

## B. Collection of information employing statistical methods.

The statistical methods used in the sample design of the survey are described in this section. The documents listed below are attached or available at the hyperlink provided. These documents are either referenced in this section or provide additional information.

*Overview of the Survey of Occupational Injuries and Illnesses Sample Design and Estimation Methodology - Presented at the 2008 Joint Statistical Meetings (10/27/08)-- <http://www.bls.gov/osmr/pdf/st080120.pdf>*

*Deriving Inputs for the Allocation of State Samples (04/01/10)*

*Methods Used To Calculate the Variances of the OSHS Case and Demographic Estimates (2/22/02)*

*Variance Estimation Requirements for Summary Totals and Rates for the Annual Survey of Occupational Injuries and Illnesses (6/23/05)*

*BLS Handbook of Methods - Occupational Safety and Health Statistics (September 2008) -- <http://www.bls.gov/pub/hom/pdf/homch9.pdf>*

### 1. Description of universe and sample.

#### **Universe**

The main source for the SOII sampling frame is the BLS Quarterly Census of Employment and Wages (QCEW). The QCEW is a near quarterly census of employers collecting employment and wages by ownership, county, and six-digit North American Industry Classification System (NAICS) code. States have an option to either use the QCEW or supply public sector sampling frames for State and local government units.

The potential number of respondents (establishments) covered by the scope of the survey is approximately 7.7 million, although only about 1 million employers keep records on a routine basis due to recordkeeping exemptions defined by OSHA for employers in low hazard industries and employers with less than 11 employees, or having no recordable cases. The occupational injury and illness data reported through the annual survey are based on records that employers in the following North American Industry Classification System (NAICS) industries maintain under the Occupational Safety and Health Act:

| Sector     | Description  |
|------------|--|
| 11         | Forestry, Fishing, and Hunting   |
| 21         | Mining   |
| 22         | Utilities  |
| 23         | Construction   |
| 31, 32, 33 | Manufacturing  |
| 42         | Wholesale Trade  |
| 44,45      | Retail Trade   |
| 48,49      | Transportation and Warehousing   |
| 51         | Information  |
| 52         | Finance and Insurance  |
| 53         | Real Estate and Rental and Leasing                                       |
| 54         | Professional, Scientific, and Technical Services                         |
| 55         | Management of Companies and Enterprises                                  |
| 56         | Administrative and Support and Waste Management and Remediation Services |
| 61         | Educational Services   |
| 62         | Health Care and Social Assistance  |
| 71         | Arts, Entertainment, and Recreation                                      |
| 72         | Accommodation and Food Services  |
| 81         | Services (except Public Administration)                                  |

Excluded from the national survey collection are:

- Self-employed individuals;
- Farms with fewer than 11 employees (Sector 11);
- Employers regulated by other Federal safety and health laws;
- United States Postal Service and;
- Federal government agencies.

Mining and Railroad industries are not covered as part of the sampling process. The injury and illness data from these industries are furnished directly from the Mine Safety and Health Administration and the Federal Railroad Administration, respectively, and used to produce State and national level estimates.

Data collected for reference year 2008 and published in calendar year 2009 marked the first time State and local government agency data were collected for all States and published for all States and the nation as a whole.

The SOII is a Federal/State cooperative program, in which the Federal government and participating States share the costs of participating State data collection activities. State participation in the survey may vary by year. Sample sizes are determined by the participating States based on budget constraints and independent samples are selected for each State annually. Data are collected by BLS regional offices for non-participating States.

For the 2011 survey, 40 states plus the District of Columbia plan to participate in the survey. For the remaining 10 states which are referred to as Non-State Grantees (NSG), a smaller sample is selected to provide data which contribute to national estimates only.

The ten NSG States for 2011 are:

|              |               |               |
|--------------|---------------|---------------|
| Colorado     | Idaho         | Massachusetts |
| Mississippi  | New Hampshire | North Dakota  |
| Ohio         | Pennsylvania  | Rhode Island  |
| South Dakota |               |               |

Additionally, estimates are tabulated for three U.S. territories- Guam, Puerto Rico, and the Virgin Islands-but data from these territories are not included in the tabulation of national estimates.

### **Sample**

The SOII utilizes a stratified probability sample design with strata defined by State, ownership, industry, and size class. The first characteristic enables all the State grantees participating in the survey to produce estimates at the State level. Ownership is defined into three categories: State government, local government, and private industry. There are varying degrees of industry stratification levels within each State. This is desirable because some industries are more prevalent in some States compared to others. Also, some industries can be relatively small in employment but have high injury and illness rates which make them likely to be designated for estimation. Thus, States determine which industries are most important in terms of publication and the extent of industry stratification is set independently within each State. These industry classifications are referred to as Target Estimation Industries (TEI).

Finally, establishments are classified into five size classes based on average annual employment and defined as follows:

| Size Class | Average Annual Employment |
|------------|---------------------------|
| 1          | 10 or less                |
| 2          | 11-49                     |
| 3          | 50-249                    |
| 4          | 250-999                   |
| 5          | 1000 or greater           |

After the units are assigned to their respective stratum, a systematic selection with equal probability is used to select a sample from each sampling cell (stratum). As mentioned earlier, a sampling cell is defined as State/ownership/TEI/size class. Prior to sample selection, units within a sampling cell are sorted by employment and then by LDB number (unique identifier assigned to each reporting unit on the QCEW) to ensure a consistent representation of all employments in each stratum. Full details of the survey design are provided in Section 2.

For survey year 2011, the sample size will be approximately 240,000 or three percent of the total 7.7 million establishments in State, local, and private ownerships.

Response rate. The survey is a mandatory survey, with the exception of State and local government units in the States listed below:

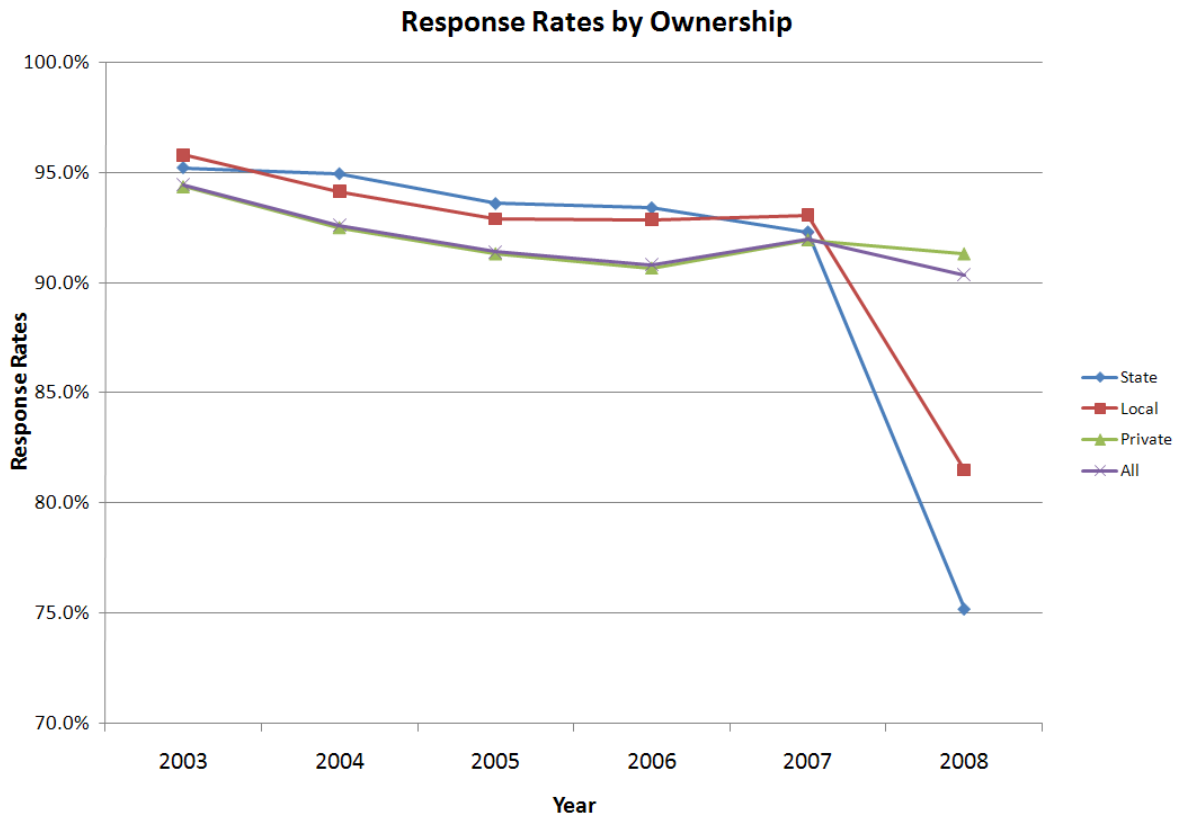
|                      |              |           |               |
|----------------------|--------------|-----------|---------------|
| Alabama              | Arkansas     | Colorado  | Delaware      |
| District of Columbia | Florida      | Georgia   | Idaho         |
| Illinois             | Kansas       | Louisiana | Massachusetts |
| Mississippi          | Missouri     | Montana   | Nebraska      |
| New Hampshire        | North Dakota | Ohio      | Pennsylvania  |
| Rhode Island         | South Dakota | Texas     |               |

Aggregate response rates in the SOII have historically been above 90% due to the mandatory nature of the survey and the excellent efforts to obtain survey data by our State and Regional partners. Thus, no specific nonresponse bias studies were conducted. Each year, respondents in the SOII survey are notified of their requirement to participate via mail. All non-respondents are sent up to two non-response mailings as a follow-up to the initial mailing. Some States choose to send a third non-response mailing to non-respondents late in the collection period. For Survey Year 2008, approximately half of the States sent an optional third non-response mailing to a majority of the non-respondents at that point in time. In addition, States may contact respondents via telephone for additional non-response

follow-up. No systematic establishment level data on the number of telephone non-response follow-up contacts is captured.

As mentioned earlier, public sector establishments were included in the 2008 survey for **all** States, including those from which no public sector data had been collected in the past. In these states, public sector establishments have no mandate to provide data to the SOII; their participation is voluntary. For survey year 2008, the rates for both State and Local government decreased, primarily due to the addition of the voluntary State and Local government establishments.

The table below illustrates the establishment level response rates from 2003-2008:



Although response rates for the SOII program have historically been high, the expansion of public sector collection in voluntary States resulted in a response rate of 75 percent in State government. Per OMB statistical guidelines, a nonresponse bias

study has been initiated. This work currently includes analysis of characteristics of respondents compared to nonrespondents using sampling frame variables such as NAICS industry, employment size class, ownership category (i.e. private vs. State government vs. Local government), size of the state, and targeted estimation industries. This work also includes an evaluation of response rates for collection of State and Local government units in mandatory collection states versus voluntary collection states. It also includes an analysis of response rates based on available modes of collection within each State and will include a multivariate analysis of nonresponse rate looking at several combinations of the variables being analyzed above. As the current analysis effort is completed, the specifics of the multivariate analysis will be specified more completely.

Additional response efforts are being conducted to analyze response rates for several key data elements collected for each establishment in the survey. Data elements for NAICS industry, SOC occupation, source, nature, part, and event for each case with days away from work are coded by BLS regional staff and/or State partners. As such, these fields are always available for collected data. Other data elements such as ethnicity, whether the event occurred before/during/after the work shift, the time of the event, and the time the employee began work may be missing from collected data. We have initiated a response analysis effort for these other data elements to identify our specific response rates and the characteristics of respondents versus non-respondents for these variables.

Regional offices are also working with States on collection practices to improve response for voluntary units.

We will continue to monitor the response rates in the next 3 years for all segments of the survey scope. We will update the analysis each year and make recommendations for improvements in the data collection process based on the results of our analysis. If response rates at the establishment level remain below 80%, we will conduct additional non-response bias studies. If response rates for any specific data element within establishments are below 70%, we will also implement additional non-response bias studies. Details for these studies will be documented as the studies begin.

## 2. Statistical methodology.

Survey design. The survey is based on probability survey design theory and methodology at both the national and State levels. This methodology provides a statistical foundation for drawing inference to the full universe being studied.

Research was done to determine what measure of size was most appropriate for the allocation module. Discussion with OSHS

Management narrowed the choices to the rates for Total Recordable Cases (TRC); Cases with Days Away from Work (DAFW); and Cases with Days Away from Work, Job Transfer, or Restriction (DART).

Rates from the 2003 SOII were studied for all 1251 TEIs for each of the above case categories. The average case rate, standard deviation (SD), and coefficient of variation (CV) for each set of rates were calculated. The CV is the standard deviation divided by the estimate, which is commonly used to compare estimates in relative terms. The results are shown below:

| <u>Description</u> | <u>Ave. Rate</u> | <u>SD</u> | <u>CV</u> |
|--------------------|------------------|-----------|-----------|
| DAFW               | 1.5540           | 1.078     | 0.69      |
| DART               | 3.0479           | 2.000     | 0.66      |
| TRC                | 5.5300           | 3.229     | 0.58      |

Based on this information it was recommended that the TRC rate be used as the measure of size for the sample allocation process for the survey. The lower CV indicates that it is the most stable indicator.

Additionally, to fulfill the needs of users of the survey statistics, the sample provides industry estimates. A list of the industries for which estimates are required is compiled by the BLS after consultation with the principal Federal users. The sample is currently designed to generate national data for all targeted NAICS levels that meet publication standards.

Allocation procedure. The principal feature of the survey's probability sample design is its use of stratified random sampling with Neyman allocation. The characteristics used to stratify the units are State, ownership (whether private or state or local government), industry code, and employment size class. Since these characteristics are highly correlated with the characteristics that the survey measures, stratified sampling provides a gain in precision and thus results in a smaller sample size.

Using Neyman allocation, optimal sample sizes are determined for each stratum within each State. Historical case data are applied to compute sampling errors used in the allocation process. Details about this process can be found in ***Deriving Inputs for the Allocation of State Samples*** (04/18/05).

The first simplifying assumption for allocation is that for each TEI  $\times$  size class stratum  $h$ , the employment in each establishment is the same, which is denoted by  $E_h$ . We also ignore weighting adjustments. In addition, we assume that the sampling of establishments in each stratum is simple random sample with replacement. (It is actually without replacement of course, but this is a common assumption to simplify the formulas.)

One consequence of these assumptions is that the estimate of the overall employment is constant and as a result the estimated incidence rate of recordable cases in the universe is the estimated number of recordable cases divided by this constant. Therefore, the optimal allocation for the total number of recordable cases and the incidence rate of recordable cases are the same. We will only consider the optimal allocation for the total number of recordable cases.

We introduce the following notation. For sampling stratum  $h$  let:

$N_h$  denote the number of frame units

$n_h$  denote the number of sample units

$W_h = N_h/n_h$  denote the sample weight

$T_h = N_h E_h$  denote the total employment in stratum  $h$

$p_h$  denote the incident rate for total recordable cases

$\hat{Y}_h$  denote the unweighted sample number of recordable cases



Also let:

$\hat{Y}$  denote the estimated number of recordable cases in the entire universe.

Then

$$\hat{Y} = \sum_h W_h \hat{Y}_h = \sum_h \frac{N_h \hat{Y}_h}{n_h} \quad (1)$$

$$V(\hat{Y}) = \sum_h \frac{N_h^2 V(\hat{Y}_h)}{n_h^2} \quad (2)$$

where  $V$  denotes variance.

Now we will obtain  $V(\hat{Y}_h)$  under two different assumptions. Assumption (a) is:

(a) All employees in stratum  $h$  have either 0 or 1 recordable cases and the probability that an employee has a recordable case is  $p_h$ .

In this case  $\hat{Y}_h$  can be considered to have a binomial distribution with  $n_h E_h$  trials and  $p_h$  the probability of success in each trial and consequently

$$V(\hat{Y}_h) = n_h E_h p_h (1 - p_h) \quad (3)$$

Assumption (b) is:

(b) The total recordable case rate for the  $n_h$  sample establishments in stratum  $h$  has a binomial distribution with  $n_h$  trials and  $p_h$  the probability of success in each trial. In that case

$$V(\hat{Y}_h) = n_h E_h^2 p_h (1 - p_h) \quad (4)$$

Although we will derive the optimal allocations under both assumptions, we prefer assumption (b) since under assumption (a) the variance of the recordable case rate among establishments in

stratum  $h$  we believe will be unrealistically small, particularly for strata with large  $E_h$ .

To derive the optimal allocation under assumption (a) we substitute (3) into (2) obtaining

$$V(\hat{Y}) = \sum_h \frac{N_h^2 E_h p_h (1-p_h)}{n_h} \quad (5)$$

Viewing (5) as a function of the variables  $n_h$  and minimizing (5) with respect to these variables by means of the method of Lagrange multipliers from advanced calculus, we obtain that (5) is minimized when the  $n_h$  are proportional to

$$N_h (E_h * p_h * (1-p_h))^{.5} \quad (6)$$

As for the preferred assumption (b), to derive the optimal allocation, we similarly substitute (4) into (2) obtaining

$$V(\hat{Y}) = \sum_h \frac{N_h^2 E_h^2 p_h (1-p_h)}{n_h} \quad (7)$$

Minimizing (7) as we minimized (5), we obtain that (7) is minimized when the  $n_h$  are proportional to

$$N_h * E_h * (p_h (1-p_h))^{.5} = T_h (p_h (1-p_h))^{.5} \quad (8)$$

which is the preferred allocation.

Sample procedure. Once the sample is allocated, the process of selecting the specific units is done by applying a systematic selection with equal probability for each sampling cell. Because the frame is stratified by employment size within each TEI before sample selection, it was felt equal probability sampling was appropriate rather than a PPS selection. PPS selection is often applied to frames that aren't stratified by size so in this case, it was felt that no additional value would be gained by selecting the sample by PPS.

The survey is conducted by mail questionnaire through the BLS-Washington and Regional Offices and participating State statistical grant agencies.



Estimation procedure. The survey's estimates of the population totals are based on the Horvitz-Thompson estimator, which is an unbiased estimator. The precision of the estimates is further improved by using the ratio estimator that utilizes available auxiliary information to improve the estimate. The estimates of the incidence rates are calculated as:  $N/EH \times 200,000$ , where:

N = number of injuries and illnesses  
EH = total hours worked by all employees during a calendar year  
200,000 = base for 100 full-time equivalent workers (working 40 hours per week, 50 weeks per year).

The estimation system has several major components that are used to generate summary estimates. The first four components generate factors that are applied to each unit's original weight in order to determine a final weight for the unit. These factors were developed to handle various data collection issues. The original weight that each unit is assigned at the time the sample is selected never changes; however, a final weight is calculated that is equal to the product of the original weight and each of the factors calculated by the estimation system. The following is a synopsis of these four components.

When a unit cannot be collected as assigned, it is assigned a **Reaggregation** factor. For example, if XYZ Company exists on the sample with 1000 employees but the respondent reports for only one of two locations with 500 employees each, it is treated as a reaggregation situation. The Reaggregation factor is equal to the target (or sampled) employment divided by the reported employment. It is calculated for each individual unit. It is initially set to 1 for all units.

In cases where a sampled unit is within scope of the survey but does not provide data, it is treated as a nonrespondent. Units within scope are considered viable units. This would include collected units as well nonrespondents. The **Nonresponse** adjustment factor is the sum of the weighted viable employment divided by the sum of the weighted usable employment for an entire stratum. One nonresponse adjustment factor is applied to each unit in a stratum. It is initially set to 1 for all units.

In some cases, collected data is so extreme that it stands apart from the rest of the observations. For example, suppose in a dental office (which is historically a low incidence industry for injuries and illnesses, poisonous gas gets in the ventilation system which causes several employees miss work for several days. This is a highly unusual circumstance for that industry. This situation would be deemed an Outlier for estimation purposes and handled with the Outlier adjustment. The **Outlier adjustment factor** is initially set to 1 for all units. If any outliers are identified and approved by the national office, the system calculates an outlier adjustment factor so that the outlier represents only itself. In addition, the system calculates outlier adjustment factors for all other non-outlier units in the sampling stratum.

**Benchmarking** is done in an effort to account for the time lapse between the sampling frame used for selecting the sample and the latest available frame information. Thus, a factor is computed by dividing the target employment (latest available employment) by the weighted reported employment for collected units. The Benchmark adjustment factor is initially set to 1 for all units.

The system calculates a final weight for each unit. The final weight is a product of the original weight and all four of the factors. All estimates are the sum of the weighted (final weight) characteristic of all the units in a stratum.

### 3. Statistical reliability.

#### Survey sampling errors.

The survey utilizes a full probability survey design that makes it possible to determine the reliability of the survey estimates. Standard errors are produced for all injury and illness counts and case and demographic data as well for all data directly collected by the survey.

The variance estimation procedures are described in detail in the attached documents mentioned earlier:

***Methods Used To Calculate the Variances of the OSHS Case and Demographic Estimates (2/22/02)***  
***Variance Estimation Requirements for Summary Totals and Rates for the Annual Survey of Occupational Injuries and Illnesses (6/23/05)***

#### 4. Testing procedures.

The survey was first undertaken in 1972 with a sample size of approximately 650,000. Since then the BLS has made significant progress toward reducing respondent burden by employing various statistical survey design techniques; the present sample size is approximately 240,000. The BLS is continually researching for methods that will reduce the respondent burden without jeopardizing the reliability of the estimates.

Responding to concerns of data users and recommendations of the National Academy of Sciences, in 1989, the BLS initiated its efforts to redesign the survey by conducting a series of pilot surveys to test alternative data collection forms and procedures. Successive phases of pilot testing continued through 1990 and 1991. Cognitive testing of that survey questionnaire with sample respondents was conducted at that time. The objective of these tests was to help develop forms and questions that respondents easily understand and can readily answer.

In 2003, the BLS introduced the Internet Data Collection Facility (IDCF) as an alternative to paper collection of data. This system has edits built in which helps minimize coding errors. The system is updated annually to incorporate improvements as a result of experience from previous years.

In survey year 2006, the SOII program conducted a one-year quality assurance study that had primarily a focus on addressing the magnitude of employer error in recording data from their OSHA forms to the different types of BLS collection forms and methods. The results showed no systematic under-reporting or over-reporting by employers. There was no strong dependence between error rates and collection methods.

Beginning in survey year 2007, the QA program was modified to evaluate the quality of the data collected in terms of proper collection methods. If improper collection methods or procedures were uncovered, they were corrected. A byproduct of this program was that each data collector would know that any form they have processed could be selected for the program.

The BLS also utilizes statistical quality control techniques to maintain the system's high level of reliability.

5. Statistical responsibility.

The Statistical Methods Group, Chief, Gwyn Ferguson is responsible for the sample design which includes selection and estimation. Her telephone number is 202-691-6941. The sample design of the survey conforms to professional statistical standards and to OMB Circular No. A46.