



## GUIDANCE ON HOW TO ANALYZE DATA FROM A SCHOOL-BASED ORAL HEALTH SURVEY JULY 2013

***Due to the technical nature of this topic, this information will be most helpful to data analysts, epidemiologists and statisticians.***

### **Has your state, territory or local health agency conducted a school-based oral health survey?**

If yes, then you probably have questions about how to appropriately weight and analyze the data to best represent the target population of your survey. The purpose of this document is to provide a basic framework for how to appropriately analyze data from a statewide school-based oral health survey that employed a complex sampling design. Because no one method is appropriate for all states/territories, we encourage you to read this document, and then contact ASTDD for additional guidance on analyzing data for your state/territory. Although this document is geared towards states and territories, the techniques are appropriate for other jurisdictions such as counties.

*Because no one method is appropriate for all states/territories, we encourage you to read this document, and then contact ASTDD for additional guidance on analyzing data for your state/territory.*

This topic is important because most oral health surveys employ a complex sampling design that may include stratification, unequal selection probabilities and clustering. To obtain valid point estimates, standard errors, and confidence intervals, analysis must account for the sampling design. Simply doing a weighted analysis using statistical procedures like SAS Proc Freq is not appropriate because the variance estimation in such programs use formulas appropriate for simple random sampling rather than complex sampling. These formulas do not account for stratification or clustering and may result in biased point estimates of population parameters (in an unweighted analysis) and/or underestimation of standard errors and confidence intervals for point estimates.

This document is limited to a discussion of data weighting and analysis. For additional information on how to conduct and use data from a school-based oral health survey, please refer to the Basic Screening Survey (BSS) tools developed by the Association of State and Territorial Dental Directors (ASTDD). These tools are available at the following website: [www.astdd.org/basic-screening-survey-tool/](http://www.astdd.org/basic-screening-survey-tool/).

### **Do you want to submit the data to the National Oral Health Surveillance System (NOHSS)?**

NOHSS ([www.cdc.gov/nohss](http://www.cdc.gov/nohss)) is a collaborative effort between CDC's Division of Oral Health and ASTDD. NOHSS is designed to monitor the burden of oral disease, use of the oral health care delivery system, and the status of community water fluoridation on both a national and state level. NOHSS tracks oral health surveillance indicators based on data sources and surveillance capacity available to most states. The Council of State and Territorial Epidemiologists (CSTE) and the Chronic Disease Directors (CDD) were instrumental in developing the framework for chronic disease surveillance indicators, including the oral health indicators in NOHSS. If you follow the guidance provided in this document and ASTDD's sampling guidance document, your oral health data will meet the specifications for inclusion in NOHSS.

*If you follow the guidance provided in this document and ASTDD's sampling guidance document, your oral health data will meet the specifications for inclusion in NOHSS.*

Only oral health survey data that meet the following specifications are included in the NOHSS data system:

- The data are from a **statewide probability sample** of elementary schools.

- If a complex sampling scheme is used, the data must be weighted for the sampling scheme.
- The NOHSS Advisory Committee strongly suggests that, at minimum, 3rd grade children be screened. Grades K-2 as well as Head Start may also be screened and are included in the NOHSS website.
- While it is not mandatory that data be adjusted for non-response, the NOHSS Advisory Committee strongly encourages states to adjust for non-response whenever possible. Both unadjusted and adjusted data may be submitted. If you follow the guidance in this document your estimates will be adjusted for both the sampling scheme and non-response.

**NOTE:** In some cases a state may be unable to follow this guidance. For example, because of small school size and confidentiality issues, an IRB may require that school identifiers not be included in the dataset. If you encounter such issues, please contact ASTDD for additional guidance.

**Once our oral health data has been collected, what steps do I need to take to prepare the data for analysis?**

There are several steps you should take to prepare for the analysis phase of your survey. Carefully review each step and decide if it is appropriate for your situation.

Step 1: Enter the data into an electronic file that can be exported to an appropriate statistical package. Good options for data entry systems include Epi Info ([www.cdc.gov/epiinfo](http://www.cdc.gov/epiinfo)) and Microsoft Access. To minimize or eliminate data entry errors it is important to have a “very smart” data entry system that can make a variety of checks on the data while it is being entered. Typically a useful data entry system checks each field for valid values, inconsistencies in data across fields, skip patterns, etc. A good data entry system, just like a good form, should be designed to be self-explanatory and easy to use. Examples of data entry systems using Epi Info and Access are available from ASTDD.

**Steps to Take to Prepare for Analysis**

- Enter the data
- Clean the data
- Determine number screened at each school
- Determine number in sampling interval
- Calculate weight factor

Step 2: Clean the data file. If you used a smart data entry system, there should be very few data entry errors. Make sure that each record includes the appropriate school code. A school code is necessary for calculating the weight factor that will be used in the analysis. For additional information on data cleaning and preemptive data cleaning techniques, refer to the following brief: [www2.sas.com/proceedings/sugi26/p015-26.pdf](http://www2.sas.com/proceedings/sugi26/p015-26.pdf). Once you have selected the data entry system and statistical software package to use, it may be helpful to read briefs or reference books specific to that system or package.

Step 3: Determine how many children were screened at each school. This can be accomplished by generating a frequency distribution for school codes. The number of children screened at each school will be used as the denominator in the weight factor calculation.

Step 4: Go back to the file you used to select the sample and determine how many children were enrolled in each sampling interval. Link the sampling interval information to the participating school codes. The number of children in the sampling interval will be partially dependent on the type of sampling strategy you used.

- If you used a probability proportional to size (PPS) sampling strategy, the number of children in the sampling interval will be the same for each sampling interval in a given strata. Refer to Example 1 and Table 1 (page 10).
- If you used a non-PPS sampling strategy, the number of children in the sampling interval will generally be different for each sampling interval. Refer to Example 2 and Table 2 (page 12).
- Refer to the school survey sampling guidance developed by ASTDD for additional information. The sampling guidance is available at the following site: [www.astdd.org/basic-screening-survey-tool/](http://www.astdd.org/basic-screening-survey-tool/)

Step 5: Calculate the weight factor using the following formula. Each child in a particular school and grade will have the same weight factor.

- Weight = (# of children in sampling interval) / (# of children screened in sampling interval)
  - This formula reflects the reduction of the overall probability calculation of:  
(# enrolled in school/# in sampling interval)\* (# children invited to participate /# enrolled in school) \* (# screened/# children invited to participate) = # of children screened in sampling interval/# of children in sampling interval
  - Note that number enrolled in school cancels out in the first and second terms and number of children invited to participate cancels out in the second and third terms
  - The analysis weight is the inverse of the reduced probability term: (# of children in sampling interval) / (# of children screened in sampling interval)
- **NOTE:** The number of children in the sampling interval is based on the sampling frame used for selecting the sample which generally will be from the school year prior to the data collection year. These numbers would be expected to be very close to current numbers.

### What statistical software package and program code should I use?

Analysis of data from surveys that employ a complex sampling design, such as a school-based oral health survey, must account for the sampling design. Several statistical software packages are either (1) designed specifically to analyze complex sample survey data or (2) have special procedures or modules to correctly analyze complex sample survey data including SUDAAN, SAS, STATA, SPSS, Epi Info and R. All of these packages are appropriate for the analysis of school-based oral health survey data; your decision for which package to use will probably be based on availability, familiarity or cost. Both Epi Info and R are available at no cost to the user.

To help you with the analysis process, we have created sample program code for each of the packages listed and have compared results from each package based on a sample data set from a recent state oral health survey of kindergarten and 3<sup>rd</sup> grade children. Information about each statistical software package, except R, was excerpted from *Software for Analysis of YRBS Data* (CDC 2012).

#### Definition of variables used in the sample program code:

- Grade – K=kindergarten, 3=third grade
- Race – 1=white, 2=black
- Cluster – a unique number for each school, primary sampling unit (PSU)
- Strata – geographic region of the state, stratification variable used in selecting sample
- Weight – analysis weight factor (# children in sampling interval / # children screened in interval)
- Untreated – does the child have untreated decay (0=No, 1=Yes)
- Treated – does the child have treated decay (0=No, 1=Yes)
- Experience – does the child have treated and/or untreated decay (0=No, 1=Yes). You will need to create the variable “Experience” from “Treated” and “Untreated”
  - If Untreated is missing and Treated is missing then Experience should be coded missing
  - If Untreated=0 and Treated=0 then Experience=0
  - If Untreated=1 or Treated=1 then Experience=1
- Sealants – does the child have dental sealants (0=No, 1=Yes)

#### General items that deserve caution:

- Missing data: Each software package uses its own special coding for missing data, for example “.” in SAS and “N/A” in R. These special codes for missing data translate to numeric values in calculations, sometimes very small and sometimes very large. Take care in recoding or creating new variables to be sure that missing data are categorized as you intended.
- Subpopulation analyses, “By” statement dropping observations from data set: With complex sample data, to get estimates for a subpopulation, such as male or Hispanic children, the statistical software

requires information about the sampling design, the strata, and primary sampling units (PSUs). Dropping observations from the data set for children who are not in the subpopulation can result in loss of information on some strata and PSUs, resulting in estimates that do not account for the correct number of strata and PSUs. Using a “By” statement to get estimates for males and females in SAS, for example, is equivalent to doing the analysis once dropping all of the females, and then again dropping all of the males. ***The “By” statement is NOT the recommended way to get estimates for subpopulations for many of the software packages.*** Sudaan version 11 is an exception. Check the documentation for your preferred statistical software to be sure you are using the correct syntax for proper subpopulation analysis.

**SAS sample code:** SAS versions 8 and higher include special sample survey procedures that are appropriate for analyzing complex survey data. These sample survey procedures use SAS syntax that will be familiar to those who already use SAS. There are three sample design statements in SAS: CLUSTER, where the name of the primary sampling unit (PSU) is placed; STRATA, where the name of the stratification variable is placed; and WEIGHT, where the name of the analysis weight variable is placed. Variables may be numeric or character. The input data file does not need to be sorted by stratum and/or PSU variables before analysis.

Univariate analysis (data not presented):

```
PROC SURVEYFREQ ;  
STRATA strata ;  
CLUSTER cluster ;  
WEIGHT weight ;  
TABLES untreated experience sealants / cl ;  
RUN ;
```

Table 3:

```
PROC SURVEYFREQ ;  
STRATA strata ;  
CLUSTER cluster ;  
WEIGHT weight ;  
TABLES grade*untreated / row cl ;  
RUN ;
```

```
PROC SURVEYFREQ ;  
STRATA strata ;  
CLUSTER cluster ;  
WEIGHT weight ;  
TABLES grade*experience / row cl ;  
RUN ;
```

```
PROC SURVEYFREQ ;  
STRATA strata ;  
CLUSTER cluster ;  
WEIGHT weight ;  
TABLES grade*sealants / row cl ;  
RUN ;
```

Table 4:

```
PROC SURVEYFREQ ;  
STRATA strata ;
```

```
CLUSTER cluster ;
WEIGHT weight ;
TABLES grade*race*sealants / row cl ;
RUN ;
```

**Epi Info sample code:** Epi Info includes a module for complex sample survey analysis. The analytic capabilities of Epi Info are limited and are oriented towards public health field work applications. Sample design information is entered into the appropriate box (Weight, PSU, Stratify by) in the dialog box that appears once an analysis (Complex Sample Frequencies, Complex Sample Tables, Complex Sample Means) has been selected. You can also use the syntax codes below. Variables may be numeric or character. The input data file does not need to be sorted by stratum and/or PSU variables before analysis. **IMPORTANT NOTE:** As of July 2013, Epi Info does not have the ability to appropriately generate subpopulation analyses. Using the “Select” statement will drop observations and may impact information about the sampling design, strata, and primary sampling units (PSUs). Dropping observations from the data set for children who are not in the subpopulation may result in loss of information on some strata and PSUs, resulting in estimates that do not account for the correct number of strata and PSUs.

Univariate analysis (data not presented):

```
FREQ untreated experience sealants STRATAVAR=strata WEIGHTVAR=weight PSUVAR=cluster
```

Table 3:

```
TABLES grade untreated STRATAVAR=strata WEIGHTVAR=weight PSUVAR=cluster
```

```
TABLES grade experience STRATAVAR=strata WEIGHTVAR=weight PSUVAR=cluster
```

```
TABLES grade sealants STRATAVAR=strata WEIGHTVAR=weight PSUVAR=cluster
```

Table 4:

```
SELECT grade="3"
```

```
TABLES race sealants STRATAVAR=strata WEIGHTVAR=weight PSUVAR=cluster
```

**R sample code:** R is an open source, freely available software. Users develop R “packages” for specific purposes. Analysis of complex sample survey data requires the package “survey” developed by Thomas Lumley at University of Washington. Details of R can be found at the R-Project website <http://www.r-project.org/> and further details of the survey package can be found at Thomas Lumley’s website <http://faculty.washington.edu/tlumley/survey/>. Variables may be numeric or character. The input data file does not need to be sorted by stratum and/or PSU variables before analysis.

```
#Describe the sample design to R
BSS <- svydesign(id=~Cluster, strat=~Strata, weight=~Weight, data=dat2)
```

Univariate analysis (data not presented):

```
uniana <- svymean(~Untreated+Experience+Sealants, BSS, na.rm = TRUE)
```

```
#Calculating Confidence Intervals
```

```
t1<- ftable(uniana)
```

```
UniTab = data.frame(Mean = t1[,1], CintLow = t1[,1]-1.96*t1[,2], CintHigh = t1[,1]+1.96*t1[,2])
```

```
#Rounding Table Values
```

```
UniTab[,1:3] <- round(100*UniTab[,1:3],1)
```

Table 3:

```
#Estimate proportions and standard errors within groups
vun<-svyby(~Untreated, ~Grade, svymean, design=BSS, keep.names=FALSE, na.rm = TRUE)
vexp<-svyby(~Experience, ~Grade, svymean, design=BSS, keep.names=FALSE, na.rm = TRUE)
vseal<-svyby(~Sealants, ~Grade, svymean, design=BSS, keep.names=FALSE, na.rm = TRUE)

#Calculating the Confidence Intervals
v1 = vexp[,1:2]
v1$Untreated = vun$Untreated
v1$cil.Untreated= v1$Untreated - 1.96*vun$se.Untreated
v1$clu.Untreated= v1$Untreated + 1.96*vun$se.Untreated
v1$Experience = vexp$Experience
v1$cil.Experience= v1$Experience - 1.96*vexp$se.Experience
v1$clu.Experience= v1$Experience + 1.96*vexp$se.Experience
v1$Sealants = vseal$Sealants
v1$cil.Sealants= v1$Sealants - 1.96*vseal$se.Sealants
v1$clu.Sealants= v1$Sealants + 1.96*vseal$se.Sealants

#Rounding Table Values
v1[,2:10] <- round(100*v1[,2:10],1)
```

Table 4:

```
v3Seal<-svyby(~Sealants, ~Grade+Race, svymean, design=BSS, keep.names=FALSE, na.rm = TRUE)

# Calculating Confidence Interval and Rounding
v3Seal$cintLow=v3Seal$Sealants - 1.96*v3Seal$se.Sealants
v3Seal$cintHigh=v3Seal$Sealants + 1.96*v3Seal$se.Sealants

v3Seal[,3:6] <- round(100*v3Seal[,3:6],1)
```

**SUDAAN sample code:** SUDAAN is specifically designed to analyze complex sample survey data. The user describes the sample survey design in three statements: (1) by specifying an option for the DESIGN keyword on the PROC statement, (2) by specifying the stratification and clustering (PSU) variables on the NEST design statement, and (3) by specifying the analysis weight variable on the WEIGHT design statement. **All variables must be numeric.** For this example, grade was changed for K and 3 to 0 and 3. Data should be sorted by the variables that appear on the NEST statement before analysis, otherwise procedure syntax must contain the NOTSORTED option.

**Univariate analysis (data not presented):**

```
proc descript data=bss design=wr conf_lim=95;
  nest strata cluster;
  weight weight;
  var  EXPERIENCE UNTREATED SEALANTS;
  catlevel 1 1 1;
run;
```

Table 3:

```
proc descript data=bss design=wr conf_lim=95;
  nest strata cluster;
  weight weight;
```

```

class GRADE;
var  EXPERIENCE UNTREATED SEALANTS;
catlevel  1      1      1;
tables GRADE;
run;

```

Table 4:

```

proc descript data=bss design=wr conf_lim=95;
nest strata cluster;
weight weight;
class RACE;
var  SEALANTS;
catlevel  1;
tables RACE;
subpopn GRADE=3;
run;

```

**STATA sample code:** STATA versions 7 or higher offers the capability to perform many statistical procedures on complex sample survey data. When performing menu-driven analyses, sample design information is entered into boxes on the MAIN (PSU and stratification variables) and WEIGHT (analysis weight variable) tabs of the dialogue box that appears after “declare survey design for data set” is chosen from the Survey Data Analysis menu. If syntax is written the information is included on the SVYSET statement. The survey design descriptors only need to be entered once at the beginning of the analysis session. Although variables in STATA data sets can be numeric or character, **all variables used in an analysis must be numeric**. The input data file does not need to be sorted by stratum and/or PSU variables before analysis.

Univariate analysis (data not presented):

```

Svyset Cluster [pweight = Weight], strata(Strata)
Svy linearized : proportion Sealants Untreated Experience

```

Table 3:

```

Svy linearized : proportion Sealants Untreated Experience, over(Grade)

```

Table 4:

```

Svy linearized : proportion Sealants Untreated Experience, over(Race Grade)

```

**SPSS sample code:** SPSS has an add-on module, SPSS Complex Samples, which includes sample selection and analysis of complex sample survey data. When performing menu-driven analysis, the sample design information is entered into a dialogue box when preparing for analysis and is saved as a sampling plan for the data set. Once the sampling plan has been created, it will be opened along with the data set at the beginning of an SPSS session. You can also use the following syntax code. Variables may be numeric or character. The input data file does not need to be sorted by stratum and/or PSU variables before analysis.

```

*Analysis Preparation Wizard.
CSPLAN ANALYSIS
/PLAN FILE='...\astdd test sample.csaplan'
/PLANVARS ANALYSISWEIGHT=Weight
/SRSESTIMATOR TYPE=WOR
/PRINT PLAN
/DESIGN STRATA=Strata CLUSTER=Cluster

```

```
/ESTIMATOR TYPE=WR.
```

Univariate analysis (data not presented):

```
* Complex Samples Frequencies.  
CSTABULATE  
/PLAN FILE='...\astdd test sample.csaplan'  
/TABLES VARIABLES=Untreated experience Sealants  
/CELLS POPSIZE TABLEPCT  
/STATISTICS CIN(95)  
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Table 3:

```
* Complex Samples Frequencies.  
CSTABULATE  
/PLAN FILE='...\astdd test sample.csaplan'  
/TABLES VARIABLES=Untreated  
/SUBPOP TABLE=Grade DISPLAY=LAYERED  
/CELLS POPSIZE TABLEPCT  
/STATISTICS SE CIN(95)  
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

```
* Complex Samples Frequencies.  
CSTABULATE  
/PLAN FILE='...\astdd test sample.csaplan'  
/TABLES VARIABLES=experience  
/SUBPOP TABLE=Grade DISPLAY=LAYERED  
/CELLS POPSIZE TABLEPCT  
/STATISTICS SE CIN(95)  
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

```
* Complex Samples Frequencies.  
CSTABULATE  
/PLAN FILE='...\astdd test sample.csaplan'  
/TABLES VARIABLES=Sealants  
/SUBPOP TABLE=Grade DISPLAY=LAYERED  
/CELLS POPSIZE TABLEPCT  
/STATISTICS SE CIN(95)  
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Table 4:

```
* Complex Samples Crosstabs.  
CSTABULATE  
/PLAN FILE='...\astdd test sample.csaplan'  
/TABLES VARIABLES=Race BY Sealants  
/SUBPOP TABLE=Grade DISPLAY=LAYERED  
/CELLS POPSIZE ROWPCT  
/STATISTICS SE CIN(95)  
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```



**Comparison of results:** Tables 3 and 4 compare the results from each of the aforementioned statistical software packages from a recent state oral health survey of kindergarten and 3<sup>rd</sup> grade children. Table 3 presents the prevalence of decay experience, untreated decay, and dental sealants by grade while the prevalence of dental sealants among 3<sup>rd</sup> grade children by race is presented in Table 4. ***Each of the six statistical software packages described in this document can be used to appropriately analyze data from a school-based survey, although Epi Info may not appropriately account for the correct number of strata and PSUs in subpopulation analyses.***

### Are there other things that I should consider or be aware of?

Yes, there are a variety of other issues that may impact your analysis or how you report your survey reports. Following is a short list that you should review. If you have additional questions or concerns please contact ASTDD.

- *Finite population correction:* If more than 10% of children from within any given strata are selected, you may want to consider using a finite population correction factor, which reduces variance yielding smaller standard errors and confidence intervals. For additional information on finite population correction refer to *Introduction to Survey Sampling*.
- *No data for a given sampling interval:* If a school refuses to participate, we encourage you to select a replacement school from the same sampling interval. Unfortunately, circumstances may result in an inability to screen the original or a replacement school in a given sampling interval. If this happens you should clearly report that you were not able to screen children in a sampling interval along with what that interval represented. For example, if you selected 70 schools but you only have data for 69, report that you are missing data from one sampling interval that represents children from region 3 attending schools where 30-40% of the children are eligible for the free or reduced price school lunch (FRL) program.
- *Reporting response rates:* For each school in your survey, you should collect the number of children enrolled in the grade of interest on the day of the screening (or the number invited to participate if you did not invite all children in a given grade). Your response rate for the survey will be the number screened divided by the number enrolled or invited to participate.
- *Stratifying results by school FRL status:* Many states use school FRL status as a surrogate measure of socioeconomic status. We recommend using the current year FRL status of the school if stratifying the results by FRL status; this information can be obtained from the school on the day of the screening.
- *Limitations of survey:* When preparing your survey report, it is important to clearly state any limitations of the survey including representativeness and response rate.
- *Confidence intervals:* Confidence intervals are important because they provide context for understanding the precision or exactness of a point estimate. The wider the confidence interval, the less exact the point value estimate becomes. Take, for example, a point estimate of 40% for the prevalence of dental caries experience. If the confidence interval of this point estimate is 35%-45%, then we can have greater certainty that the true prevalence is near 40% than if the confidence interval was 10-70%. For your data to be included in NOHSS, confidence intervals must be included (unless you screened all children in your target population).

### Where can I get additional help?

ASTDD can help you with the survey analysis process. Please contact us if you have any questions.

Association of State & Territorial Dental Directors  
Kathy Phipps, Data and Surveillance Coordinator  
Phone: 805-776-3393, Email: [kathyhipps1234@gmail.com](mailto:kathyhipps1234@gmail.com)

### Acknowledgements

Supported by Cooperative Agreement 5U58DP001695-05 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of

CDC. ASTDD would like to thank Kathy Phipps, Michael Manz, Laurie Barker, Eugenio Beltran, Mei Lin, Liang Wei and Srdjan Lesaja for their assistance in developing and reviewing this guidance.

#### References and additional resources

- Software for Analysis of YRBS Data, Division of Adolescent and School Health, Centers for Disease Control and Prevention, [www.cdc.gov/healthyyouth/yrbs/pdf/YRBS\\_analysis\\_software.pdf](http://www.cdc.gov/healthyyouth/yrbs/pdf/YRBS_analysis_software.pdf)
- Brogan D, Sampling error estimation for survey data. In: United Nations, Department of Economic and Social Affairs, Household Sample Surveys in Developing and Transition Countries, 2005. Available at: [http://unstats.un.org/unsd/hhsurveys/pdf/Household\\_surveys.pdf](http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf)
- Kalton G. Introduction to survey sampling. Quantitative Applications in the Social Sciences. 1988. Sage Publications, Beverly Hills. Series/Number 07-035. Isbn 0-8039-2126-8.
- Kish L. Statistical Design for Research. 2004. John Wiley & Sons, Hoboken.

**Example #1**  
**Weight factor calculation for a survey that used systematic PPS sampling  
with implicit stratification by region, urban/rural status and F/R lunch**

This example shows the steps from sample selection to weight factor calculation when a probability proportional to size (PPS) sampling strategy is used. Based on available resources, the decision was made to include 70 schools in the “Utopia” oral health survey of 3<sup>rd</sup> grade children. The following sampling steps were employed:

- The sampling frame list was sorted by region then by urban/rural status within each region
- Schools were then sorted by percent of children participating in the FRL program within urban/rural school categories.

**Calculations used for selecting the systematic PPS sample:**

- Sampling interval for sampling = (total 3<sup>rd</sup> grade enrollment) / (# of schools to be screened)
  - $53,320 / 70 = 761.7$
- Random start = random number between 0 and interval (761.7) = **148.0**
  - This is the first school selection number
  - There are a variety of methods for selecting a random number including, but not limited to, Excel and [www.random.org](http://www.random.org)
- Select the school with the 148<sup>th</sup> child. Add the sampling interval (761.7) to 148 to get the next school (909.7). Continue adding the sampling interval repeatedly until all 70 school selections are made.

148.0, 909.7, 1671.4, 2433.1, 3194.8, 3956.5, 4718.2, ...
---

- These numbers are matched to the cumulative enrollment numbers in the sampling list. The schools with enrollment intervals containing the sample selection numbers are selected into the sample. The sampling frame list and the selected schools are shown in Table 1.

**Weight factor calculation (Table 1):**

- Weight = (number of children in sampling interval) / (number of children screened)
- When PPS sampling is used the number of children in the sampling interval will always be the sampling interval used when selecting the sample, in this case, 761.7.

**Table 1: Systematic PPS sampling with implicit stratification by region, urban/rural and F/R lunch**

Region	Urban/ Rural	School Name	Free/Reduced Lunch Percent	3rd Grade Enrollment	Cumulative Enrollment	Selected School	# of Children in Sampling Interval (A)	Number Screened (B)	Weight (A)/(B)
1	Rural	KEOWEE	37.1%	38	38				
1	Rural	WALHALLA	39.8%	92	130				
1	Rural	RAVENEL	52.3%	91	221	148.0	761.7	52	14.648
1	Rural	LAKEVIEW	52.4%	94	315				
1	Rural	NINETY SIX	52.7%	130	445				
1	Rural	NORTHSIDE	52.8%	95	540				
1	Rural	MCCORMICK	53.9%	56	596				
1	Rural	FAIR-OAK	55.7%	117	713				
1	Rural	HICKORY TAVERN	56.8%	61	774				
1	Rural	HOLLYWOOD	57.2%	66	840				
1	Rural	CHEROKEE TRAIL	57.8%	53	893				
1	Rural	PINECREST	60.0%	100	993	909.7	761.7	75	10.156
1	Rural	MERRYWOOD	60.1%	90	1,083				
1	Rural	DIAMOND HILL	60.3%	38	1,121				
1	Rural	SPRINGFIELD	60.9%	89	1,210				
1	Rural	TAMASSEE-SALEM	61.7%	41	1,251				
1	Rural	WESTMINSTER	67.2%	62	1,313				
1	Rural	HODGES	67.5%	36	1,349				
1	Rural	LAURENS	67.7%	92	1,441				
1	Rural	GRAY COURT OWINGS	68.7%	58	1,499				
1	Rural	E B MORSE	68.7%	95	1,594				
1	Rural	CLINTON	69.0%	87	1,681	1,671.4	761.7	67	11.369
1	Rural	WESTWOOD	69.8%	120	1,801				
1	Rural	ORCHARD PARK	71.5%	61	1,862				
1	Rural	WARE SHOALS PRIMARY	72.0%	55	1,917				
1	Rural	JOANNA-WOODSON	72.2%	50	1,967				
1	Rural	JAMES M BROWN	73.9%	98	2,065				
1	Rural	OAKLAND	74.8%	78	2,143				
1	Rural	BLUE RIDGE ELEMENTARY	76.6%	90	2,233				
1	Rural	WATERLOO	77.5%	37	2,270				
1	Rural	EASTSIDE	78.9%	67	2,337				
1	Rural	WOODFIELDS	80.8%	97	2,434	2,433.1	761.7	35	21.763
1	Rural	SALUDA	81.0%	107	2,541				
1	Rural	MATHEWS	83.3%	84	2,625				
1	Rural	JOHN C CALHOUN	89.9%	36	2,661				
1	Rural	FORD	92.8%	81	2,742				
1	Urban	MIDWAY SCHL	16.7%	142	2,884				
1	Urban	WREN	26.5%	100	2,984				
1	Urban	WRIGHT	28.6%	28	3,012				
1	Urban	POWDERSVILLE	31.2%	173	3,185				
1	Urban	CONCORD	31.5%	133	3,318	3,194.8	761.7	96	7.934
1	Urban	HUNT MEADOWS	39.8%	75	3,393				
1	Urban	MT LEBANON	41.9%	55	3,448				
1	Urban	MERRIWETHER	48.4%	120	3,568				
1	Urban	SPEARMAN	51.3%	60	3,628				
1	Urban	LA FRANCE	51.3%	52	3,680				
1	Urban	BELTON	52.2%	160	3,840				
1	Urban	STARR	53.2%	57	3,897				
1	Urban	CENTERVILLE	55.5%	117	4,014	3,956.5	761.7	85	8.961
1	Urban	WEST PELZER	55.6%	69	4,083				
1	Urban	HONEA PATH	55.6%	97	4,180				
1	Urban	CEDAR GROVE	55.9%	90	4,270				
1	Urban	PALMETTO	61.9%	90	4,360				
1	Urban	TOWNVILLE	63.5%	36	4,396				
1	Urban	W E PARKER	64.9%	88	4,484				
1	Urban	MCLEES	65.1%	118	4,602				
1	Urban	IVA	67.2%	70	4,672				
1	Urban	CALHOUN ACADEMY	67.6%	120	4,792	4,718.2	761.7	74	10.293
1	Urban	WHITEHALL	69.9%	75	4,867				
1	Urban	NEW PROSPECT	72.3%	66	4,933				
1	Urban	JOHNSTON	73.5%	49	4,982				
1	Urban	HOMELAND PARK	74.7%	52	5,034				
1	Urban	PENDLETON	74.7%	52	5,086				
1	Urban	FLAT ROCK .	76.5%	67	5,153				
1	Urban	NEVITT FOREST SCHOOL	80.7%	66	5,219				
1	Urban	DOUGLAS	82.4%	48	5,267				
1	Urban	VARENNES ACADEMY	90.8%	65	5,332				
2	Rural	SPARTANBURG	43.2%	45	5,377				
2	Rural	LOCKHART	58.8%	23	5,400				
2	Rural	BUFFALO	70.9%	103	5,503				

**Example #2**  
**Weight factor calculation for a survey that used systematic non-PPS sampling with implicit stratification by region, urban/rural status and F/R lunch**

This example shows the steps from sample selection to weight factor calculation when a non-PPS sampling strategy is used. Based on available resources, the decision was made to include 70 schools in the “Utopia” oral health survey of 3<sup>rd</sup> grade children. The following sampling steps were employed:

- The sampling frame list was sorted by region then by urban/rural status within each region
- Schools were then sorted by percent of children participating in the FRL program within urban/rural school categories.

**Calculations used for selecting the systematic non-PPS sample:**

- Sampling interval for sampling = (number of schools in sampling frame) / (# of schools to be screened)
  - $700 / 70 = 10.0$
- Random start = random number between 1 and interval (10) = 6.0
  - This is the first school selection number
  - There are a variety of methods for selecting a random number including, but not limited to, Excel and [www.random.org](http://www.random.org)
- Select the 6<sup>th</sup> school. Add the sampling interval (10.0) to 6 to get the next school (16.0). Continue adding the sampling interval repeatedly until all 70 school selections are made.

6.0, 16.0, 26.0, 36.0, 46.0, 56.0, 66.0, ...

- These numbers are matched to the sequential number of schools in the sampling list to identify the schools selected into the sample. The sampling frame list and the selected schools are shown in Table 2.

**Weight factor calculation (Table 2):**

- Weight = (number of children in sampling interval) / (number of children screened)
- When non-PPS sampling is used the number of children in the sampling interval will vary from one interval to another. The number of children in the sampling interval is the total of all 3<sup>rd</sup> grade children in the given interval.
- **NOTE:** In this example, dividing the number of schools by the number of schools to screen produced a whole number. Please contact ASTDD if you need more information on how to appropriately calculate weights if a fractional sampling interval was used.

**Table 2: Systematic sampling (non-PPS) with implicit stratification by region, urban/rural status and F/R lunch**

Region	Urban/ Rural	School Name	Free/Reduced Lunch Percent	3rd Grade Enrollment	Cumulative # of Schools	Sampling Interval	Selected School	# of Children in Sampling Interval (A)	Number Screened (B)	Weight (A)/(B)
1	Rural	KEOWEE	37.1%	38	1	1		840		
1	Rural	WALHALLA	39.8%	92	2	1				
1	Rural	RAVENEL	52.3%	91	3	1				
1	Rural	LAKEVIEW	52.4%	94	4	1				
1	Rural	NINETY SIX	52.7%	130	5	1				
1	Rural	<b>NORTHSIDE</b>	<b>52.8%</b>	<b>95</b>	<b>6</b>	<b>1</b>	<b>6</b>		<b>52</b>	<b>16.154</b>
1	Rural	MCCORMICK	53.9%	56	7	1				
1	Rural	FAIR-OAK	55.7%	117	8	1				
1	Rural	HICKORY TAVERN	56.8%	61	9	1				
1	Rural	HOLLYWOOD	57.2%	66	10	1				
1	Rural	CHEROKEE TRAIL	57.8%	53	11	2		659		
1	Rural	PINECREST	60.0%	100	12	2				
1	Rural	MERRYWOOD	60.1%	90	13	2				
1	Rural	DIAMOND HILL	60.3%	38	14	2				
1	Rural	SPRINGFIELD	60.9%	89	15	2				
1	Rural	<b>TAMASSEE-SALEM</b>	<b>61.7%</b>	<b>41</b>	<b>16</b>	<b>2</b>	<b>16</b>		<b>25</b>	<b>26.360</b>
1	Rural	WESTMINSTER	67.2%	62	17	2				
1	Rural	HODGES	67.5%	36	18	2				
1	Rural	LAURENS	67.7%	92	19	2				
1	Rural	GRAY COURT OWINGS	68.7%	58	20	2				
1	Rural	E B MORSE	68.7%	95	21	3		771		
1	Rural	CLINTON	69.0%	87	22	3				
1	Rural	WESTWOOD	69.8%	120	23	3				
1	Rural	ORCHARD PARK	71.5%	61	24	3				
1	Rural	WARE SHOALS	72.0%	55	25	3				
1	Rural	<b>JOANNA-WOODSON</b>	<b>72.2%</b>	<b>50</b>	<b>26</b>	<b>3</b>	<b>26</b>		<b>40</b>	<b>19.275</b>
1	Rural	JAMES M BROWN	73.9%	98	27	3				
1	Rural	OAKLAND	74.8%	78	28	3				
1	Rural	BLUE RIDGE	76.6%	90	29	3				
1	Rural	WATERLOO	77.5%	37	30	3				
1	Rural	EASTSIDE	78.9%	67	31	4		915		
1	Rural	WOODFIELDS	80.8%	97	32	4				
1	Rural	SALUDA	81.0%	107	33	4				
1	Rural	MATHEWS	83.3%	84	34	4				
1	Rural	JOHN C CALHOUN	89.9%	36	35	4				
1	Rural	<b>FORD</b>	<b>92.8%</b>	<b>81</b>	<b>36</b>	<b>4</b>	<b>36</b>		<b>63</b>	<b>14.524</b>
1	Urban	MIDWAY	16.7%	142	37	4				
1	Urban	WREN	26.5%	100	38	4				
1	Urban	WRIGHT	28.6%	28	39	4				
1	Urban	POWDERSVILLE	31.2%	173	40	4				
1	Urban	CONCORD	31.5%	133	41	5		898		
1	Urban	HUNT MEADOWS	39.8%	75	42	5				
1	Urban	MT LEBANON	41.9%	55	43	5				
1	Urban	MERRIWETHER	48.4%	120	44	5				
1	Urban	SPEARMAN	51.3%	60	45	5				
1	Urban	<b>LA FRANCE</b>	<b>51.3%</b>	<b>52</b>	<b>46</b>	<b>5</b>	<b>46</b>		<b>38</b>	<b>23.63</b>
1	Urban	BELTON	52.2%	160	47	5				
1	Urban	STARR	53.2%	57	48	5				
1	Urban	CENTERVILLE	55.5%	117	49	5				
1	Urban	WEST PELZER	55.6%	69	50	5			850	
1	Urban	HONEA PATH	55.6%	97	51	6				
1	Urban	CEDAR GROVE	55.9%	90	52	6				
1	Urban	PALMETTO	61.9%	90	53	6				
1	Urban	TOWNVILLE	63.5%	36	54	6				
1	Urban	W E PARKER	64.9%	88	55	6				
1	Urban	<b>MCLEES</b>	<b>65.1%</b>	<b>118</b>	<b>56</b>	<b>6</b>	<b>56</b>	<b>79</b>		<b>10.759</b>
1	Urban	IVA	67.2%	70	57	6				
1	Urban	CALHOUN ACADEMY	67.6%	120	58	6				
1	Urban	WHITEHALL	69.9%	75	59	6				
1	Urban	NEW PROSPECT	72.3%	66	60	6		570		
1	Urban	JOHNSTON	73.5%	49	61	7				
1	Urban	HOMELAND PARK	74.7%	52	62	7				
1	Urban	PENDLETON	74.7%	52	63	7				
1	Urban	FLAT ROCK	76.5%	67	64	7				
1	Urban	NEVITT FOREST	80.7%	66	65	7				
1	Urban	<b>DOUGLAS</b>	<b>82.4%</b>	<b>48</b>	<b>66</b>	<b>7</b>	<b>66</b>		<b>25</b>	<b>22.800</b>
1	Urban	VARENNES ACADEMY	90.8%	65	67	7				
2	Rural	SPARTANBURG	43.2%	45	68	7				
2	Rural	LOCKHART	58.8%	23	69	7				
2	Rural	BUFFALO	70.9%	103	70	7				

**Table 3: Results from analyses of an oral health dataset using SAS, Epi Info, R, SUDAAN, SPSS and Stata**

Oral health variable	Kindergarten			3 <sup>rd</sup> Grade		
	Estimated %	95% CI Lower limit	95% CI Upper limit	Estimated %	95% CI Lower limit	95% CI Upper limit
<b>Decay experience (% yes)</b>						
SAS 9.3	43.1	38.6	47.6	57.6	54.2	61.0
Epi Info 7	43.1	38.6	47.6	57.6	54.2	61.0
R	43.1	38.7	47.5	57.6	54.2	60.9
SUDAAN	43.1	38.7	47.6	57.6	54.1	61.0
SPSS	43.1	38.7	47.6	57.6	54.1	61.0
Stata	43.1	38.6	47.6	57.6	54.2	61.0
<b>Untreated decay (% yes)</b>						
SAS 9.3	19.7	16.8	22.5	21.3	18.8	23.8
Epi Info 7	19.7	16.8	22.5	21.3	18.8	23.8
R	19.7	16.9	22.5	21.3	18.8	23.7
SUDAAN	19.7	16.9	22.7	21.3	18.9	23.9
SPSS	19.7	16.9	22.7	21.3	18.9	23.9
Stata	19.6	16.8	22.5	21.2	18.7	23.7
<b>Dental sealants (% yes)</b>						
SAS 9.3	NA			29.0	25.6	32.4
Epi Info 7				29.0	25.6	32.4
R				29.0	25.7	32.3
SUDAAN				29.0	25.8	32.5
SPSS				29.0	25.7	32.5
Stata				29.0	25.6	32.4

**Table 4: Results from analyses of an oral health dataset using SAS, Epi Info, R, SUDAAN, SPSS and Stata**

Oral health variable	3 <sup>rd</sup> Grade Students					
	Non-Hispanic White			Non-Hispanic Black		
	Estimated %	95% CI Lower limit	95% CI Upper limit	Estimated %	95% CI Lower limit	95% CI Upper limit
<b>Dental sealants (% yes)</b>						
SAS 9.3	31.0	26.5	35.6	25.5	21.5	29.6
Epi Info 7	31.0	26.5	35.6	25.5	21.5	29.6
R	31.0	26.6	35.5	25.5	21.6	29.5
SUDAAN	31.0	26.7	35.8	25.6	21.7	29.8
SPSS	31.0	26.7	35.7	25.5	21.7	29.8
Stata	31.0	26.5	35.6	25.6	21.5	29.7