Slide 1 of 18

Title Slide: Getting Started with the NHES Data

Slide 2 of 18

This module introduces users to the data collected through the NHES and the resulting micro-level data files. It provides information needed to get started with the NHES data. It describes the data that are available, how to obtain the data, and some considerations to ensure quality in your NHES analyses. It also describes the types of variables available on the data files, the data file ordering, and variable naming conventions. In addition, this module also describes the resources that are available to learn more about the data files.

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

Slide 3 of 18

The most recent NHES data are from the 2011-2012 school year. The 2012 NHES included two Topical surveys—one on early childhood programs and one on family involvement in schools. Data were collected in the winter and spring of 2012. The unit of analysis in both surveys is the sampled child. This means that when you make population estimates or describe the data, your estimates for a given characteristic will describe the number or percentage of children with that characteristic. For example, you can estimate the percentage of children with a particular household income level or the percentage of children whose families participate in school-related activities. The estimates do not describe households or families.

Information about the sampled child was provided by a parent or another adult in the household who was knowledgeable about the child. The surveys were conducted by mail using a paper-based questionnaire.

Slide 4 of 18

Past NHES surveys from 1991 to 2007 were conducted by telephone with interviews that ranged from 15 to 30 minutes in length. In this table, you can see the various survey topics conducted under the NHES program by years administered. Like the 2012 survey, the child is the unit of analysis in past NHES surveys about young children and school-age children. The NHES also fielded topical surveys of adults at the same time as the child-focused surveys. In the adult surveys, the sampled adult is the unit of analysis and population estimates are of non-institutionalized adults in the United States.

Slide 5 of 18

NCES public-use data, including the NHES, are available for free online and can be downloaded directly from the study's website – which is accessible by clicking on the corresponding underlined screen text. Most data collected in the NHES are public-use data. Public-use NHES data can also be downloaded from the Education Data Analysis Tool, or eDAT. The eDAT can be accessed by clicking on the corresponding underlined screen text.

Restricted-use NHES data are available to researchers who obtain an NCES restricted-use data license. Generally, restricted-use data files include data that may potentially identify an individual in some way, such as verbatim responses and detailed geographic information like zip code, and a child's state of residence.

The data file user's manuals explain which variables are restricted. The published questionnaires for administrations prior to 2012 may also be a useful resource, as they also identify restricted-use variables. Variables that are only available in restricted files are marked with an "R."

More information about obtaining access to restricted-use data is available in the Common Module titled "Acquiring Micro-level NCES Data," which is accessible by clicking on the corresponding underlined screen text.

Slide 6 of 18

NHES data files from 1999 to 2012 are provided online at the study's website in ASCII format along with the syntax code necessary to import the data into SPSS, SAS, and Stata. Each data file is accompanied by a data file user's manual and a complete codebook which shows each variable name, variable description, question and question number, position order, response values, and weighted and unweighted frequencies.

Slide 7 of 18

This slide shows the years for which selected NHES repeating Topical survey data files are available online from the study's website. Each of the Topical survey data files are accessible by clicking on the underlined screen text, 'Surveys.' The data files from the Adult Education Survey conducted in 2001, 2003, and 2005; the After-School Programs and Activities Survey conducted in 1999, 2001, and 2005; the Early Childhood Program Participation Survey conducted in 1999, 2001, 2005, and 2012; the Parent and Family Involvement in Education Survey conducted in 1999, 2003, 2007, and 2012; and the School Readiness Survey conducted in 1999 and 2007.

Slide 8 of 18

NHES data files from 1991 to 2012 are available through eDAT, which is accessible by clicking on the corresponding underlined screen text. It is important to note that NHES data files from 1991 to 1996 are **only** available through the eDAT as the NHES: 1991 to 1999 CD-ROM is no longer available through EDPubs.

Slide 9 of 18

This slide explains the types of variables you will find on the NHES data files. Most of the variables are numeric, but some are character variables, which are also known as string variables. The data file user's manual identifies which variables are character variables. These variables are also identified in the syntax code.

NHES variables can also be categorized by their source and function. These types of variables are: case identifiers, questionnaire item variables, derived variables, household characteristics variables, appended variables, weighting variables, and imputation flag variables.

The case identifier variable is the numeric case ID. On most NHES files this variable is called B-A-S-M-I-D, or base-M-I-D. Questionnaire item variables are linked directly to a questionnaire item and its response categories. They can be found next to each question on the questionnaires available in the data file user's manual. Derived variables are variables created by NCES by recoding item variables to create commonly used analysis variables. Household characteristics variables are often derived variables and are shown in the data file user's manuals. Some household characteristics variables are also direct questionnaire variables, like household income.

Appended variables have been merged into the NHES data files from other NCES and Census Bureau data sources. They contain information about the child's school and characteristics about the household's geographic location, such as percent of families in poverty in that zip code area.

Weighting variables include the person-level weight plus replicate weights that are used to weight the estimates. These weights are also used to calculate appropriate standard errors that account for the complex survey design. Lastly, imputation flag variables indicate which variables have been imputed and how many cases for a particular variable were imputed.

Slide 10 of 18

All variables in the NHES have names that roughly correspond to the variable description. For example, RACEETH2, spelled R-A-C-E-E-T-H-2, is a derived variable that indicates a sampled child's race and ethnicity and P1AGE is a variable that indicates the age of the first parent reported about in the questionnaire.

Variables with the same name in multiple years are usually directly comparable. However, variables with similar but different names are not directly comparable.

As mentioned earlier, variable names and descriptions for the questionnaire items can be found in the annotated questionnaires in the data file user's manuals, as well as in the codebooks. Variable labels include the question number that corresponds to the variable. If a variable is derived, the question number is replaced with the letter "D". Sometimes a variable appears on more than one survey. When this happens, a letter corresponding to each survey the question is on is included in the description. For

example, in the 2012 data file for the Parent and Family Involvement surveys, the letter "E" indicates the question number on the survey for enrolled students, and the letter "H" indicates the question number on the survey for homeschooled students.

The naming conventions for imputation flag variables, weight variables, replicate weights, and stratum and cluster identification variables are described in the module titled 'NHES Sample Design, Weights, Variance, and Missing Data' which can be accessed by clicking on the corresponding underlined screen text.

Slide 11 of 18

As mentioned previously, B-A-S-M-I-D, or base-M-I-D, is the NHES case identifier. In some NHES data files there may also be household or person identifiers to allow linking of persons in the same household who appear on different data files. However, most analysts will use a single identifier.

Another useful variable on the child surveys is the variable that denotes the main population groups of interest. It identifies, for example, if the child is an elementary school, middle school, high school, or homeschooled student. This variable is called PATH in recent NHES surveys or MAIN-R-S-L-T in earlier surveys.

Slide 12 of 18

In addition to data collected through the NHES surveys, some data from the U.S. Census Bureau and NCES school databases are appended to the files. Data from the Census Bureau provide additional information about the zip code area where the NHES household is located, such as percent of families in poverty. Data from NCES school databases provide information about children's schools or school districts, such as percent of students receiving full or reduced price lunches.

Variables appended from these external sources are described in detail in the data file user's manuals. Many of these variables are only available to NHES restricted-use data license holders, but several are included on the publicly available data files.

Slide 13 of 18

It is important to note that all missing data in the NHES are imputed. This makes it easier to conduct analyses on the full NHES data file for all variables. Each variable that contains imputed data will be accompanied by an imputation flag variable. The data provided in the imputation variables allow analysts to recode missing data using a different imputation methodology if they wish. However, it is recommended that imputed data be used. Note also that missing responses are different from "not applicable" responses. Missing responses are denoted in the data file with a -1. A detailed discussion of missing data and imputation is presented in the module titled, 'NHES Sample Design, Weights, Variance, and Missing Data,' which can be accessed by clicking on the underlined screen text, 'missing data.'

Slide 14 of 18

When you first get the NHES data, there are some basic checks you should do before you run any statistical programs.

First, if you are accessing restricted-use NHES data via CD ROM, make sure the data CD is inserted. Some NHES CD packages come with two CDs, one containing the data and one containing codebooks and other information. If you have downloaded the data from the study website, make sure the name and location of the data file are correct in the read-in program.

When you begin analysis, start with and keep an original, write-protected file. This way, if you delete cases or variables and then decide you need them later, you won't have to start over.

It is also useful to organize your code. Even if you are using the point-and-click feature of a software program, paste the code into a file that you can rerun later if you want to make a change. It is helpful to separate read-in code from recodes and recodes from analysis code. It is also helpful to make notations in your program about what your syntax does or what cases it applies to.

Slide 15 of 18

The data file user's manuals should be your first stop for information. The user's manuals explain the study methodology, sample sizes, response rates, data file organization and key data considerations. You should also check the survey website regularly.

When using the NHES data, it is important to understand the unit of analysis. As described earlier in this module, in NHES surveys about children, the child is the unit of analysis. In surveys about adults, the adult is the unit of analysis.

Remember that the NHES underwent a mode change between the 2007 and 2012 surveys. Prior to 2012, the NHES was conducted by telephone interviews. In 2012 the NHES mode changed to a self-administered mail survey. This mode change is important to consider when making comparisons of NHES data collected using different modes. Although many estimates will not be affected by differences in mode, some may be. Review the questionnaire wording carefully when making comparisons across time and always account for the possibility that change over time could be due to the change in mode.

Even when using data from a single year, it is always necessary to review the questionnaires and question wording. The questionnaires will tell you exactly how a question was asked. It will also tell you who was asked the question. Some respondents are directed to intentionally skip questions that do not apply to them or their child.

Lastly, every NHES data file contains derived variables. Derived variables are recodes of common analysis variables that NCES has added to the data file because they are

often of analytic interest. Some examples of these include parents' highest level of education, child and parent race and ethnicity, and household composition. To help save time, check the user's manuals to locate and examine these derived variables before you begin your analyses.

Slide 16 of 18

Once you begin your analyses, there are several steps you can take to check the accuracy of your programs. These steps can help you find inconsistencies and errors. Run descriptive statistics and look for outliers and unlikely values that you might want to consider recoding or excluding from your analyses.

All NHES data have undergone internal NCES review and any data entry errors have been corrected. Unlikely responses that are determined to be in error have also been corrected. However, if there is no evidence that an unlikely response is wrong, it is kept as a valid response. As an analyst, you might decide the value is unlikely, erroneous or otherwise unusable, and choose to exclude it from your analysis. The data file user's manual documents many data anomalies, so always check there if you come across something unusual as there may be an explanation for it in the documentation. For example, two children are reported to be living with a birth parent and a foster parent. As we are not sure if these are correct responses or whether respondents gave incorrect information, researchers interested in analyses of foster parents may want to consider how to include these cases in their analyses.

You should also run your analysis both weighted and unweighted. Using unweighted data helps you more easily identify odd cases. It will also help you see where you might have high variance due to small cell sizes for some variable combinations. However, always run your final analyses using weights to make sure the complex sample design is accounted for and that estimates produced generalize to the target population.

Slide 17 of 18

When using the NHES data, it is important to use the appropriate weights in your analysis to account for the complex survey design when producing estimates and standard errors. All NHES data files contain replicate weights that can be used to calculate appropriate standard errors. When recoding variables, you should perform cross-tabulations and look at unweighted frequencies to check for any coding errors. It is also important to check results of your analyses for plausibility. When possible, try to verify your estimates against external sources. If your results are surprising, it is a good idea to check for potential coding or programming errors.

Detailed information regarding weighting NHES data is provided in the module titled, 'NHES Sample Design, Weights, Variance, and Missing Data,' which can be accessed by clicking on the underlined screen text, 'weights.'

Slide 18 of 18

This module introduced users to the data collected through the NHES and the resulting micro-level data files. It also described the resources that are available to learn more about the data files.

Additionally, important resources that have been provided throughout the module summarized here 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.