Slide 1 of 29

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

Slide 2 of 29

This module provides information about the weights and variables associated with estimates from for the ECLS-K data.

The ECLS-K 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 in relation to analyses of data from the ECLS-K.

For information about weighting and calculating appropriate standard errors more generally, please view the relevant common modules, if you have not already done so.

Additionally, this module describes how to handle missing data in the ECLS-K 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 3 of 29

The ECLS-K sample was selected using a multistage, stratified, clustered design, meaning that 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 first stratified based on selected characteristics such as region, median household income, and metropolitan status. Then the PSUs were sampled within strata with probability proportional to size, where size was operationalized as the number of 5-year-olds in the PSU.

In the second stage of sampling, public and private schools offering kindergarten programs or educating kindergarten-aged children were selected within sampled PSUs.

The third-stage sampling units were children in kindergarten or of kindergarten age, who were selected within each sampled school.

Slide 4 of 29

Because of clustering, the sampled children lived in closer geographical proximity to each other and were more likely to attend the same school than would have been the case if simple random sampling had been conducted.

Children who live near one another or attend the same school tend to be more similar to one another on many characteristics than do children who live in a wider area or attend different schools.

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

Slide 5 of 29

In addition to the clustered design, the ECLS-K oversampled children of Asian or Pacific Islander decent. Such children were sampled at 2.5 times the rate of non-Asian/Pacific Islander children.

Additionally, after the first year of the study, which is referred to as the base year, subsampling of the full sample was used in different ways to reduce field costs. The fall first-grade data collection was conducted with a 30 percent subsample of schools that participated in the base year. Subsampling was also used for some children who moved from their original sampled school after the base year. While ideally the study would have followed all base-year respondents who moved from their original schools after the spring of kindergarten, it would have been expensive to do. Significant effort had to be made to locate students in their new schools and to obtain permission to assess them in their new schools. Also, as the study progressed, student mobility had a more serious impact on the cost of collecting data because the number of schools children attended increased. For these reasons, about half of the students who changed schools were followed for continuation in the study. The probability of being followed did depend on a few factors, including language minority status and nonresponse in prior rounds.

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

Slide 6 of 29

There are many sampling weights available for use in analysis of ECLS-K 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-K in order to make estimates produced from the sample representative of the target population. In the ECLS-K, the sampling weights are designed to make the sample representative of the population of children enrolled in kindergarten in the United States during the 1998-99 school year. Due to sample

freshening in the first grade data collection round, the sample is also representative of children in the first grade in the United States during the 1999-2000 school year. In the base-year, the data are representative of kindergarten teachers and schools with kindergarten programs or educating kindergarten-aged children in ungraded settings.

As discussed earlier in this module, due to clustering primary sampling units that were selected with probability proportional to size, clustering of schools within sampled PSUs, and the use of oversampling for certain small population groups, the members of the ECLS-K population did not all have the same chance of being selected into the study sample.

Additionally, the weights account for subsampling and 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 were not followed or did not respond to the survey.

Slide 7 of 29

Multiple weights are provided for analysis using data from each round of data collection. Given that ECLS-K weights are adjusted for nonresponse, ideally there would be a weight available that is adjusted for nonresponse to every component of every round of data collection. However, in longitudinal studies like the ECLS-K that have multiple components across multiple rounds of data collection, there is a large number of possible weights that could be created for analysis of data within and across rounds. It is not economical, nor useful in a practical sense, to create a weight that accounts for nonresponse to every combination of components across every round of data collection. Therefore, researchers will need to decide which weight is the best one to use given their research question.

Slide 8 of 29

There are three main considerations when deciding which weight to use in a given analysis: the level, or unit, of analysis; the rounds of collection from which the data are taken; and the components that provide the data.

The ECLS-K sample design makes it possible to conduct analyses of the base-year data at three levels: the child level, the teacher level, and the school level. After the base year, analyses can only be conducted at the child level, because the data are not representative of teachers or schools after the base year.

With respect to the round of data collection, a weight should be selected that corresponds to the rounds from which you are using data. Some ECLS-K weights were designed for use in analyses of data collected in a single round, while other weights were designed for use in longitudinal analyses that use data from multiple rounds. For example, if an analysis includes data from the spring of kindergarten, spring of first grade, and spring of third grade, a weight associated with rounds 2, 4, and 5 should be chosen.

The last consideration is the component or components from which the data are taken. A weight should be chosen that is adjusted for nonresponse to the components included in the analysis. Some weights are adjusted for nonresponse to the child assessment, while others are adjusted for nonresponse to the parent interview or the teacher questionnaire. 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.

The next few slides discuss specific details of the weight naming conventions and procedures for selecting an appropriate weight for analysis. Several of the weights are described in more detail to illustrate these selection criteria more clearly.

Slide 9 of 29

The names of the weight variables provide information about the level of analysis for which the weights are designed to be used, the rounds of data they are to be used with, and the components for which they adjust for nonresponse.

As noted earlier, data collected in the base year may be analyzed at the child, teacher, or school level. The first element in the variable names for weights to be used with base-year data indicates the level of analysis for which the weight is to be used. Weights designed for use in school-level analyses begin with “S”; weights designed for use in teacher-level analyses begin with “B”; weights designed for use in child-level analysis with data from either the fall kindergarten round or spring kindergarten round alone begin with “C”; and weights designed for use in longitudinal child-level analysis, meaning analysis of data from both the fall and spring of kindergarten, begin with the letters “BY” for “base year.”

Data collected in the first- through eighth-grade rounds only generalize to the child level. Therefore, all weights for use with data collected from first through eighth grade, even if base-year data are also included in the analysis, begin with the letter “C,” with one exception: the weight named Y2COMW0, which is designed for child-level analyses that include assessment data from rounds 1, 2, and 4, AND parent or teacher data from the spring first-grade data collection, AND parent and/or teacher data from one or both rounds of the base year.

Slide 10 of 29

The second element in a weight name is a number indicating the round or rounds of data for which the weight was designed to be used.

Analyses of ECLS-K data may focus on either one round of data collection or be longitudinal and incorporate data from more than one round. For example, an analyst may want to conduct analyses of data from the fall of kindergarten or the spring of third grade only. Weights designed for analyses of data from only one round have a single number as the second element of the weight name. This number ranges from 1 to 7 and identifies the round of data collection for which the weight was designed to be used.

Longitudinal analyses may include data collected from two time points in the same school year, for example, both fall and spring kindergarten data, or from two or more time points across multiple years, for example, from the spring kindergarten, spring first-grade, spring third grade, and spring fifth grade collections. The second element of the names of weights designed for longitudinal analyses includes two or more numbers to indicate the rounds of data for which the weight was designed to be used. For example, a “45” indicates that the weight should be used for analysis of data from rounds four and five. The nomenclature “1_7F,” indicates that the data are for rounds 1, 2, 4, 5, 6, and 7. The “F” indicates this is a weight developed for rounds that include the full sample, that is, it excludes data collected in round 3, the fall of first grade, because this round included only a subsample of the full ECLS-K sample. When the second element of the weight name includes an “S,” such as “1_5S,” this indicates the weight is designed to be used with data collected during the fall of first grade, or round 3, in addition to data from the other rounds identified in the weight name.

It is important to remember that analyses of data collected after first grade are not representative of children at that grade. For example, the data collected in the spring of third grade are not representative of third-graders. They are representative of children who were enrolled in kindergarten in the 1998-99 school year or in first grade in the 1999-2000 school year in the spring of 2002, when the majority of the sample was in third grade.

Slide 11 of 29

The third element in a weight variable name indicates the source or sources of data for which the weight is adjusted for nonresponse.

For analyses of child assessment data alone or in conjunction with a limited set of child characteristics, such as age, sex, and race/ethnicity, a weight that includes the letter “C” alone in the third element of the weight name is appropriate.

For analyses of data collected from the parent interview, regardless of whether child assessment data are also included, a weight that includes the letter “P” alone in the third element of the weight name is appropriate.

Analyses that use child assessment data in conjunction with data collected from the parent interview and the classroom teacher questionnaire should include a weight that has the letters “CPT” in the third element of the weight name. Additionally, for the fifth-grade weights, “CPT” is followed by an “R,” “M,” or “S” indicating that the teacher data for which the weight is adjusted for nonresponse were collected from the reading, math, or science teacher, respectively. Similarly, for the eighth-grade weights, “CPT” is followed by an “E,” “M,” or “S” indicating that the teacher data for which the weight is adjusted for nonresponse were collected from the English, math, or science teacher, respectively. Separate weights for use with data from teachers of different subjects were needed for the fifth- and eighth-grade data collections because of the procedures that were used in those rounds to collect information from teachers. More information on

those procedures is provided in the module titled ‘Considerations for Analysis of ECLS-K Data.’

Slide 12 of 29

There are many sources of data in the ECLS-K, but only three of them impact the decision about which weight to select for analysis. This is because the ECLS-K weights are adjusted for nonresponse only to the child assessments, the parent interviews, and the teacher-/classroom-level teacher questionnaires. Therefore, users should make their weight selection based on the combination of these three components included in their analysis.

Slide 13 of 29

Inclusion of data from other components such as the school administrator questionnaire, the school facilities checklist, the child-level classroom teacher questionnaire, the special education teacher questionnaires, the student record abstract data, the Head Start data, and the salary and benefits data do not affect the choice of the weight because none of the weights are adjusted for nonresponse to these components. Analysts using data from these sources are encouraged to keep in mind that the weights are not adjusted for differential nonresponse to these sources and may want to examine the data for evidence of nonresponse bias.

Slide 14 of 29

This example illustrates the weight naming conventions and how the weight name indicates the kind of analysis for which the weight was designed to be used.

The “C” in weight “C23PW0” indicates that this weight is designed for use in a child-level analysis, that is, analysis that generalizes to children.

The “23” indicates that this weight is designed for use with data from rounds 2 and 3, the spring of kindergarten and the fall of first grade.

The “P” indicates that this weight is adjusted for nonresponse to the parent interview and, because of the numbers in the second element, we know it is adjusted for nonresponse to the parent interview at rounds 2 and 3.

Therefore, this weight is most appropriate for a child-level analysis of parent interview data from both rounds 2 and 3. It does not matter whether the analysis also includes child assessment data or data from other sources that do not affect the selection of a weight, such as data from the school administrator questionnaire.

Slide 15 of 29

In this example, the “C” in weight “C7CPTM0” indicates that this weight is designed for use in a child-level analysis.

The “7” indicates that this weight is designed for use with data from round 7, the spring of eighth grade.

The “CPTM” indicates that this weight is adjusted for nonresponse to the child assessment, the parent interview, and the math teacher teacher-level questionnaire. Again, because of the numbers in the second element, we know this weight is adjusted for nonresponse to these components in round 7.

Therefore, this weight is most appropriate for a child-level analysis of data from these three sources in eighth grade. Again, it does not matter whether the analysis also includes data from other sources that do not affect the selection of a weight, such as data from the school administrator questionnaire.

Slide 16 of 29

There are two weight variables for which the names do not adhere to the general principles discussed on the previous few slides: BYCOMW0 and Y2COMW0.

The weight BYCOMW0 was designed for analyses of direct child assessment data from both the fall and spring kindergarten data collections, in conjunction with parent or teacher data, or both, from one or both base-year data collections.

Also, as noted previously, the weight Y2COMW0 was designed for analyses of direct child assessment data from fall and spring kindergarten and spring first grade, in conjunction with parent or teacher data, or both, from spring first grade, and parent or teacher data, or both, from one or both base-year data collections.

Slide 17 of 29

It may not be possible to find the perfect weight, or one that is adjusted 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 be adjusted for 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 for 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.

If there is no weight that corresponds to all the rounds from which you are using data, the weight selected should, at the very least, be adjusted for nonresponse to components from the latest round of data included in the analysis. Such a weight will account for the subsampling of children who moved from their original sample schools across time and will be adjusted for the largest amount of nonresponse in the data. Finally, if data from the fall first-grade subsample are included in the analysis, it is critical that a weight with adjustments for round 3 be used. Such a weight will account for the fact that only about a third of the study schools and children were included in that round of data collection.

Slide 18 of 29

Now that we've reviewed the purpose of weights and discussed some of the weights that are available for analysis of the ECLS-K data, let's take a look at the actual weight variables in the ECLS-K data. Here is a very small sample of ECLS-K cases and their values on three weights created for analysis of the data collected in the first round of data collection, the fall of kindergarten.

Cases that are considered respondents to a given survey component have a valid, non-zero weight value for that component. The actual value of a weight indicates how many children in the target population that case is representing. For example, the first case, with CHILD ID 0002024C, represents about 227 children in the population when using weight C1CW0, while the second case, with CHILD ID 0005001C, only represents about 92 children in the population when using weight C1CW0. 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 have an analytic weight of zero, or sometimes a system missing code, for all weights that are adjusted for response to that component. For example, looking at the weight C1CW0, which is adjusted for nonresponse to the fall kindergarten child assessment only, we see that there are two cases in this subset that have a weight of zero. These are the cases that did not have any fall kindergarten child assessment data. Similarly, the final case in this subset, with ID 0005012C, has a value of zero for C1PW0. This means that this case did not have any parent interview data at round 1. Other responding cases with similar characteristics that do have a completed parent interview will have a larger analytic weight to compensate for missing data for cases that are nonrespondents.

Since C1CPTW0 is adjusted for nonresponse to the child assessment, parent interview, and teacher questionnaire, children who do not have valid data for all three components do not have a valid value for C1CPTW0. For example, case 0005008C has a weight of zero for weight C1CW0 because the child does not have valid child assessment data. As a result, the case also has a weight of zero for C1CPTW0, even though the case has data for the parent interview, as evidenced by the fact that it has a valid weight for C1PW0.

Slide 19 of 29

This table shows the means and sums of the same three weights created for the fall kindergarten round of the ECLS-K. 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 total. In the ECLS-K, the child-level weights sum to the general target population of children in kindergarten in the 1998-99 school year or first grade in the 1999-2000 school year, which was about 3.9 million children.

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, 19,173 cases have a valid value for the weight that is adjusted for nonresponse to the child assessment. These are cases that have valid child assessment data. The average weight for these cases, that is, the average number of children in the population each case represents, is about 202. Fewer cases have valid parent interview data than have child assessment data. Therefore, the mean value of C1PW0 is higher than that of C1CW0. On average, each case with parent interview data represents about 214 cases in the population. Similarly, even fewer cases have valid data on all three components, that is the child assessments, parent interview, AND teacher questionnaire. Therefore, the mean value of C1CPTW0 is higher than the mean value for the weights that are adjusted for nonresponse to just one of the components.

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 different weights that seem equally applicable for a given analysis, an analyst may want to select the weight that is valid for more cases in order to have a larger analytic sample.

Slide 20 of 29

In the common modules, two standard error calculation procedures were discussed: Replication Techniques and Taylor Series Linearization. This portion of the ECLS-K training modules describes important information regarding the inclusion of study design variables in your statistical package to compute appropriate 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 compute standard errors, you need to select replicate weight variables that are associated with your sampling weight variable. Each sampling weight variable in the ECLS-K has a set of replicate weight variables that are associated with it. For example, the replicate weights associated with the sampling weight variable C1CW0 are C1CW1 through C1CW90.

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

Slide 21 of 29

The second method for calculating standard errors is the Taylor Series Linearization Method. This method uses PSU and strata identifiers to calculate 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 weight variables, each sampling weight variable has PSU and strata variables associated with it. For example, the Taylor Series variables associated with the sampling weight variable BYCW0 are BYCWPSU and BYCWSTR.

Unlike the replicates weight variables, the strata and PSU variable names are not always identical to the sampling weight variable name. For example, the strata and PSU identifiers associated with the sampling weight variable C1CW0 are C1TCWSTR and C1TCWPSU, respectively.

Slide 22 of 29

Turning now to a discussion of the missing data, data can be missing across the ECLS-K for a variety of reasons. Standard missing data codes are used in the ECLS-K 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 spring kindergarten collection, a case will have a system missing code for all parent interview variables from that round of data collection.

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, the variable for that question will have a missing data value, while variables for other questions from the parent interview will have non-missing values.

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, a self-administered teacher questionnaire – subsequent items that are tied to the skipped variable are coded as not applicable and assigned a missing value code of -1.

A -7 missing data code indicates that a participant refused to answer an item. A missing data code of -8 indicates that a participant indicated that he or she did not know the answer to a question.

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 also be used if a respondent did not provide an answer to a question on a hard-copy questionnaire, and a composite variable indicating the sample child's race would be coded -9 if there was no information available from the parent or the school about the child's race.

The last code listed here, -2, is somewhat different than the other codes just discussed. While this code indicates data are missing, the data are not missing because a respondent failed to or was not asked to provide them. Rather, a -2 is used when data are suppressed on the public-use file for disclosure reasons. Researchers who find that several or key variables of interest for their analyses have been coded to -2 may want to consider applying for the restricted-use version of the data.

Slide 23 of 29

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 missing 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 three questions from the teacher questionnaire fielded in the fall of kindergarten:

Question 13 asks "Do any of the children in your class(es) speak a language other than English?"

Question 15 asks "Do you have any children with Limited English Proficiency (LEP) in your class(es)?"

Question 16 asks "How many children with Limited English Proficiency (LEP) do you have in each of your classes?"

Question 15 is only supposed to be answered when teachers said that, yes, children in their classes spoke a language other than English for question 13. Data for respondents who said no, they did not have children in their classes who spoke a language other

than English, are coded -1, not applicable, for question 15 because those respondents were instructed to skip to question 20.

Similarly, question 16 is only supposed to be answered when teachers said that, yes, they had children with limited English proficiency in their classes for question 15. Data for respondents who said no, they did not have children with limited English proficiency in their classes, are coded -1, not applicable, for question 16 because those respondents were instructed to skip to question 18.

There are different options available for recoding missing data values for questions 15 and 16. You can choose 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 responses to the previous question.

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

Slide 24 of 29

Here are two research questions someone may be interested in investigating with the ECLS-K data that generalize to kindergarten teachers in the 1998-99 school year.

For research question number 1, “What is the average number of limited English proficient, or LEP, students in AM kindergarten teachers’ classrooms?” the analysis should be conducted using all AM kindergarten teachers as the analytic group.

For research question number 2, “Of those AM kindergarten teachers who have limited English proficient, or LEP, students in their classrooms, how many LEP students, on average, do they have in their class?” the analysis should be conducted using only the subpopulation of AM kindergarten teachers who report that that they have limited English proficient students in their classrooms.

Slide 25 of 29

To address research question 1 about the average number of limited English proficient students in AM kindergarten teachers’ classrooms, cases coded not applicable, or -1, on question 16 can be recoded to a valid value based on the responses to questions 13 and 15.

Specifically, if the answer to either question 13 or 15 is “NO,” then the -1 codes on question 16 about the number of students can be recoded to a value of 0, because we know that the teachers did not have any children in their classrooms who were LEP or spoke a language other than English.

The remaining missing data values for question 16, that is -1, -7, -8, and -9, can be identified as missing data so that they are not treated as valid values.

Slide 26 of 29

To address research question 2 about the average number of limited English proficient students in AM kindergarten teachers' classrooms among AM kindergarten teachers who have LEP students in their classrooms, all missing values can be identified as missing data so that they are not treated as valid values. This way, the average will be calculated using only responses to question 16.

Slide 27 of 29

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

The first estimate, a mean of 4.03 LEP students, is the result for all AM kindergarten teachers, with the -1 missing data value recoded to a valid value of 0 for those teachers who reported that they did not have any LEP students or students who spoke a language other than English in their classes.

The second estimate, a mean of 6.22 LEP students, is the result for only those AM kindergarten teachers who did have LEP students in their classes, with all missing data values identified as missing data.

The third estimate, a weighted mean of -.36 LEP students, 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 erroneously decreasing the calculated mean.

Slide 28 of 29

This module provided information about the weights and variables associated with calculating estimates and standard errors for the ECLS-K data. When analyzing data from the ECLS-K, 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. Also, missing data must be handled appropriately to produce accurate estimates.

Slide 29 of 29

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 click the exit button to return to the landing page.