Slide 1 of 18

Title Slide: Getting Started with the HSLS:09 Data

Slide 2 of 18

This module introduces users to the data collected across the High School Longitudinal Study of 2009, or HSLS:09. It describes the process of obtaining public-use HSLS:09 data from the HSLS:09 website and the resources that are available within the Restricted-use DVD.

The module describes the contents of the Base Year and First Follow-up data files, as well as their variables, and variable naming conventions. This module also provides resources that are available to learn more about the study, the data, and the data files. Information presented in this module will be helpful in understanding some of the more detailed information presented in subsequent HSLS:09 modules. For this reason, users who are planning to proceed through the following HSLS:09 modules and use HSLS:09 data for analytic purposes are strongly encouraged to complete this module first.

Slide 3 of 18

Across the span of the study, HSLS:09 collects data at multiple points in time, from multiple sources, using multiple methods.

HSLS:09 began in 2009 with a nationally representative sample of 9th graders. In 2012, HSLS:09 followed up with most of the sampled students in their 11th grade year. Subsequent data collections are planned to occur in 2013, 2016, and 2025 to follow students across their young adult years.

In the 9th grade year, also referred to as the base year, data were collected from the students, their parents, mathematics teachers, science teachers, school counselors, and school administrators.

In 2012, data were collected from the students, their parents, school counselors, and school administrators but were **not** collected from the students' teachers.

The 2013 follow-up collected information about the student from the students themselves, or their parents; as will the planned 2016 and 2025 follow-up data collections.

Slide 4 of 18

HSLS:09 data are available in both public- and restricted-use formats.

Public-use data can be accessed by anyone; they can be downloaded as full data files from the Data and Documentation page on the HSLS:09 website. They are available in three software-based formats: SPSS, SAS, and Stata.

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 to use them. The restricted-use file for a given round of data collection contains more data and a wider range of data values than are included for that round in the public-use file.

For most users, the public-use files provide all the data necessary for most analyses, though some users may find that only the restricted-use files contain 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 to ascertain whether their specific analytical objectives can be met using that data file. All restricted-use variables within the public-use file have been suppressed and contain -5 codes throughout.

Slide 5 of 18

There are two ways to download the pubic-use HSLS:09 data. You can go to the Education Data Analysis Tool, or eDAT, website at http://nces.ed.gov/edat, or you can access it by clicking on the underlined screen text 'eDAT'. Follow the instructions on the screen to start the eDAT. Click on the plus sign next to HSLS (High School Longitudinal Study of 2009), select HSLS:09, and then click on next at the bottom.

This process will guide you through downloading the syntax file in the statistical language of your choice. You can also access the public-use data from the 'Data and Documentation' page on the HSLS:09 website.

Regardless of how you access the data, once you have downloaded the public-use data file, be sure to download the data file documentation, which can be accessed by clicking on the underlined screen text 'Data File Documentation'.

Slide 6 of 18

If you have access to restricted-use HSLS:09 data, you will find important resources available within the restricted-use DVD. You should always start with the most important resource on the DVD which is the README file. This file contains information about the DVD contents and structure. It also provides you with installation information.

Slide 7 of 18

Once the DVD is installed and you have reviewed the README file, you can use the Electronic Codebook, or ECB. The ECB helps users create customized data files with only the variables needed for analysis, which may be easier to work with and analyze than the full data files. Users can view descriptions of all the variables in the data file, along with the variable categories and frequency distributions, in the electronic codebook. Functions within the ECB allow users to identify which variables they want to include in their customized data files and then create a syntax file to be used within a statistical software package to generate their analysis data file. Syntax files can be generated for SAS, SPSS, or Stata.

Note that the ECB does **not** produce a data file. More information on the ECB, including a guide to using it, can be found in the user's manual.

Methodological and/or analytical reports, reference materials, and source documents are also available as resources on the DVD.

Slide 8 of 18

Whether you are using a public-use or restricted-use dataset, you will need to understand the contents of the base year data file.

Because the base year study design supports estimates of both students in 9th grade **and** schools that offer 9th grade, there are two data files associated with the base year data file release.

(1) A student level file with all data collected as part of the study merged to the student level. For this data file, student is the unit of analysis and when appropriately weighted, is representative of the approximately 4 million students enrolled in 9th grade in the fall of 2009.

(2) A school level file with data collected from schools and school counselors merged to the school level. For this data file, school is the unit of analysis and when appropriately weighted, is representative of the approximately 23,000 schools offering 9th grade in the fall of 2009.

Slide 9 of 18

The First Follow-up data file contents are structurally very similar to the base year file contents. The two main differences are that 1) there was no school file for the first follow-up, and, 2) there were no teacher questionnaires collected in the first follow-up. All base year data have been appended to the first follow-up data. So the school file and the teacher questionnaire data, along with all other base year data, appears on the first follow-up DVD.

Slide 10 of 18

The HSLS:09 2013 Update Data File includes derived variables, student-level analytic weights, the student and parent questionnaires, high school transcript files, and high school transcript course files.

Slide 11 of 18

The variables in the HSLS:09 data file are named using a standard set of conventions. Data users should find naming conventions for variables, flags, and weights intuitive.

The first character of the variable name is the component identifier.

The second character of the variable name indicates the round of data collection. All base-year variables are “1” and subsequent rounds will follow sequentially. For example, the first follow-up is “2.”

Characters 3 through 12 indicate a descriptive name for the variable. Variable names have been expanded beyond the eight characters used by previous data products because SAS, SPSS, and Stata now support longer variable names.

Slide 12 of 18

Variable labels offer more description than the variable names. For convenience the first two characters of the variable name have been retained in the variable label to indicate the component and the round.

For example, let’s consider the variable ‘P1JOBNOW1’, which is spelled P-1-J-O-B-N-O-W-1. The ‘P’ in the first character position indicates that this variable is from a parent questionnaire, and the ‘1’ in the second character position, indicates that it is from the base-year. The variable label associated with ‘P1JOBNOW1’ is ‘P1 C05 Parent 1 currently holds a job.’ The first two characters of the variable name, ‘P1’, are retained in the variable label. The next part of the variable label will describe the section and the question number within that section, if applicable, from which the variable originates. In this example, the variable ‘P1JOBNOW1’ comes from the fifth question in section C of the parent questionnaire, or ‘C05’. The last part of the variable label is a text description of the item. In this case the fifth question in section C of the parent questionnaire administered within the base-year asks if, ‘Parent 1 currently holds a job.’

Slide 13 of 18

This slide provides examples of the variable naming conventions.

The first letter of the variable name indicates the survey component from which it comes. In the case of X1SEX and X1RACE, the X in the first character of the variable name indicates a composite variable. In the case of S1SUREHSGRAD, the S indicates that the variable is from the student questionnaire. The other variables shown here come from other components including the parent questionnaire, as indicated by the first character P; the mathematics teacher questionnaire, indicated by the letter M in the first character position; and the science teacher, administrator, and counselor questionnaires, as indicated by the letters N, A, and C. Lastly, the variable W1STUDENT begins with a W indicating that it is a weight variable.

Each of the first character component indicators in the variable names are followed by a numeric round identifier. All of the variables featured here are from the first round of data collection – or the base year. Finally, the third through twelfth characters of the variable name provide a shortened description of the contents of the variable. For example in the case of C1FTCNSL, the letters FTCNSL represent a shortened version of full-time counselors as the item on the base year counselor questionnaire inquires about the number of full-time high school counselors at the sampled student’s school.

Slide 14 of 18

Based on the assessment and questionnaire data, HSLS:09 data contain computed or derived variables. These are often referred to as composite variables. As mentioned, these variable names begin with an “X.”

Composite or derived variables are variables that were not directly collected as part of the instruments. They were created for users using information from two or more variables, two or more sources, or both.

Sometimes the composite variables were derived using information that is not available to the researcher in the data file. For example, the items from the direct math assessment are not available on the data file.

Composite variables also save time and effort. Detailed descriptions of all the composite variables can be found in the data file documentation.

For example, the X1RACE characterizes the sample member’s race/ethnicity by summarizing the following six dichotomous race/ethnicity composites: X1HISPANIC, X1WHITE, X1BLACK, X1ASIAN, X1PACISLE, and X1AMINDIAN. The dichotomous race/ethnicity composites are based on data from the student questionnaire, if available; if not available from the student questionnaire, they are based on, in order of preference, data from the school-provided sampling roster or data from the parent questionnaire.

In addition to this detail, the SAS code used to produce these variables can be found in the variable description windows of the ECB and EDAT for your reference.

Slide 15 of 18

Examples of composite variables derived from the algebra assessment are provided in the table shown here. As shown previously, the X in first character position of these variable names indicates that these variables are composite variables, and the number 1 in the second character position indicates that they are derived from the first round, or base year, data collection.

There are many variables across the study data collections and data sources that may appear similar, but are actually very different. It is imperative to always consult the README file and the codebook associated with your derived variable of interest to ensure proper analysis of the data.

Slide 16 of 18

Examples of composite variables derived from the questionnaires and related to family characteristics are provided in the table shown. Once again, the X in first character position of these variable names indicates that these variables are composite variables, and the number 1 in the second character position indicates that they are derived from the first round, or base year, data collection. The characters in positions three through twelve provide a shortened description of the contents of the variable.

Slide 17 of 18

These modules have been designed to help you understand and analyze high school longitudinal study data. While completing these modules will provide a solid foundation of knowledge, the modules alone do not provide sufficient detail to ensure successful analyses that reflect a comprehensive understanding of the data and what they mean.

First and foremost anyone interested in using the data should read through the data file documentation, the questionnaires, and the First Look reports.

In addition, it is always a good idea to check the study website occasionally for information about new data releases or other important information that users should know.

Slide 18 of 18

This module has provided users with an overview of the data collected across the HSLS:09. This module also described how to download the HSLS:09 public-use data and has described the resources available within the restricted-use DVD. It has also described the contents of the Base Year and First Follow-up Data Files and has described variable information and naming conventions. Finally, the module has provided resources that are available to learn more about the study, the data, and the data files.

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