Slide 1 of 31

Title Slide: ECLS-B Sample Design, Weights, Variance, and Missing Data

Slide 2 of 31

This module provides information about the weights and variables associated with estimates from the ECLS-B data. The ECLS-B study 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 the data are representative of the population and that hypothesis tests are accurate. Specifically, weights must be applied and standard errors must be computed taking the complex sample design into account. This module discusses these issues specifically in relation to analyses of data from the ECLS-B.

Additionally, this module describes how to handle missing data in the ECLS-B data, specifically which missing data codes are used in the data file and some tips for handling missing data to ensure accurate data analysis.

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

Slide 3 of 31

The ECLS-B sample was selected using a multi-stage, clustered design, meaning that data were not collected using a simple random sample (or SRS) of the target population. Instead, cases were clustered within primary sampling units (or PSUs) in order to reduce field costs.

In the first stage of sampling, the country was separated into primary sampling units, which were counties or groups of contiguous counties. The PSUs were sampled with probability proportional to size and with stratification based on selected characteristics such as region, median household income, percent minority, and metropolitan status.

In the second stage of sample selection, birth certificates were sampled from within the sampled PSUs.

The ECLS-B includes an oversample of American Indian/Alaska Native children. Therefore, in addition to the primary PSU sample selected from among all PSUs in the country, a supplementary sample of PSUs was selected from among areas that had a high concentration of American Indian/Alaska Native births. This supplementary sample of American Indian/Alaska Native PSUs was included to ensure that the study selected sufficient numbers of children who were American Indian/Alaska Natives to support reliable estimates for this group of children.

Slide 4 of 31

Because of clustering, the sampled children live in closer geographical proximity to each other than would have been the case if simple random sampling had been conducted.

Children who live near one another, for example within one PSU, tend to be more similar to one another on many characteristics than do children who live in a wider area, for example in different PSUs.

Therefore, as a result of clustering, variation in the sample is lower than it would be if the children were selected through a simple random sample.

Slide 5 of 31

In addition to the clustered design, the ECLS-B oversampled children with certain characteristics.

As mentioned when describing the sampling of PSUs, the ECLS-B oversampled American Indian/Alaska Native children. Additionally, the ECLS-B oversampled children born with low or very low birth weight, Chinese and other Asian American or Pacific Islander children, and twins.

Because of the clustering and oversampling in the ECLS-B, users will need to use weights and appropriate procedures to compute standard errors to account for the complex sample design when doing hypothesis testing.

Slide 6 of 31

There are many sampling weights available for use in analysis of ECLS-B data. Information presented in this section of the module will help you understand how to choose the most appropriate weight for your analysis.

As was discussed in Common Module 3, weights should be used in analyses of data from studies like the ECLS-B in order to make estimates produced from the sample representative of the target population. In the ECLS-B, the sampling weights are designed to make the sample representative of the population of children born in the United States in 2001.

As discussed earlier in this module, due to clustering primary sampling units that were selected with probability proportional to size, and the use of oversampling for certain small population groups, the members of the ECLS-B population did not all have the same chance of being selected into the study sample. The weights account for these differential probabilities of selection.

Additionally, the weights 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 31

For every sampled child, multiple weights are provided for analysis using data from each round of data collection. ECLS-B weights sum across the sample cases to the population of children born in the United States in 2001. For components that have additional eligibility requirements, the sum of the sampling weights is the number of individuals in the target population who meet the criteria for eligibility for the given study component. Every ECLS-B weight is adjusted for nonresponse to one or more of the data collection components used to collect information about these children and their experiences over time. Estimates produced using these nonresponse adjusted weights are representative of the characteristics and experiences of the population of children born in the United States in 2001 even though some of the sampled children did not participate in the assessments and some of their parents and care providers did not participate in the study.

Slide 8 of 31

Development of the ECLS-B weights reflects the longitudinal design of the study and the requirements for participation. To be eligible for participation in a round after the 9-month data collection, a parent interview must have been completed in the prior round. Also, to be eligible for the different components within a round of data collection (for example, the child assessment, the father questionnaires, the child care provider interview, and the child care observation in the 2-year data collection), a parent interview must have been completed for that round. Thus, the process of developing weights for each round of data collection in the ECLS-B was sequential and began with the development of a weight that sums to the population of children and has adjustments for nonresponse to the parent interview in that round. The first weight ever created for the study was the weight adjusted for nonresponse to the 9-month parent interview. This became the base weight for other weights developed for the 9-month collection. For example, the weight adjusted for nonresponse to the child assessment is the weight that is to be used when analyzing data from the child assessment in the 9-month data collection. As a result of this sequential weight development process, all weights for a given round of data collection are adjusted for nonresponse to the parent interview at that round.

The weight adjusted for nonresponse to the parent interview in any round is also used as the base weight for the weight adjusted for nonresponse to the parent interview in the subsequent round, meaning that the parent interview nonresponse adjustments are cumulative. Thus, all of the ECLS-B weights include nonresponse adjustments for the parent interview in a given round, as well as for prior rounds, with one exception related to American Indian and Alaska Native children, which is discussed later in this module.

Slide 9 of 31

As will be seen in the discussion that follows, selection of a weight is driven primarily by two things: the sources of data about the children and their experiences that are being used in the analysis; and consideration of which weight is adjusted for nonresponse to those sources. Given that the ECLS-B weights are adjusted for nonresponse, ideally there would be a nonresponse adjusted weight available for every component of every round of data collection. However, in longitudinal studies like the ECLS-B that have multiple components across multiple rounds of data collection, there is 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 weight is the best one to use, given a specific research question.

Slide 10 of 31

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 ECLS-B weight name begins with the letter “W” to indicate it is a weight.

Other letters used in a weight name indicate the component or components of the ECLS-B for which nonresponse adjustments are made.

As previously mentioned, all the ECLS-B weights are designed to make the ECLS-B sample representative of the target population: children born in the United States in 2001. The different weights available in the data file are adjusted for nonresponse to different data collection components or respondent types, but the unit of analysis is always the child.

As can be seen here, an “R” indicates the weight is adjusted for nonresponse to the parent interview. A “C” indicates the weight is adjusted for nonresponse to the direct child assessment, a “J” indicates the weight is adjusted for nonresponse to one of the early care and education provider interviews, a “P” indicates the weight is adjusted for nonresponse on the early care and education setting observation, an “F” indicates the weight is adjusted for nonresponse to one of the father questionnaires (either the resident or nonresident questionnaires, or both, depending on which ones the child was eligible for), and a “D” indicates the weight is adjusted for nonresponse to the resident father questionnaire. The “D” is used for the preschool round because there was no nonresident father questionnaire fielded in that round of data collection.

Slide 11 of 31

Numbers used in weight variable names indicate a few different things. The second character in a weight name is always a number, and this number indicates the round of data collection for which the weight was developed. It also indicates the latest round from which data in a given analysis should be used. For example, if an analysis includes data from the 9-month, 2-year, and preschool rounds, the latest round of data in the analysis is preschool, or round 3. A weight that has "3" as the second character should be selected.

Other numbers indicate the round of data collection for the components for which nonresponse adjustments have been made. As can be seen here, a "1" indicates the variable is associated with round 1, or the 9-month data collection, a "2" indicates the variable is associated with round 2, or the 2-year data collection, and so on. The exception to this rule can be seen in the last row; a "K" is used to indicate that the variable is associated with data or analyses that pertain to the child's year of kindergarten entry, which, as discussed in the introductory module, could have been either the kindergarten 2006 or the kindergarten 2007 round of data collection.

The next slide uses specific examples to clarify the naming conventions described here.

Slide 12 of 31

Here are some of the weights that are available for use in analyses of ECLS-B data. The first weight listed is W1R0. The "W" indicates that this is a weight variable. The "1" indicates this weight was developed for analysis of data from round 1 and that round 1 is the latest round from which data in the analysis should be taken. This means that this weight should only be used with analyses that include just data from the 9-month round.

The "R" indicates that this is the weight adjusted for nonresponse to the parent interview. Since this is the first round of data collection, by default the adjustments are for the parent interview at round 1. Notice that this weight ends with a zero; although this is a number, it is not associated with a round of data collection. Rather, the zero indicates that this is a main sample weight, distinct from replicate weights which are discussed later in this module.

The second weight listed is WKR0. Again, the "W" indicates that this is a weight variable. The "K" indicates this is a weight that is adjusted for nonresponse to the parent interview in the round in which the study child was in kindergarten for the first time. The "R" indicates that this weight is adjusted for nonresponse to the parent interview. So, this weight is for the kindergarten 2006 parent interview for children who were in kindergarten for the first time in the 2006 data collection and the kindergarten 2007 parent interview for children who were in kindergarten for the first time in the 2007 data collection. Again, the zero indicates this is the main sample weight. As discussed earlier, WKR0 includes adjustments for nonresponse to interviews in prior rounds of data collection. Any case had to have a complete parent interview in the previous round

for inclusion in the next round. Therefore, analyses that use the WKR0 weight may also include parent interview data from rounds 1, 2, or 3.

The weight W3D0 is adjusted for nonresponse to the resident father questionnaire in preschool (or round 3). Also, although there is no “R” in the weight name, this weight is adjusted for nonresponse to the parent interview in rounds 1, 2, and 3 as well.

Some ECLS-B weights account for nonresponse on multiple components simultaneously and this can be seen in the weight name. For example, the last weight listed here is W523J0. The “5” indicates that this weight is adjusted for nonresponse to a component in round 5. This is the second character, indicating this is the latest round for which nonresponse adjustments were made. The 2 and 3 indicate that nonresponse adjustments were also made for components at rounds 2 or 3, or both. The “J” indicates that adjustments were made for nonresponse to the early care and education provider telephone interview. Thus, this weight adjusts for nonresponse to the child care provider interview at round 2 and the early care and education provider interview at round 3. Since it is a “W5” weight, it also adjusts for nonresponse to the parent interview through wave 5.

While these naming conventions were adhered to, it is not always particularly easy to determine the exact components for which nonresponse adjustments were made from the variable name alone. The user documentation contains detailed descriptions of the nonresponse adjustments for each weight, and users are strongly encouraged to read that documentation when selecting a weight for analysis.

Slide 13 of 31

There are two main considerations when deciding which weight to use in a given analysis: the rounds of collection from which the data are taken and the components that provide the data.

With respect to the round of data collection, a weight should be selected that corresponds to the latest round of data included in the analysis. For example, if an analysis includes data from the preschool round, or round 3, and does not include any data from the later kindergarten rounds, a weight associated with the preschool round should be chosen. The preschool weights all begin with “W3.”

The second consideration is the component or components from which the data are taken. A weight should be chosen that adjusts for nonresponse to the components included in the analysis. If data are taken from more than one component, the weight that is adjusted for nonresponse to the greatest number of components should be selected.

Some examples of selecting a weight are provided in later slides in this module.

Slide 14 of 31

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 also choose a weight with nonresponse adjustments to fewer components than are included in the analysis. This would result in a larger analytic sample and, therefore, 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.

Slide 15 of 31

The user documentation includes a series of tables to help users decide which weight to select for their analysis. This table shows the table listing the four weights developed for the 9-month data collection. A “yes” in a cell indicates that nonresponse adjustments were made for that component. When selecting a weight, it is best to choose the one with a “yes” in the largest number of cells corresponding to the components in your analysis.

For example, consider an analysis of the children’s experiences at birth that includes data collected in the 9-month parent interview. The ideal weight would have a “yes” in the 9-month parent interview column, and only that column, because the “yes” indicates that the weight is adjusted for nonresponse to only that component. Looking at the table, there is a “yes” in just the parent interview column for weight W1R0, so that is the best weight to use. An analysis that includes data from the parent interview, child assessment, and resident father questionnaire should use weight W1FC0, because that weight is adjusted for nonresponse to all three of those components. If an analysis includes data from just the parent interview and resident father questionnaire, W1F0 would be a better weight to use than W1FC0. Although both of these weights are adjusted for nonresponse to these two components, using W1FC0 would result in the analysis excluding children who did not also have child assessment data. Since child

assessment data are not being used in the analysis, use of W1FC0 limits the analytic sample unnecessarily.

Note that the 9-month parent interview column is “yes” for all the weights listed here, reflecting the point made earlier that the 9-month weight adjusted for nonresponse to the parent interview serves as the weight from which all other 9-month weights are developed.

Slide 16 of 31

These tables show the weights developed for the 2-year data collection. As can be seen here, after 9 months, the number of combinations of components across rounds becomes much larger. As a result, the tables become somewhat more complex to use, but the basic method of using information in the table remains the same: choose the weight with a “yes” in the largest number of cells corresponding to the components in your analysis.

For example, consider an analysis of the early care and education arrangements children had when they were about 2 years old that includes data collected in the parent and child care provider interviews. The ideal weight would have a “yes” in the 2-year parent interview and 2-year care provider interview columns, and only those columns, because the “yes” indicates that the weight is adjusted for nonresponse to only these components. Looking at the table, there is a “yes” in these columns, and only these columns, for weight W22J0, so that is the best weight to use. While there are several other weights that have a “yes” in the parent and care provider columns, those weights also have a “yes” in at least one other column, indicating the weights are adjusted for nonresponse to at least one other component that is not included in the analysis. Selecting those weights would limit the analytic sample unnecessarily, because the analysis would exclude cases that have data for the parent and care provider interviews but not for the other component that is not contributing data.

Again, note that the 9-month parent interview column is “yes” for all the weights, reflecting the point made earlier that all weights are adjusted for nonresponse to the parent interview in earlier rounds. For this reason, presence of “yes” in the 9-month parent interview column is not of concern in selection of the ideal weight for analyses involving 2-year data.

Tables such as these for weights from all rounds of data collection can be found in the user documentation that accompanies the restricted-use data file.

Slide 17 of 31

In the 2-year, preschool, and kindergarten rounds of data collection, the response patterns for the child assessment component were determined to be so close to the response patterns for the parent interview that no separate weights needed to be developed for use in analyses that include child assessment data from just one of those rounds. That is, nonresponse to the parent interview was very similar to nonresponse to the child assessment. Said another way, if a parent interview was completed, in almost

all cases a child assessment was also completed. Therefore, when analyzing child assessment data from just one of these later rounds, the “R” weight is sufficient for analysis.

In contrast, the response pattern for the 9-month child assessment did differ significantly from that of the parent interview at 9 months and a separate weight that was adjusted for nonresponse to the 9-month child assessment was developed. This is the W1C0 weight. Any analysis that includes data from the 9-month child assessment should use a weight that is adjusted for nonresponse to that assessment. All of the “C” weights include such an adjustment.

The “C” weights are also adjusted for nonresponse to the child assessment at every round up through the round indicated by the second character of the weight name. For example, W3C0 is adjusted for nonresponse to the child assessments at 9 months, 2 years, and preschool.

A researcher may want to use a “C” weight when analyzing child assessment data for more than one round, even if 9-month child assessment data are not included, since the “C” weight is adjusted for nonresponse to the child assessment at multiple rounds.

To illustrate these points, consider an analysis that includes data from the parent interview in rounds 1 and 4, as well as child assessment data from round 4. If the only child assessment data included in such an analysis come from round 4, then the weight that is adjusted for nonresponse to the round 4 parent interview, W4R0, should be selected. Remember that this weight has implicit adjustments for nonresponse to the parent interview at round 1. If round 1, or 9-month, child assessment data are added to the analysis, then the best weight to use becomes W4C0, because this weight is adjusted for nonresponse to the 9-month child assessment. This weight is also adjusted for nonresponse to the 2-year, preschool, and kindergarten 2006 child assessments. Even though child assessment data from those rounds are not included in the analysis, this weight should still be selected because the estimates could be biased without adjustments for nonresponse to the 9-month assessment. If the analysis includes 2-year and preschool child assessment data but not 9-month child assessment data, an analyst may still want to use W4C0 because it is adjusted for nonresponse to the child assessment in the 2-year, preschool, and kindergarten 2006 data collections.

Slide 18 of 31

As mentioned in the introductory module, to be eligible for the subsequent follow-up in the ECLS-B, a case had to have a completed parent interview from the prior round. The one exception to this is American Indian/Alaska Native, or AIAN, children. To preserve the sample of AIAN children, all AIAN children were eligible for the preschool and kindergarten data collections if they had a complete 9-month parent interview, regardless of whether they had a complete parent interview in other rounds. For example, a complete parent interview in the 2-year data collection was not needed for AIAN cases to be eligible for the study in the preschool round, as long as they had a 9-month parent interview.

Since the primary parent interview, or "R," weight developed for the data was adjusted for parent interview nonresponse at every round of data collection, a specific analytic weight had to be developed for the AIAN cases that was not contingent on response to every data collection after the 9-month parent interview.

For example, the preschool weight W31R0 is positive for children with a parent interview at both the 9-month and preschool rounds of data collection, regardless of the response status at 2 years, and includes 2-year American Indian/Alaska Native nonrespondents who responded at the preschool wave. Similar weights are created for analysis of these children at kindergarten entry. This weight is intended for use by researchers specifically interested in analyzing American Indian/Alaska Native children. It is not recommended that analysts who are not specifically interested in AIAN children use these weights because they adjust for less nonresponse than the other weights.

Slide 19 of 31

Now that we've reviewed the purpose of weights and discussed the weights that are available for the ECLS-B, let's take a look at the actual weight variables in the ECLS-B data. Here is a very small sample of ECLS-B cases and their values on the four weights created for analysis of the data collected in the first round of data collection, when the children were about 9 months of age.

Cases that are considered respondents to a given survey component have a valid weight value for that component. For example, looking at the weight W1R0, which adjusts for nonresponse to the 9-month parent interview, we see that every case has a valid weight. This is because a case had to have a complete parent interview at 9 months to be considered a study participant and have the data included in the data file. The actual value of a weight indicates how many children in the target population that case is representing. For example, case 1 represents about 1002 children in the population, while case 6 represents about 63 children in the population. The weight assigned to any given case depends on a variety of factors, including the number of children with similar characteristics in the population and the number of children with similar characteristics in the study sample who were nonrespondents for a given component.

Cases that did not respond to a particular component do not have a valid analytic weight for analysis of data from that component. For example, sampled children whose resident or nonresident father did not complete a father questionnaire do not have a valid W1F0 weight. This is indicated with the period in the cell for weight W1F0 in the table shown here. Other responding cases with similar characteristics that do have a completed father questionnaire will have a larger analytic weight to compensate for missing data for cases that are nonrespondents.

Since W1FC0 adjusts for nonresponse to both the child assessment and the father questionnaire, children who do not have valid data for **BOTH** components do not have a valid value for W1FC0.

Slide 20 of 31

This table shows the means and sums of the same four weights created for the 9-month round of the ECLS-B. The weights are designed to make the sample data representative of the target population. Therefore, the weights for each case in the sample sum to the population totals. In the ECLS-B, the weights W1R0 and W1C0 sum to the general target population of children born in the United States in 2001, which was about 3.9 million children. For components that have additional eligibility requirements, the sum of the sampling weights is the number of individuals in the target population who meet the criteria for eligibility for the given study component. For example, the W1F0 and W1FC0 weights sum to the population of children born in 2001 who had a resident father or a nonresident father who met the eligibility requirement for the nonresident father questionnaire. Note that the number of cases with a valid W1F0 weight is lower than the number of cases with a valid W1R0 weight because not all children lived with a resident father or had a nonresident father who met the eligibility requirements for the nonresident father questionnaire.

The last column in the table shows the mean and sum for the weight adjusted for nonresponse to the parent interview in the 2-year data collection, W2R0. The number of cases with a valid W2R0 weight is lower than the number of cases with a valid W1R0 weight because not all children who participated in the 9-month data collection participated again in the 2-year data collection.

This table also illustrates that, as the number of cases with valid data on a given component or combination of components decreases, the average weight for those cases that do have data on that component must increase in order to still sum to the population total. As shown here, about 10,700 cases have a valid value for the weight that adjusts for nonresponse to the parent interview at 9 months, W1R0. These are cases that have valid 9-month parent interview data. The average weight for these cases, that is, the average number of children in the population each case represents, is about 374. Fewer cases had valid parent interview data in the 2-year collection, so the mean value of W2R0 is higher than that of W1R0. On average, each case with 2-year parent interview data represents about 403 cases in the population.

As a last point to note here, the number of cases with a valid weight can be a useful piece of information to look at when selecting a weight. If a decision must be made between two different weights, this information indicates how many cases would be in the analytic sample if a given weight were selected.

Maximization of the analytic sample size also explains why an analyst should not select a weight with nonresponse adjustments for rounds of data collection after the latest round of collection from which data are being used. For example, if weight W2R0 was used in an analysis using only 9-month parent interview data, meaning an analysis for which 2-year nonresponse adjustments are unnecessary, the analytic sample size would be 9,850, rather than 10,700.

Slide 21 of 31

In Common Module 4, two standard error calculation procedures were discussed: Replication Techniques and Taylor Series approximation. This portion of the ECLS-B 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 the ECLS-B has a set of 90 replicate weights that are associated with it. For example, the replicate weights associated with the main sampling weight W1R0 are W1R1 through W1R90, while those for the child sampling weight are W1C1 through W1C90.

The paired-jackknife, or JK2, method is the appropriate replication technique to use when computing standard errors for ECLS-B data, given the specific features of the ECLS-B complex sample design. JK2 should be specified as the replication method within the statistical software used for analysis.

Slide 22 of 31

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. If this is the method you will use to compute standard errors, you need to select the PSU and strata identifiers that are associated with your sampling weight variable. As with the replicate weights, each sampling weight variable has PSU and strata variables associated with it. For example, the Taylor Series variables associated with the main sampling weight W1R0 are W1PSU and W1STR.

Slide 23 of 31

Data can be missing across the ECLS-B for a variety of reasons. Standard missing data codes are used in the ECLS-B data file to identify different types of missing data.

When a data value for a case is blank, or system missing, this is an indication that there are no data for an entire instrument and can be considered unit nonresponse. For example, if a parent interview was not completed in the 2-year data collection, a case will have a system missing code for all variables for all components in that round of data collection. If a direct child assessment was not completed in the 2-year data collection but a parent interview was, a case will have a system missing code for all variables associated with the child assessment but none of the variables associated with the parent interview will have a system missing code.

The other codes listed here are used for item-level missing data. These are data that are missing within an otherwise complete interview, questionnaire, or assessment, for

example if a respondent did not know the answer to one question in a section about her spouse's employment but was able to answer the other questions in that section.

A missing data code of -1 indicates that an item was not asked because of a planned skip pattern. Most often this code is used because the item skipped was not applicable for that particular case. For example, an item asking about the number of hours a child spends in center-based nonparental care would be skipped if the parent indicated that the child did not have a center-based child care arrangement. In addition, if a lead item was left blank on a paper-based data collection instrument – for example, if a lead item is skipped – subsequent items that are tied to the skipped variable are coded as not applicable and assigned a missing code of -1.

A missing data code of -4 is used in rare instances in which there was a problem in the administration of an item that led to a high proportion of cases having missing data on the affected item, and the data that were collected were not meaningful. Although the administration error typically did not affect all cases, the -4 missing data code is assigned to all cases, whether or not the specific case had a response or was missing data due to the error. This was done to prevent users from running analyses on highly problematic data.

A -7 missing data code indicates that a participant either stated or, in the case of self-administered instruments wrote, that he or she refused to answer an item. A missing data code of -8 indicates that a participant either stated or wrote that he or she did not know the answer to a question. The -7 and -8 missing data values rarely occur in the ECLS-B data file.

A missing data code of -9 indicates that the response to an item was not ascertained or, in the case of composite variables, not possible to calculate. For example, a -9 code would be used if a respondent did not provide an answer to a question on a hard-copy questionnaire and a composite score for the NCATS assessment would be coded -9 if there were too few items for which data were available to calculate the score.

Slide 24 of 31

There are a few exceptions in the use of standard missing data codes in the ECLS-B dataset. Missing values for the T-scores derived from the BSF-R are coded as -99. The birth certificate data are taken from the National Center for Health Statistics natality files and retain the reserve codes from those files. Those data use several different missing value codes with labels that vary across variables. More information on these codes can be found in the technical documentation of birth certificate variables on the NCHS natality file. For the fine and gross motor child assessment items, codes of 95, 96, and 97 were used to identify cases for which the item was uncodeable, not administered, or had no response respectively.

Slide 25 of 31

The missing value codes will not be specified as missing data when you initially create your data set. You must identify them as missing values through recoding so that the statistical software does not treat them as valid values. It is important to understand why data are missing in order to appropriately recode such data for your analysis. This last section of the module provides an example of how different approaches to recoding missing data can impact results.

Consider these two questions from the 9-month Parent Interview.

Question CH010, says: “Please think back to when [CHILD/TWIN] was born. As a newborn, did [CHILD/TWIN] have to stay longer in the hospital because of medical problems?”

Question CH015, asks “How many days did [CHILD/TWIN] stay in the hospital because of medical problems?”

Question CH015 is only asked of respondents who said that, yes, their child had to stay in the hospital longer due to medical problems in question CH010. Data for respondents who said no, their child did not have to stay in the hospital longer due to medical problems are coded -1, not applicable, for question CH015.

There are two different options available for recoding missing data values for CH015. You can elect to recode all missing values, including not applicables, to be treated as missing data, or you can recode some to missing and convert others to valid values based on the response to CH010.

The type of missing data recode you choose should be based on your research question of interest.

Slide 26 of 31

Here are two research questions someone may be interested in investigating with the ECLS-B data.

For research question number 1, “What is the average number of days that infants stayed in the hospital after birth because of medical problems?”, the analysis should be conducted using all infants as the analytic group.

For research question number 2, “Of those infants who had to spend some time in the hospital after birth because of medical problems, how many days on average did they spend in the hospital?”, the analysis should be conducted using only the subpopulation of infants that spent time in the hospital after birth due to medical problems as the analytic group.

Slide 27 of 31

To address research question 1 about the average number of days that infants stayed in the hospital after birth because of medical problems, cases coded not applicable, or -1, on CHQ015, can be recoded to a valid value based on the response to question CHQ010 .

Specifically, if the answer to CH010 is 'No,' then the -1 code for CH015 about the number of days in the hospital can be recoded to a value of 0, because we know that the child did not spend extra time in the hospital due to medical problems.

The remaining missing data values for CHQ015, that is -1, -7, -8, and -9, can be identified as missing data so that they are not treated as valid values.

Slide 28 of 31

To address research question 2 about the average number of days that infants stayed in the hospital after birth because of medical problems among infants who did stay in the hospital, all missing values for question CHQ015 can be identified as missing data so that they are not treated as valid values. This way, the average will be calculated using only valid responses to question CHQ015.

Slide 29 of 31

The table provided here shows the different estimates that are obtained when missing data are treated different ways.

The first estimate, a weighted mean of 1.9 days, is the result for ALL infants, with the -1 missing data value recoded to a valid value of 0 for those children who spent no time in the hospital after birth because of medical problems.

The second estimate, a weighted mean of 14.7, is the result for only those infants who did have to spend extra time in the hospital, with all missing data values identified as missing data.

The third estimate, a weighted mean of 1 day, is incorrect. It is the result that is obtained when no missing data are recoded. In this analysis, -1, -7, -8, and -9 are all treated as valid values, which has the effect of decreasing the calculated mean.

Slide 30 of 31

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

Additionally, this module described how to handle missing data in the ECLS-B data, specifically which missing data codes are used in the data file and some tips for handling missing data to ensure accurate data analysis.

Slide 31 of 31

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