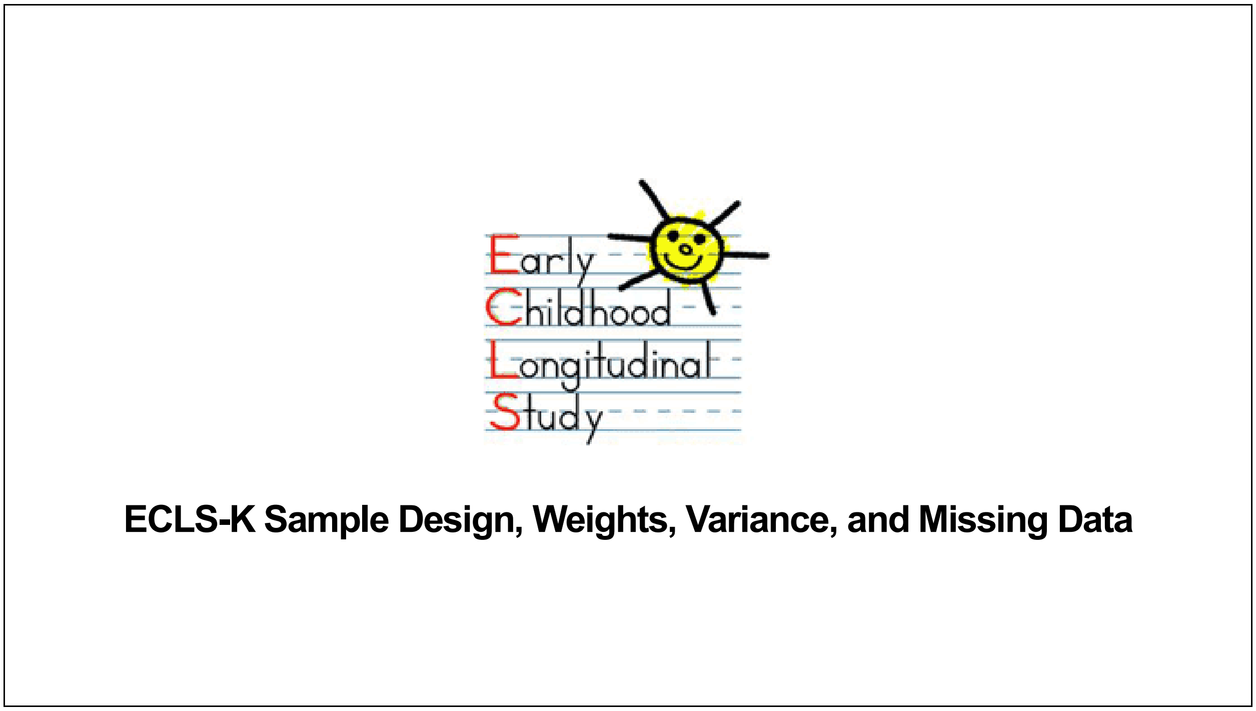
Slide 1 of 29



ECLS-K Sample Design, Weights, Variance, and Missing Data

Slide 2 of 29

Module Objectives

- Describe the ECLS-K weights that must be applied to assure data are representative of the target population
    - Summarize the ECLS-K sample design as it relates to study weights
- Describe the procedures for calculating appropriate standard errors
    - Identify the ECLS-K study design variables that must be used in statistical software to compute the correct standard errors
- Describe the missing data codes used in the ECLS-K data and how to handle missing data to ensure accurate data analysis

Slide 3 of 29

The ECLS-K Sample Design: Clustering

- The sample for the ECLS-K is not a simple random sample (SRS) of the target population
- Cases were clustered within primary sampling units (PSUs) to reduce field costs
- Multistage, stratified, clustered design
  - Stage 1: Country separated into PSUs, which were counties or groups of contiguous counties
    - Sampled within strata with probability proportional to size, where size was operationalized as the number of 5-year-olds in the PSU
  - Stage 2: Schools were sampled within sampled PSUs
  - Stage 3: Children were sampled within schools

Slide 4 of 29

The ECLS-K Sample Design: Clustering (Continued)

- 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, variation in the clustered sample is lower than it would be in an SRS

Slide 5 of 29

The ECLS-K Sample Design: Oversampling and Subsampling

- The ECLS-K includes an oversample of Asian/Pacific Islander children
    - Children identified as Asian/Pacific Islander were sampled at 2.5 times the rate of non-Asian/Pacific Islander children
- After the base year, subsampling was used 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
    - Children who moved from their base-year school were subsampled for follow-up in the first- through fifth-grade rounds

Slide 6 of 29

Purpose of Weights - Review

- Weights are used to make sample data representative of the target population
    - Data are representative of children enrolled in kindergarten in the United States during the 1998-99 school year
    - Data are representative of children in first grade in the United States during the 1999-2000 school year
    - Base-year data are representative of kindergarten teachers in the United States during the 1998-99 school year
    - Base-year data are representative of schools with kindergarten programs or educating kindergarten-aged children in ungraded settings in the United States during the 1998-99 school year
- Weights account for differential selection probabilities and differential patterns of response/nonresponse

Slide 7 of 29

Sampling Weights for the ECLS-K

- Multiple weights are provided for analysis using data from each round of data collection
- Ideally there would be a weight available that is adjusted for nonresponse to every component of every round of data collection
- 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
- Researchers must decide which weight is the best one to use given their research question

Slide 8 of 29

Selection of an ECLS-K Weight

Three main considerations when deciding which weight to use: the level, or unit, of analysis; the rounds of collection from which the data are taken; and the components that provide the data

- Identify the *level of analysis*: child, teacher, school
- Consider the *round(s)* of data in the analysis
  - Choose a weight that is adjusted for nonresponse at the rounds from which you are analyzing data
- Consider the *components providing the data* in the analysis
  - Choose a weight that is adjusted for nonresponse to the greatest number of components providing data for the analysis

Slide 9 of 29

Weight Names Indicate the Level of Analysis

- The first element in a weight variable name indicates the level of analysis
    - For the base year
        - Weights for school-level analyses begin with “S”
        - Weights for teacher-level analyses begin with “B”
        - Weights for child-level analyses with data from one round begin with “C”
        - Weights for child-level analyses with data from both kindergarten rounds begin with “BY”
- Weights for analyses of data from first through eighth grade all begin with “C” and are designed for child-level analyses only
    - The one exception to the naming convention is Y2COMW0

Slide 10 of 29

Weight Names Indicate the Round(s) of Data Included in the Analyses

The **second element** in a weight variable name indicates the round(s) of data

- Weights designed for analyses of data from only one round have a single round number: 1, 2, 3, 4, 5, 6, or 7
- Weights designed for longitudinal (i.e., across-round) analyses have two or more numbers, for example
    - “45” for rounds 4 and 5
    - “124” for rounds 1, 2, and 4
    - “1_4” for rounds 1, 2, 3, and 4
    - “1_7F” for rounds 1, 2, 4, 5, 6, and 7 (F = full sample, no round 3)
    - “1_5S” for rounds 1, 2, 3, 4, and 5 (S = subsample, round 3)

Slide 11 of 29

Weight Names Indicate the source of the Data

The third element in a weight variable name indicates the source(s) of data for which the weight is adjusted for nonresponse

- Weights designed to be used in analysis of child assessment data (alone or in conjunction with any combination of a limited set of child characteristics, e.g. age, sex, race/ethnicity) include only a “C” in the third element of the weight name
- Weights designed to be used in analysis of parent interview data (with or without child assessment data) include only a “P” in the third element of the weight name
- Weights designed for use in analyses of child AND parent AND teacher data include the letters “CPT” in the weight name
    - Fifth-grade weights: “CPT” is followed by an “R,” “M,” or “S” for reading, math, or science teacher
    - Eighth-grade weights: “CPT” is followed by an “E,” “M,” or “S” for English, math, or science teacher

Slide 12 of 29

Sources of Data That Affect Weight Selection

Sources that affect the choice of weight

- Child assessments
- Parent interview
- Teacher-/classroom-level teacher questionnaires

Slide 13 of 29

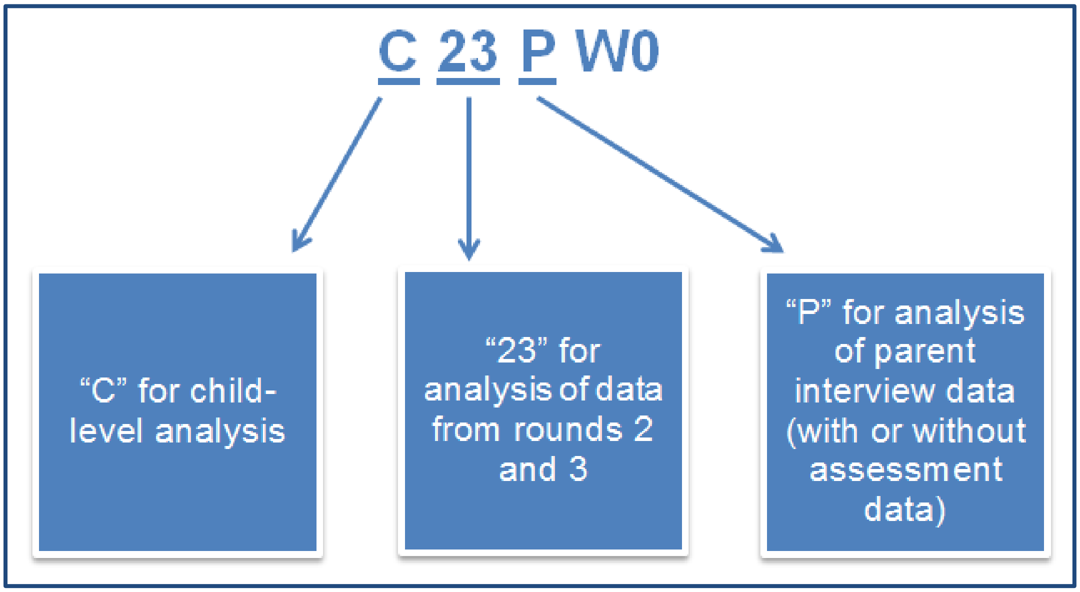
Sources of Data That Do Not Affect Weight Selection

Sources that do not affect the choice of weight

- School administrator questionnaire
- Facilities checklist
- Classroom teacher questionnaire (child-specific data)
- Special education teacher questionnaires
- Student record abstract data
- Head Start data
- Salary and benefits data

Slide 14 of 29

Weight Example



C 23 P W0

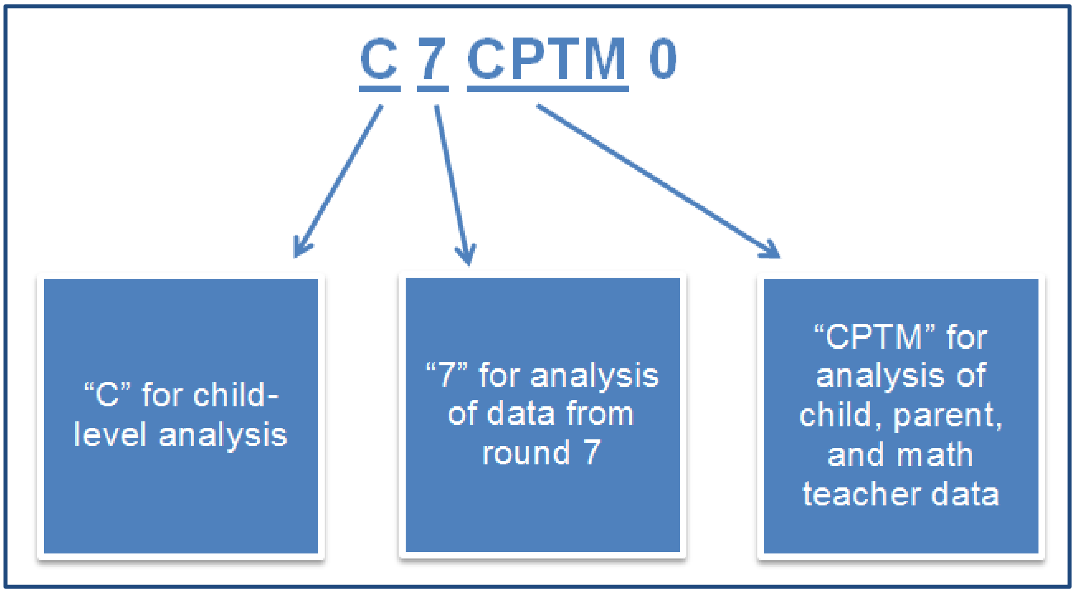
"C" for child-level analysis

"23" for analysis of data from rounds 2 and 3

"P" for analysis of parent interview data (with or without assessment data)

Slide 15 of 29

Weight Example (Continued)



C 7 CPTM 0

“C” for child-level analysis

“7” for analysis of data from round 7

“CPTM” for analysis of child, parent, and math teacher data

Slide 16 of 29

Two Exceptions to the Rule

- BYCOMW0: Direct child assessment data from Fall and Spring kindergarten, in conjunction with parent or teacher data, or both, from one or both base-year data collections
- Y2COMW0: 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

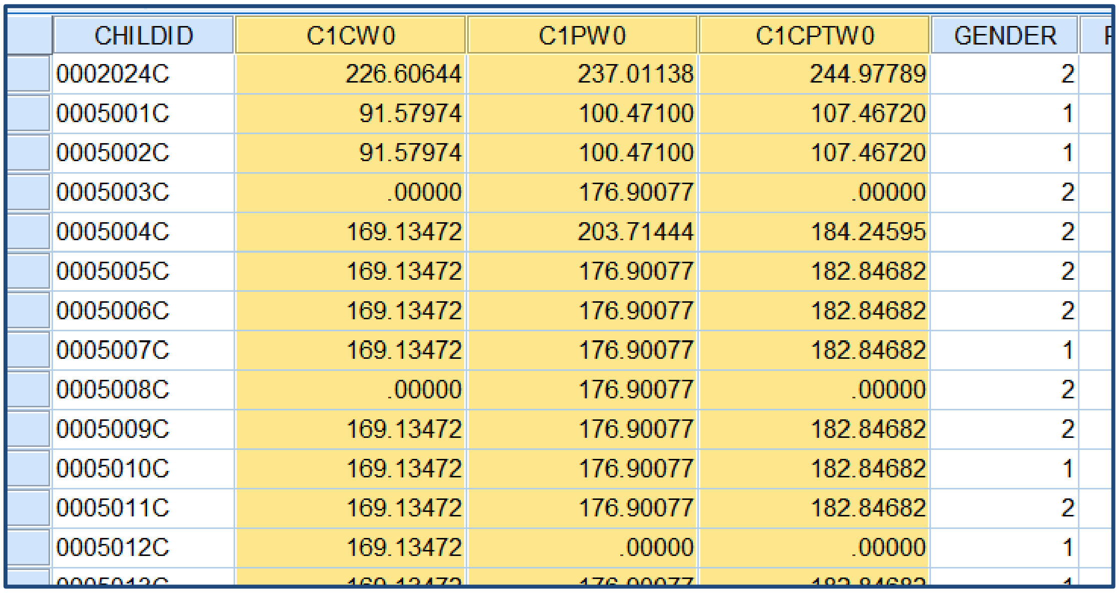
Selection of a Weight

Selection of a weight for analysis when the available weights do not correspond to your analysis

- 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
    - However, the weight will be adjusted for nonrandom nonresponse associated with each of the components that it covers, including the ones of interest in the analysis
- A researcher may choose a weight with nonresponse adjustments for fewer components than are included in the analysis to have a larger analytic sample. The degree to which key data are missing should be examined to get a sense of how likely the data are to be biased.
- At the very least, choose a weight that is adjusted for nonresponse to components from the latest round of data included in the analysis
- If data from the fall first-grade subsample are included in the analysis, choose a weight with adjustments for round 3

Slide 18 of 29

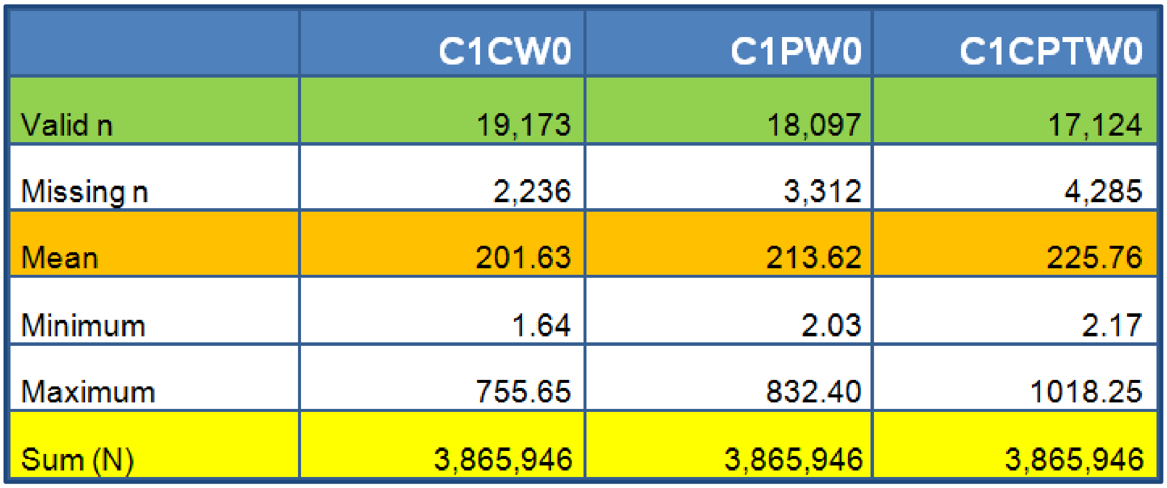
A Look at Weights in the ECLS-K Data



CHILDID	C1CW0	C1PW0	C1CPTW0	GENDER	F
0002024C	226.60644	237.01138	244.97789	2	
0005001C	91.57974	100.47100	107.46720	1	
0005002C	91.57974	100.47100	107.46720	1	
0005003C	.00000	176.90077	.00000	2	
0005004C	169.13472	203.71444	184.24595	2	
0005005C	169.13472	176.90077	182.84682	2	
0005006C	169.13472	176.90077	182.84682	2	
0005007C	169.13472	176.90077	182.84682	1	
0005008C	.00000	176.90077	.00000	2	
0005009C	169.13472	176.90077	182.84682	2	
0005010C	169.13472	176.90077	182.84682	1	
0005011C	169.13472	176.90077	182.84682	2	
0005012C	169.13472	.00000	.00000	1	
0005013C	169.13472	176.90077	182.84682	1	

Slide 19 of 29

Weighted Data Sum to the Population



	C1CW0	C1PW0	C1CPTW0
Valid n	19,173	18,097	17,124
Missing n	2,236	3,312	4,285
Mean	201.63	213.62	225.76
Minimum	1.64	2.03	2.17
Maximum	755.65	832.40	1018.25
Sum (N)	3,865,946	3,865,946	3,865,946

Slide 20 of 29

Standard Error Calculation in ECLS-K: Replication Techniques

- These methods calculate appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples (replicates)
- Select replicate weights that are associated with your sampling weight (e.g., C1CW1 to C1CW90 for weight C1CW0)
- The paired-jackknife, or JK2, method is the appropriate replication technique to use when computing standard errors for ECLS-K data

Slide 21 of 29

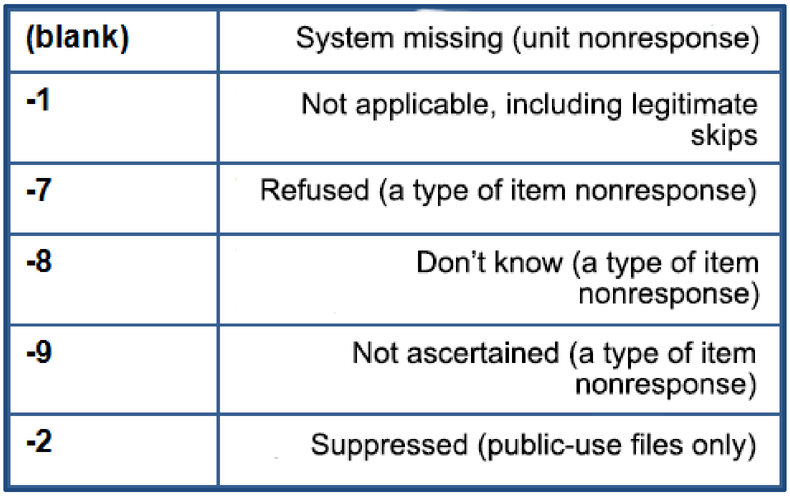
Standard Error Calculation in ECLS-K: Taylor-Series Linearization

- This method uses primary sampling unit (PSU) and strata identifiers to calculate appropriate standard errors
- Each sampling weight variable has PSU and strata variables associated with it
  - The Taylor Series variables associated with the sampling weight variable BYCW0 are BYCWPSU and BYCWSTR
- Unlike the replicate weight variables, the strata and PSU variable names are not always identical to the sampling weight variable name
  - The strata and PSU identifiers associated with the sampling weight variable C1CW0 are C1TCWSTR and C1TCWPSU

Slide 22 of 29

Missing Data Values Used in the ECLS-K Data

- Missing data are retained for most variables
- Standard ECLS-K missing data codes



(blank)	System missing (unit nonresponse)
-1	Not applicable, including legitimate skips
-7	Refused (a type of item nonresponse)
-8	Don't know (a type of item nonresponse)
-9	Not ascertained (a type of item nonresponse)
-2	Suppressed (public-use files only)

Slide 23 of 29

Example: Recoding Missing Data

Fall Kindergarten Teacher Questionnaire questions

Q13: Do any of the children in your class(es) speak a language other than English? IF NO, THEN SKIP TO Q20

Q15: Do you have any children with Limited English Proficiency (LEP) in your class(es)? IF NO, THEN SKIP TO Q18

Q16: How many children with Limited English Proficiency (LEP) do you have in each of your classes?

Slide 24 of 29

Example: Recoding Missing Data (Continued)

Question 1: What is the average number of limited English proficient (LEP) students in AM kindergarten teachers' classrooms?

Question 2: Of those AM kindergarten teachers who have limited English proficient (LEP) students in their classrooms, how many LEP students, on average, do they have in their class?

Slide 25 of 29

Example: Recoding Missing Data (Continued)

Q1: What is the average number of limited English proficient (LEP) students in AM kindergarten teachers' classrooms?

Recoding steps

- If the answer to Q13 (any children speak a language other than English?) or Q15 (any children with limited English proficiency in class?) is "NO," then recode values of -1 on Q16 (number of LEP students in class) to 0
- Then recode all remaining missing value codes (-1, -7, -8, -9) for Q16 to missing

Slide 26 of 29

Example: Recoding Missing Data (Continued)

Q2: Of those AM kindergarten teachers who have limited English proficient (LEP) students in their classrooms, how many LEP students, on average, do they have in their class?

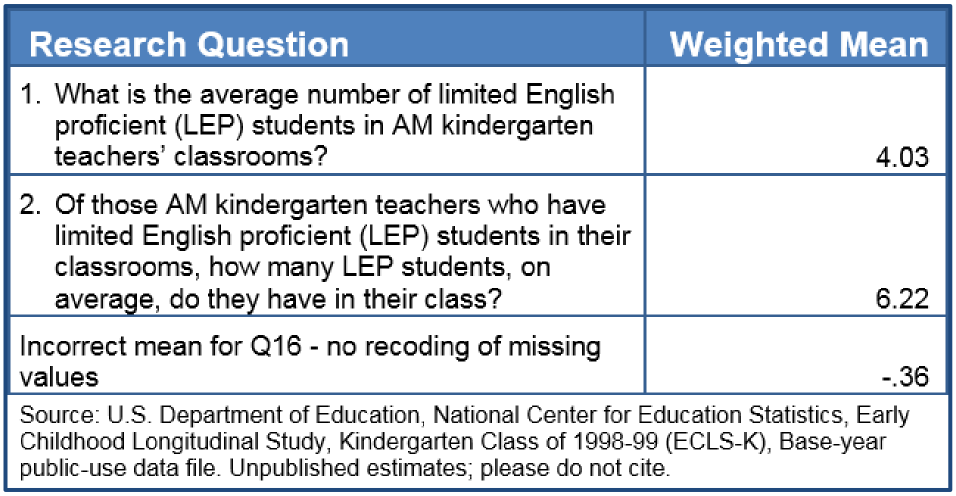
Recoding steps

- Only those cases with a response of 'Yes' to Q15 (any children with limited English proficiency in class?) should be included in the analysis
- All missing value codes (-1, -7, -8, -9) for Q16 should be recoded to missing

Slide 27 of 29

Example: Recoding Missing Data (Continued)

Weighted Results for Research Questions



Research Question	Weighted Mean
1. What is the average number of limited English proficient (LEP) students in AM kindergarten teachers' classrooms?	4.03
2. Of those AM kindergarten teachers who have limited English proficient (LEP) students in their classrooms, how many LEP students, on average, do they have in their class?	6.22
Incorrect mean for Q16 - no recoding of missing values	-.36

Source: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Base-year public-use data file. Unpublished estimates; please do not cite.

Slide 28 of 29

Module Summary

- Described the ECLS-K weights that must be applied to assure data are representative of the target population
  - Summarized the ECLS-K sample design as it relates to study weights
- Described the procedures for calculating appropriate standard errors
  - Identified the ECLS-K study design variables that must be used in statistical software to compute the correct standard errors
- Described the missing data codes used in the ECLS-K data and how to handle missing data to ensure accurate data analysis

Slide 29 of 29

Module Resources

- Applying weights
- Clustering
- Calculating appropriate standard errors
- Replication techniques
- Taylor series linearization