

Slide 1 of 22

Title Slide: SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

Slide 2 of 22

SSOCS uses a complex design for selecting the sample from which data are collected. As discussed in the common training modules, when analyzing data from sample surveys, certain procedures must be used to ensure that estimates made from the data are representative of the population. Weights must be applied correctly and standard errors must be calculated using appropriate procedures. This module discusses these issues specifically in relation to analyses of data from SSOCS.

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

In addition to an overview of the sampling procedures used to select the schools to participate in SSOCS and a description of how weights are calculated, this module will describe how imputation procedures were applied to missing and anomalous data in the SSOCS data file.

Slide 3 of 22

A sampling frame defines the population from which a sample is drawn for research purposes. The sampling frame for SSOCS is based on the most recent Common Core of Data (CCD) Public Elementary/Secondary School Universe data file available at the time the sample is developed. For the 2010 SSOCS, this was the 2007-2008 CCD.

More information about the CCD can be accessed by clicking on the corresponding underlined screen text.

Slide 4 of 22

Certain types of schools from the CCD Public Elementary/Secondary School Universe file are excluded from the SSOCS sampling frame, including schools located outside of the 50 states and the District of Columbia, Bureau of Indian Education schools, special education schools, vocational schools, alternative schools, ungraded schools, and schools with a high grade of kindergarten or lower. Regular schools, charter schools, and schools that have partial or total magnet programs are included in the frame.

Researchers analyzing the SSOCS data need to be aware of these exclusions, and should be careful to indicate the population parameters when drawing conclusions about the data.

Slide 5 of 22

The same general sample design was used to select a sample of schools for SSOCS in 2000, 2004, 2006, 2008 and 2010. As in the prior collections, the objective of the 2009–2010 sample design was twofold: (1) to obtain overall cross-sectional and subgroup estimates of important indicators of school crime and safety and (2) to develop precise estimates of change in various characteristics relating to crime between the SSOCS administrations.

Slide 6 of 22

To attain these objectives, researchers drew a stratified sample of regular public schools in each year of SSOCS.

Stratification is used to ensure that different subgroups are adequately represented in the sample. Stratification involves dividing the sampling frame into relevant subgroups, or “strata”, prior to sample selection. Stratification helps to increase accuracy when estimating population parameters for these subgroups by ensuring that different subgroups of a population are represented adequately in the sample.

Slide 7 of 22

For sample allocation and sample selection purposes, SSOCS strata were defined by school level, locale, and enrollment size. School level indicates whether the school enrolls students in primary, middle or high school grades, or some combination. Locale – characterized as City, Suburb, Town, or Rural – is based on the Census-defined geographic area in which the school is located. Enrollment size is categorized into four ranges based on the number of students attending. These three explicit stratification variables have been shown to be related to school crime and thus create meaningful subgroups for this survey.

Slide 8 of 22

In addition, region and percent White enrollment were used as implicit stratification variables by sorting schools by these variables within each explicit stratum before sample selection. In this case, region is the Census region of the school, categorized as the Northeast, South, Midwest, and West. The percent White enrollment represents the percentage of students enrolled in the school who are identified as White, non-Hispanic.

Note that prior to the 2008-2009 data collection, the locale variable used to stratify the sample was derived from an 8-level metro based code. In 2008, the variable was switched to one derived from a 12-level place based code consolidated into four categories: city, suburb, town and rural. This variable has been used since then to stratify the sample.

Slide 9 of 22

To determine how many schools to sample, NCES researchers begin by defining an optimal goal for the number of completed surveys to collect in order to meet their objectives. To reach that goal, potential non-response must be taken into account, and the initial sample selected must be greater than the final goal. Each round of data collection, then, uses the previous round's response rate to calculate how many additional schools must be sampled to compensate for non-response. For example, in 2010, the goal was to collect data for at least 2,550 schools by the end of the data collection and follow-up period. To reach this goal, a stratified sample of 3,476 regular public schools was drawn for SSOCS.

Within each school level (primary, middle, high and combined), the overall desired sample of schools was allocated among the 16 cells formed by the cross-classification of enrollment size and locale.

In some instances, such as when analyses are to be calculated for very small subgroups, it may be advantageous to have units from a particular stratum over-represented in the sample. This is known as oversampling, or disproportionate sampling. It is achieved by drawing more sampling units from that stratum than would be sampled in a simple random sample. The effect of disproportionate sampling on the final estimate is then corrected through weighting.

For SSOCS, the simplest method of assigning schools to the different sampling strata would be to allocate them proportionally to the U.S. public school population. However, while the majority of U.S. public schools are primary schools, the majority of school violence is reported in middle and high schools. Therefore, a larger proportion of the desired sample of schools is allocated to middle and high schools for SSOCS. A sample of a size proportional to the number of schools assigned to each stratum is then drawn using systematic sampling within each stratum.

Slide 10 of 22

Because of the complex nature of the SSOCS sample design, weighting of the data is necessary to obtain population-based estimates, minimize non-response bias, and reduce sampling error. Sample weights allow for inferences to be made about the total population of schools from which the sampled schools are drawn. The method for calculating weights has not varied from year to year.

Base weights for the SSOCS data are calculated using the ratio of the number of schools available in the sampling frame to the number of schools selected. In order to reduce the potential of bias due to non-response, weighting classes were determined by using chi-square automatic interaction detection, or CHAID. CHAID partitions the sample so that schools within a weighting class have the same probability of responding. The base weights were adjusted for nonresponse so that the weighted distribution of the responding schools resembled the initial distribution of the total sample.

SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

The nonresponse-adjusted weights were then poststratified to calibrate the sample to known population totals. Two-dimensional margins were set up for the poststratification based on instructional level and school enrollment size; and instructional level and locale. These data from SSOCS were then compared to the corresponding data in the CCD through an iterative process known as the raking ratio adjustment, which brought the weights into agreement with the known control totals in the CCD.

The final weight is labeled "FINALWGTT" in the data files. This variable must be used to produce national estimates from the sample data in the file.

Slide 11 of 22

After the data are collected, a unit response rate is calculated for SSOCS. The unit response rate is the ratio of completed eligible respondents to the total count of eligible respondents.

For the SSOCS data, there are three measures to evaluate response: the completion rate, the unweighted response rate and the weighted response rate.

The simplest of these measures is the completion rate. The completion rate is defined as the number of completed surveys (C) divided by the total sample size (T). In the 2010 SSOCS, this was calculated as: 2,648 completed surveys divided by 3,476 sampled schools which equaled a 76.2 percent completion rate.

While this figure represents the quality of the SSOCS data collection operations, it does not necessarily represent the quality of the data. To determine this, all schools selected for the study must be considered. A conservative measure, the unweighted response rate, divides the number of completed surveys (C) by the total initial sample size (T), subtracting known ineligible schools (I) from the denominator. For the 2010 SSOCS, this was calculated as: 2,648 completed surveys divided by (3,476 sampled schools minus 49 ineligible schools) which equaled a 77.3 percent unweighted response rate.

While unweighted unit response rates generally measure the proportion of the sample that produced usable information for analysis, weighted unit response rates can be used to estimate the proportion of the total population covered by the units that responded. The weighted unit response rate is calculated by applying the base sampling weights described earlier and substituting the result in the equation for the unweighted unit response rate.

For SSOCS, the weighted response rate is calculated by dividing the weighted number of completed surveys (Cw) by the weighted total initial sample size (Tw), subtracting the weighted number of known ineligible schools (Iw) from the denominator. In 2010, this was calculated as: 66,918.7 weighted completed surveys divided by (84,614.7 weighted sampled schools minus 1,761.1 weighted ineligible schools) which equaled an 80.8 percent weighted response rate.

Slide 12 of 22

Schools that do not respond to the survey potentially introduce bias into survey estimates, depending on the magnitude of the nonresponse and whether differences exist between responding and nonresponding schools in characteristics related to the estimates of interest. NCEs standards require analysis of nonresponse bias for any survey with a base-weighted response rate less than 85 percent. SSOCS 2010 had a base-weighted response rate of 80.8%. As a result, a unit level nonresponse bias analysis was conducted to evaluate the extent of this bias in SSOCS data. Responding and nonresponding schools were compared across the characteristics available for both groups from the CCD, including: school level, enrollment size, locale, percent White enrollment, region, number of FTE teachers, student-to-teacher ratio, and percentage of students eligible for free or reduced-price lunch. This analysis indicated that there were no measurable differences between the responding schools and the full sample of schools, leading to the determination that nonresponse bias may not be an issue for these data.

For more detailed information on the nonresponse bias analysis, please refer to the User's Manual, which can be accessed by clicking on the underlined screen text, 'non-responding schools'.

Slide 13 of 22

Just as schools sometimes choose not to respond to the SSOCS survey request, those that do respond occasionally do not answer all of the survey questions. As with unit response rates, researchers evaluated the impact of item response rates on the data. Unweighted item response rates are calculated by dividing the number of sampled schools responding to an item by the number of schools asked to respond to the item. For example, if only 1,000 schools responded to a question on the survey designed to be asked of all survey respondents, the unweighted response rate would be:

1,000 responses to the question divided by 2,476 schools asked to respond to the question which equaled a 40 percent unweighted item response rate for that question.

Some questions are not designed to be answered by all respondents. In those cases, response rates are calculated using the number of eligible respondents in the denominator. In the data file, variable response values of -1 indicate items appropriately skipped by respondents based on other criteria.

Once the unit base weight described earlier is applied, weighted item-level response rates in SSOCS are generally high. In the 2010 data file, item response rates range from 81 to 100 percent.

For each survey item with a response rate below 85 percent, NCEs requires that an item-level bias analysis be performed. Researchers should refer to the user's manual for each year of data they are including in their analysis to determine which items fall below the 85 percent response rate. In the 2010 data file, three variables were found to have response rates below 85 percent. However, an analysis of non-response bias

SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

comparing those schools that completed the questions to non-respondents concluded that there was little potential for non-response bias for these three items.

Slide 14 of 22

In the common module titled, 'Statistical Analysis of NCES Datasets Employing a Complex Sample Design', two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization. The SSOCS data file includes variables that can be used to calculate standard errors using these two procedures. Data users should be aware that the use of different approximation methods or software packages in the calculation of standard errors may result in slightly different results. Standard errors computed using the jackknife replication method and Taylor Series linearization are nearly always very similar, but not identical. Standard errors for selected SSOCS variables are calculated and reported on the NCES website for your reference and can be accessed by clicking on the underlined screen text, '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. SSOCS 2010 uses the jackknife method of replication, which divides the sample into 50 subsamples for the computation of the replicate weights. These replicate weights are included in the SSOCS 2010 data file as the variables labeled 'REPWGT1' through 'REPWGT50'. These weights can be used to calculate standard errors for estimates in a number of statistical software packages, such as R, SAS, SAS-callable SUDAAN, SPSS, and Stata.

Slide 15 of 22

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 will select the PSU and strata identifiers named 'SCHID' and 'STRATA', and specify Taylor Series linearization as the method within the statistical software used for analysis.

The 2010 SSOCS Data File User's Manual contains sample code for calculating standard errors for SSOCS data using these methods in SAS, SAS-callable SUDAAN, Stata, and SPSS statistical software programs.

Slide 16 of 22

Files containing missing data can be problematic for research. Depending on how the missing data are treated, analysis of incomplete datasets may cause users to arrive at different conclusions. Another problem with missing data is that certain groups of respondents may be more likely than others to leave some survey items unanswered, creating bias in the survey estimates. To minimize the impact of missing data in SSOCS two processes were used: editing and imputation. These processes were maintained from year to year of SSOCS data collection.

SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

Editing data took the form of two procedures: computer program-based editing, and data retrieval follow-up interviewing.

Computer-based editing begins with the process of running the returned questionnaire through a program to check whether it is complete. Then, editing programs check the data for valid data ranges, consistency, and skip patterns.

During the data collection period and shortly after, data retrieval follow-up interviews were conducted for those **responses** that did not meet the criteria for being “complete” or that had responses to critical survey items that were flagged during data processing as out of range or missing.

Slide 17 of 22

Even after computer program-based editing and data retrieval interviewing, completed SSOCS surveys still contain some level of item nonresponse. As a result, imputation procedures are used to create values for all questionnaire items with missing information.

It should be noted that the base-weighted item response rates for SSOCS 2010 were generally high. The mean weighted item response rate was about 98 percent, which is relatively high for a mailed, self-administered questionnaire. In fact, after editing, 99 percent of items in the 2010 data file had weighted response rates of over 85 percent. As such, while imputation was used for the remaining missing data, most of the data in the SSOCS file are the originally entered values from the respondents.

The imputation methods used in SSOCS 2010 were tailored to the nature of each survey item. Three methods were used: aggregate proportions, best match, and clerical. Imputation flags for each variable in the survey indicate which method was used when a value for that item was imputed.

Slide 18 of 22

To impute data, values are estimated using data from respondents matched on key characteristics to the respondent with the missing data. When imputing counts of incidents or disciplinary actions in SSOCS data, it is important to maintain relationships between survey items and school characteristics in the sample frame. Therefore, rather than imputing counts from a single donor or a mean count from a group of donors, proportions were imputed using two methods. The imputed proportions were derived for most items from aggregate proportions found by summing across all donor schools within an imputation class, defined by instructional level and enrollment size category, and dividing by the sum of the number of enrolled students within that donor class.

For a select number of items, donor groups were formed by selecting five donor schools with the identical instructional level and enrollment size category as the recipients. Regardless of how the donors were selected, the donor proportion was assigned to recipient schools in that imputation class, and the proportion was multiplied by a known value for the recipient school, such as number of students. Unlike mean imputation, this

SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

method maintains variability. Since the proportion is based on multiple donors, the result is also more stable than if it had been based on a single donor. By using more stable, aggregate proportions, imputation of outlier values is also minimized.

Slide 19 of 22

Categorical variables and several of the continuous variables in the survey are imputed using the “best-match” method. Donor classes were defined by instructional level, enrollment size category, locale, and the three categorical survey variables most strongly correlated with the variable to be imputed. Whenever possible, a recipient received data from a “perfect” donor – that is, one that matched on all of the variables used to define the imputation class. If more than one perfect donor was available, the donor was randomly assigned. If a perfect donor was not available, the least correlated questionnaire variable was dropped, and another search was conducted. The process of first dropping least correlated questionnaire variables and then dropping imputation class variables continued until a suitable donor was found. Imputation flags indicate whether a perfect donor was available or whether criteria had to be relaxed to find a suitable donor.

Slide 20 of 22

In some instances, missing data are available from other sources. For example, the CCD sampling frame was used to impute values for those schools missing student enrollment data for item 25. CCD data were also available on school type for item 31 and the percentage of students eligible for free or reduced-price lunch, item 26a. In other instances, research was done on school administrative records to estimate logical values for missing data.

Slide 21 of 22

The interrelationships between the items in the SSOCS survey require that a specific imputation order be followed. For example, because item 25 – student enrollment – is used in imputation for other variables, this item is imputed first. In some cases, values for a particular variable are limited by the values of other variables. For example, the matrix of disciplinary actions by offenses in item 23 is related to the total number of offenses recorded and the total number of disciplinary actions recorded in earlier questions. If values must be imputed for these earlier items, this is done before item 23.

Slide 22 of 22

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

Additionally, this module described how imputation procedures were applied to missing and anomalous data in the SSOCS data file.

SSOCS Sample Design, Weights, Variance, and Imputed and Missing Data

Important resources that have been provided throughout the module are summarized in this slide 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.