

Slide 1 of 16

Title Slide: Weights, Variance, and Missing Data

Slide 2 of 16

HSLs:09 used a complex design for selecting the sample from which data were collected. As discussed in the common modules, when analyzing data from sample surveys, certain procedures must be used to assure that estimates made from the data are representative of the population and that hypothesis tests are accurate. Specifically, weights must be applied and standard errors must be calculated using appropriate procedures. This module discusses these issues specifically in relation to analyses of data from HSLs:09.

For information about weighting and standard error calculation more generally, please view Common Module 3: Analyzing NCES Complex Survey Data and Common Module 4: Statistical Analysis of NCES Datasets Employing a Complex Sample Design, if you have not already done so.

Slide 3 of 16

HSLs:09 is a sample that when appropriately weighted is nationally representative of students enrolled in 9th grade in the fall of the 2009-2010 school year. Additionally, in the base year, HSLs:09 when appropriately weighted is nationally representative of schools with 9th and 11th grades.

HSLs:09 is not a simple random sample of 9th graders in the United States in 2009.

Slide 4 of 16

In the base year survey of HSLs:09, students were sampled through a two-stage process. First, stratified random sampling and school recruitment resulted in the identification of 1,889 eligible schools. A total of 944 of these schools participated in the study, resulting in a 55.5 percent weighted response rate.

Slide 5 of 16

In the second stage of sampling, students were randomly sampled from ninth-grade school enrollment lists, with 25,206 eligible students selected (or about 27 students per school). A total of 21,444 students participated, resulting in an 86 percent weighted response rate.

All students who met the target population definition were deemed eligible for the study. However, not all students were capable of completing a student questionnaire or algebra assessment. Students who, for reasons of language barriers or severe disabilities, were unable to participate directly in the study were retained in the sample and contextual data were obtained for them. Their ability to complete the study

Weights, Variance, and Missing Data

instruments was reassessed in the first follow-up. Of the 25,206 eligible students, 24,658 were classified as questionnaire-capable and 548 as questionnaire-incapable.

HSLs:09 school and student samples are nationally representative and also state-representative for a subset of 10 states. More information about the subset of states for which state-representative data are available is presented in the module titled, 'Considerations for Analysts of HSLs:09 Data' and can be accessed by clicking the underlined screen text.

Slide 6 of 16

There are several weights available for use in analysis of HSLs:09 data. Information presented in this section of the module will help data users understand how to choose the most appropriate weight for their analysis.

As was discussed in Common Module 3, weights should be used in analyses of data from studies like HSLs:09 in order to make estimates produced from the sample representative of the target population. In HSLs:09, the sampling weights are designed to make the sample representative of the population of children enrolled in 9th grade in the United States during the 2009-2010 school year.

As discussed earlier in this module, HSLs:09 is not a simple random sample. The weights account for probability of selection.

Additionally, not everyone sampled responded to HSLs:09. Therefore, the weights also account for differential response patterns within a given round, as well as across time, to assure that the data are still representative of the population even though some members of the sample do not respond to the survey.

Slide 7 of 16

Ideally, there would be a nonresponse adjusted weight available for every component of every round of data collection. However, in longitudinal studies like HSLs:09 that have multiple components across multiple rounds of data collection, there are a large number of possible weights that can be created for analysis of data within and across rounds. It is not economical, nor useful in a practical sense, to create nonresponse adjusted weights for every combination of components across every round of data collection. Therefore, each researcher will need to decide which of the weights provided is the best one to use, given a specific research question.

Slide 8 of 16

Before proceeding into a discussion about the process of weight selection, it is helpful to review the study conventions for naming weight variables because the discussion that follows makes reference to the weighting variables by their names. Every HSLs:09 weight name begins with the letter "W" to indicate it is a weight. The second character of the weight variable name indicates the round of data collection. A "1" indicates a base year weight, a "2" indicates a first follow-up weight, and so on.

Slide 9 of 16

Here is a table of some of the weights that are available for use in the analysis of HSLs:09 data. The table shows the weights available on the HSLs:09 data file and gives their variable name on the data file, a description of the weight, and the component for which the weight has adjusted for nonresponse.

Earlier in the module it was noted that the set of weights available does not include weights that account for nonresponse to every component, or every combination of components, for every round of data collection. Therefore, it may not be possible to find the perfect weight, or one that adjusts for nonresponse to every component, and only those components, at every round from which data are being used.

If no weight corresponds exactly to the combination of components included in the analysis, researchers might prefer to use a weight with nonresponse adjustments for more components than are included in the analysis. Such a weight may result in a slightly smaller analytic sample than the perfect weight would, because it factors in nonresponse to one or more additional components. However, the weight will adjust for nonrandom nonresponse associated with each of the components that it covers, including the ones of interest in the analysis.

Alternatively, a researcher may choose a weight with nonresponse adjustments to fewer components than are included in the analysis. This would result in a larger analytic sample and may be the better, or even necessary, decision when the analysis focuses on a relatively small subgroup. In this situation, the researcher should keep in mind that estimates produced using such a weight could potentially be biased since the weight does not include nonresponse adjustments for all components in the analysis. The degree to which key data are missing should be examined to get a sense of how likely the data are to be biased.

For example, the first student analysis weight in the table is W1STUDENT. This weight accounts for (1) base-year school nonresponse and (2) student nonresponse in the base-year study. All records for sample students who participated in the base year will have a positive (non-zero) weight. Estimates generated with this base-year student weight are associated with the HSLs:09 target population of ninth-grade students. This weight can be used for the analysis of base-year student assessment scores or survey data, alone or in combination with the school characteristics or administrator/counselor data.

Now, consider W2STUDENT. This weight accounts for (1) base-year school nonresponse and (2) student nonresponse in the first follow-up only (regardless of the student's base-year response status). All records for sample students who participated in the first follow-up will have a positive (non-zero) weight. The estimates generated with these weights are associated with the HSLs:09 target population of ninth-grade students. This weight can be used for the analysis of first follow-up student assessment scores or survey data, alone or in combination with the school characteristics, administrator/counselor data from either round of HSLs:09, or teacher data from the

Weights, Variance, and Missing Data

base year. If the analysis includes base year student data, researchers are encouraged to consider W2W1STU.

For more information on the development of the analytic weights and how to select a weight for your analysis please see Chapter 6 of the data file documentation, also referred to as the DFD.

Slide 10 of 16

In Common Module 4, two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization. This portion of the HSLs:09 training modules describes important information regarding the inclusion of study design variables in your statistical package to calculate standard errors.

Replication is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates. If this is the method you will use to calculate standard errors, you need to select replicate weights that are associated with your main sampling weight. Every main sampling weight in HSLs:09 has a set of replicate weights that are associated with it. For example, the replicate weights associated with the main sampling weight W2STUDENT are W2STUDENT001 through W2STUDENT200.

HSLs:09 replication weights use a Balanced Repeated Replication method so that is the method that should be specified within the statistical software used for analysis.

Slide 11 of 16

The second method for calculating standard errors is the Taylor series linearization method. This method uses PSU and strata identifiers to compute the appropriate standard errors. This method of standard error calculation is only available to users with access to restricted-use data, as the stratum and PSU variables are not available on the public-use file to minimize disclosure risk. If this is the method you will use to compute standard errors, you will select the PSU and strata identifiers named 'PSU' and 'STRAT_ID', and specify Taylor Series linearization as the method within the statistical software used for analysis.

Slide 12 of 16

Reserve codes represent different types of missing data. The reserve code values listed here apply to all variables throughout HSLs:09. A negative value scheme for the reserve codes has been adopted so that users can easily exclude missing data without having to identify each value for each variable explicitly.

Researchers should note that different studies within NCES may use different reserve codes to mean different things.

Weights, Variance, and Missing Data

For HSLs:09, a reserve code of -5 indicates that data have been suppressed on the public-use data file. Information regarding how to obtain restricted-use data is provided in Common Module 5.

A reserve code of -7 indicates items that are programmatically skipped based on rules in the questionnaire and are not applicable to those respondents. Negative eight reserve codes indicate that data are not available because of unit nonresponse or the interview component did not apply (for example, student has no math class, thus the math teacher interview does not apply). Reserve codes of -9 indicate item level missing where the question may apply to the respondent but is not answered, or the question is not administered because the gate or introductory question is not answered.

-1, -2, -3, -4, and -6 are reserved for subsequent rounds where new reserve code values may apply.

Slide 13 of 16

Let's examine two variables in HSLs:09 to illustrate missing data, the use of reserve codes, and proper handling of missing data in your analyses. The first variable, S1EDUEXPECT, is a student questionnaire item from the base year (as the first two characters provide component and round identifiers). Further examination of the variable label tells us that S1EDUEXPECT was the first item within section G of the student questionnaire. The second variable we will examine, S1SURECLG, is the second item in section G of the base year student questionnaire.

Slide 14 of 16

First let's examine S1EDUEXPECT. Students were asked to indicate how far in school they think they will get. There were 308 students that did not answer this question, as indicated by the reserve code of -9. Valid values of 1 through 11 range from less than high school to completion of a professional degree, but also include students who indicated they 'Don't know'.

Now let's move on to S1SURECLG. Students were asked how sure they were that they would go to college to pursue a BA or BS. We see that 397 students did not answer the question, as indicated by the reserve code of -9. We also see that this item was not applicable (or legitimately skipped, meaning respondents were routed away before this question based on their response to a previous question) for 8,546 students. Additionally we can see the range of student responses – from very sure about going to very sure about not going – with valid values of 1 through 4.

Now let's consider these variables together to better understand HSLs:09 sample students' anticipated college plans at the base year data collection. The 8,546 students for which S1SURECLG was skipped are the same students, who in item one, reported that they expected to complete less than high school, complete high school or earn a GED, start an Associate's degree, complete an Associate's degree, or didn't really know how far they would get in school.

Weights, Variance, and Missing Data

Now that we have examined these variables together, we know more about the two types of missing data presented in this table. The first missing data – those who did not provide a response to S1SURECLG - are coded as -9; the second – those who don't expect to go to college – are coded as -7.

In this last table, you can see that the analyst has recoded the students who have the reserve code of -7 on S1SURECLG to 'does not expect to go' which more accurately details the intent of the sampled 9th graders to go to college to pursue a BA or BS. This example stresses the importance of understanding the data relevant to your research question, as well as properly handling missing data to ensure accurate analysis and reporting.

Slide 15 of 16

This slide discusses some common questions about the use of weights in analyses of HSLs:09 data.

When selecting a weight, you do not need to subset your data. Cases without a valid or non-zero weight will automatically be excluded from your weighted analysis. It may, however, be useful to limit your preliminary, exploratory, unweighted analysis to the cases that have a valid non-zero weight in order to better understand the sample sizes and relationships between the variables that will be in your final weighted analysis.

Your weight selection will not change if you are analyzing a subsample of the full population (for example, Hispanic children or children living in single-parent households). As discussed above, your weight selection should be based on the components and rounds that are the sources of the data in your analyses.

Weight selection also does not vary by the type of statistical procedure being used for analysis. All analyses, both univariate and multivariate, should be weighted.

Slide 16 of 16

This module provided information about the weights and variables associated with calculating estimates and standard errors for HSLs:09 data. This module detailed how weights must be applied and standard errors must be calculated to assure that estimates made from the data are representative of the population and that hypothesis tests are accurate specifically in relation to analyses of data from HSLs:09.

Additionally, this module described the importance of handling missing data in HSLs:09, specifically which missing data codes are used in the data file and an example of how to ensure accurate data analysis.

Additionally, important resources that have been provided throughout the module are summarized here along with the module's objectives for your reference.

You may now proceed to the next module in the series or exit the module.