary transcript data is only available within the restricted-use files.

Slide 10 of 14

Across all ELS data files, both public- and restricted-use, weights must be used to calculate appropriate estimates and standard errors to ensure accurate analysis of ELS data. Weight variables are found within the files nested under the Student File within

the eDAT. The weight variables associated with each round of data collection are provided within the corresponding Weights and Composites file. For example, the weights associated with base year data are found within the file 'Weights and Composites for BY.' In addition to these round-specific weights, analysts will also include either replicate weights or strata and PSU variables for the proper calculation of standard errors. Replicate weights are found within files that begin with 'Weight Replicates' and strata and PSU variables are found within the ID and Universe Variables file.

More information about these weight variables is provided in the module titled, 'ELS:2002 Sample Design, Weights, Variance, and Missing Data.'

Slide 11 of 14

In ELS:2002, there were several reserve codes used across all rounds of data collection. A reserve code of {-1} indicates the person who filled out the questionnaire selected or wrote in "don't know." A {-2} indicates that the person refused to answer the question. A legitimate skip, or {-3}, meant that, due to the person's response on a previous question, they were not given this question. In other words, they were routed away from this question due to skip logic. For example, sample members who answered that English was their first language were not asked to specify what other language was their first language. A {-4} indicates that the sample member did not respond, or, in the case of linked data, that the person supplying that component did not respond. Reserve code values of {-5} indicate the response a person provided was out-of-range. Sample members who responded that they were not actually in the grade that the study sampled would be noted with a {-5} on the data file. In addition, data suppressed on the public-use file will include a {-5} code. A reserve code of {-6} indicates the respondent circled or filled in multiple answers to a question. A {-7} indicates the respondent stopped filling out the survey (or broke off) at or before he or she got to a question. The {-8} reserve code is used when the respondent legitimately did not respond to a section of the questionnaire based on their answers to a gate or routing question earlier in the survey. For example, sample members who indicated they had graduated from high school would not be asked questions from the "reasons for dropping out" section of the survey. Finally, a reserve code of {-9} indicates a missing answer, usually indicating an item was skipped by the respondent.

Slide 12 of 14

No matter which student-level ELS data files you examine or build, the same number of cases, 16,197, will appear in each file. However, keep in mind that data are not available for every student on the data file for every ELS variable. Individual items may be missing or entire student cases may be missing. The ID and Universe variables are essential to understanding the status of sample members within and across each round of ELS data collection. Running descriptive statistics on the ID and Universe variables will provide you with an indication of how many cases should appear in each of your unweighted analysis runs.

Once the appropriate weights are applied, cases that should not be included in your analysis will be omitted – as the weight variable will be set to zero. Additionally, it is important to consider the reserve codes as decisions are made regarding recoding missing data and overall data handling.

Slide 13 of 14

ELS:2002 data are released in two formats, public-use files via the eDAT and restricted-use files accessible via the ECB. No matter which tool you use to access the data files you will generate syntax files to create a file for your analytic purposes.

Analysts should note that the eDAT will output files in seven statistical package formats: R, SAS, SPSS, Stata, S-Plus, ASCII, and CSV. The ECB will only output files in three statistical package formats: SAS, SPSS, and Stata. Output files from both eDAT and ECB can be used to generate the syntax files used to produce corresponding system files for analysis. Remember, only public-use data are available within eDAT. Analysts using restricted-use data with a statistics package other than SAS, SPSS, and Stata will need to translate the output files into the statistical programming language of their choice.

Slide 14 of 14

This module introduced users to the data collected across the Education Longitudinal Study of 2002 and described the contents of the data files and their variables. It also described the process of obtaining public-use ELS data from the eDAT website and restricted-use ELS data using the ECB provided on the restricted-use data CD/DVD.

This module also described the resources that are available to learn more about the study, the data, and the data file. These included the eDAT website and the ECB, the variable list and available data page and other data file documentation provided on the ELS:2002 website.

Remember, no matter which student-level ELS data files you examine or build, the same number of cases will always appear in each file: 16,197. Once the appropriate weights are applied, cases that should not be included in your analysis will be omitted.

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

You may now proceed to the next module in the series, or click the exit button to return to the landing page.