

Slide 1 of 14

Title Slide: Getting Started with the Postsecondary Education Data

Slide 2 of 14

This module describes the resources that are available to users of NPSAS, BPS, and B&B restricted-use micro-level data files, including the contents, file formats, file structure, and variable information and conventions. Additionally, this module describes the resources available to learn more about the study, data, and data files. Information presented in this module will be helpful in understanding some of the more detailed information presented in subsequent modules related to postsecondary longitudinal studies and sample surveys.

Slide 3 of 14

Restricted use data files are delivered to approved data users on an encrypted CD. In order to access the contents of the CD, you will need a passphrase from the IES Data Security Office.

The README file on the restricted-use CD is the most important resource available to new data users. Many researchers' first instincts are to begin exploring the data; however, we **STRONGLY** recommend that you begin by getting to know the structure of the restricted-use data files by reading the accompanying README file.

The README file details the name, location, and brief description of each file; the number of student, institution, or other type of record each file contains; the key source or sources of data represented by a given file; and the year or years of the interview, administrative record, or other type of data the file represents.

The README file has everything you need to know about getting started with your restricted-use data CD. For some CDs, this also includes installation instructions for specific software needed to use the data files.

Finally, all CDs include a codebook or codebooks, questionnaire facsimiles, and other important background documents that can help you fully understand the micro-level data file for proper analysis. These may include PDFs of the study's methodology report, analytical reports or table compendia, and syntax used to create derived variables.

Slide 4 of 14

The NPSAS, BPS, and B&B data files come in a variety of formats, depending on when the restricted-use data was released.

Most recent releases include data as CSV files. NCES provides SAS, STATA, and SPSS syntax files that read-in those CSV files, and then use an accompanying metadata file to name, format and label each variable.

Getting Started with the Postsecondary Education Data

In a few cases, the data may be provided as formatted SAS, STATA, and SPSS files.

Prior to 2013, many NCES restricted-use CDs included the proprietary ECBW, or Electronic Codebook for Windows software. That software was used to output SAS and SPSS syntax for extracting data from fixed-width files stored on the CD. Installation instructions for the ECBW software are included with your CD if applicable.

Slide 5 of 14

Each NPSAS, BPS, and B&B restricted-use CD contains multiple files broadly divided into three categories: derived files, source files, and weight files. The derived file contains all of the variables created – or derived – by NCES, as well as sampling weight, PSU, and strata variables. The derived file contains all the variables needed by the overwhelming majority of researchers and therefore should be used in almost every scenario.

In addition to the derived files, the raw source files used to create derived variables, as well as those used in sampling and weighting, are contained on the CD. All files on the CD contain ID variables to permit matching.

Users should always use caution in combining files, because multiple seemingly similar ID variables may exist. Additionally, some files contain multiple records per student or contain records at the institutional-level, matching may produce unexpected results. Users should carefully check the results of their merges, including those using syntax created by the ECBW software, to ensure correct matching.

Slide 6 of 14

Each study's derived file is "ready-to-use." That is, the data are fully edited, value labels are clearly defined, missing or implausible values are imputed when applicable, and all key data elements, including variables needed to analyze complex survey samples, are included. NCES's publicly-facing web tool, PowerStats, uses the same derived file as contained on the restricted-use CD to generate estimates, making it possible for analysts to check their results against an external source. To learn more about how these derived files, and the variables they contain, are built, click on the corresponding underlined screen text.

Slide 7 of 14

As noted earlier, the derived file is built from a set of source files, including student interviews, institutional information from the Integrated Postsecondary Education Data System (or IPEDS), student information provided by the institution, and Department of Education administrative records. When possible, NCES uses multiple sources of data on the same or similar data element to determine the most plausible value. As a result, values one might expect to be identical between (and even within) files and derived files may not be consistent.

Getting Started with the Postsecondary Education Data

The process by which NCES creates the derived file from source files is made available to users through the inclusion of source data and syntax files on the restricted-use CD. By reviewing that syntax and the accompanying documentation, users can see how source files were combined to produce a final derived value. For a detailed description of the source data files and the location of syntax files, see the README file on the restricted-use CD.

Slide 8 of 14

As you have already learned, analyzing NCES datasets requires the use of statistical software that can account for each study's unique sampling design. The derived file will typically include each student's final analysis weight, along with variables representing primary sampling units and strata.

Replicate weights, used to estimate variance by some statistical packages, are typically included in their own weight file. That weight file may also contain weight components, that is variables representing the stages of weighting and adjustment that are multiplied together to form the final analysis weight. Analysts conducting multi-level models can use these weight components to separate variability attributable to institutions from the students nested within them.

Slide 9 of 14

The README file on each restricted-use CD includes brief, one-sentence descriptions of each file that can be found there.

Derived files can be identified by some part of the word "derived" in its title. For example, in the BPS:04/09 study, the final derived file that contains all variables created through 2009 is referred to as "F9DER." Because the BPS:04/09 study was the third and final wave of the BPS study, it also includes derived files built at the end of the base year (2004) and at the end of the second follow-up (2006). These files include N4DER and B6DER in their names, reflecting they were based upon the NPSAS:2004 and BPS:2004/2006 study.

Slide 10 of 14

In general, because derived files are cumulative, there is no reason to use prior-wave derived files. The final derived file, F9DER in this example, should be used for all analyses. Some CDs may contain multiple derived files, representing students, institutions, and, if applicable, transcript records.

Slide 11 of 14

Each CD contains one or more codebooks that lists each variable's name, description, label, and other important details. The codebook also provides an unweighted frequency distribution for the variable that includes missing value codes if applicable. When missing values or other special values, known as reserve codes, are used in data

Getting Started with the Postsecondary Education Data

files, they are represented by negative numbers. Examples include -3 for legitimate skips and -9 for user missing values.

Codebooks are lengthy and navigation can be difficult. Recent codebooks are in PDF form, bookmarked by the first letter of the variable name or variable subject, to ease navigation. Older codebooks were typically simple text files.

Analysts using the ECBW software can use the software itself to search through the codebook. However, one note of caution: the same variable name may appear across multiple files. When using the ECBW to select your variables, be sure you are selecting all of them from the final derived file.

Slide 12 of 14

These modules have been designed to help you understand and analyze postsecondary education data. While completing these modules will provide a solid foundation of knowledge, the modules alone do not provide sufficient detail to ensure successful analyses.

The README file and codebook will facilitate a good start at understanding the postsecondary education data. However, researchers interested in using the restricted-use micro-level data should read through the relevant user's manuals and technical reports, which provide a wealth of information about the study and on study members.

Data from these postsecondary studies are also available through the public-facing DataLab website, which includes a table library of existing publications and access to QuickStats and PowerStats, which allows users to create custom tabulations. QuickStats allows users to generate simple tables with one row variable and one column variable. PowerStats allows users to generate more complex tables with multiple row and column variables. These tools enable researchers to explore the types of postsecondary data that are available for analysis before requesting restricted-use micro-level data for analysis. It is possible that custom tables or analyses in QuickStats or PowerStats can address your research question.

Lastly, it is always a good idea to check the study websites occasionally for information about new data releases or other important information that users should know. For additional information about available resources, click on the corresponding underlined screen text.

Slide 13 of 14

This module describes the resources that are available to users of NPSAS, BPS, and B&B restricted-use micro-level data files. Information presented in this module will be helpful in understanding some of the more detailed information presented in subsequent modules related to postsecondary longitudinal studies and sample surveys.

Slide 14 of 14

In addition it has provided resources that can be accessed through the DLDT system and/or on the NCES website.

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