NATIONAL OCCUPATIONAL EXPOSURE SURVEY

SAMPLING METHODOLOGY

Wm. Karl Sieber, Jr.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control
National Institute for Occupational Safety and Health
Division of Surveillance, Hazard Evaluations and Field Studies
Cincinnati, Ohio 45226

February 1990

DISCLAIMER

The contents of this report are expressed as received from the contractor.

Mention of a company name or product does not constitute endorsement by the National Institute for Occupational Safety and Health.

DHHS(NIOSH) Publication No. 89-102

FOREWORD

The National Occupational Exposure Survey (NOES) was a nationwide observational survey conducted in a sample of nearly 5,000 establishments from 1981-1983. The goal of the NOES was to compile data on the types of potential exposure agents found at the workplace, and the kinds of safety and health programs which had been implemented at the plant level. The sample of establishments included in the survey was designed to represent those segments of American industry covered under the Occupational Safety and Health Act of 1970.

This volume describes the method used to select the sample of plants to be surveyed, and the estimation techniques used to project survey data to national estimates.

I. ABSTRACT

The National Occupational Exposure Survey (NOES) of 1981-1983 was initiated by NIOSH to address a critical and continuing need for information on nationwide patterns of occupational exposures to potential health hazards. The NOES consisted of on-site observational surveys in a sample of nearly 5,000 establishments which had been selected to represent most sectors of the American workforce covered by the Occupational Safety and Health Act.

A two-stage sampling strategy was employed to construct the sample of establishments to be surveyed. The first stage resulted in the selection of 98 geographical areas, or primary sampling units. The geographical areas chosen in the first stage had relatively higher concentrations of those industries which were included in the target population. The second stage of sampling produced lists of establishments to be surveyed in the first-stage geographical areas. Establishments with 2,500 or more employees were not included in the first stage of sampling, and were treated separately in order to maintain more nearly equal probabilities of selection across establishments.

First stage selection of geographical areas was accomplished by random selection from strata defined by geography, number of employees, and concentration of establishments included in the target population. Second stage selection of establishments employed systematic sampling from a list of establishments ordered by number of employees and Standard Industrial Classification (SIC). The second stage sample was enlarged by 25 percent, and establishments in this enlarged sample were screened by telephone to determine eligibility for inclusion in the survey. A total of 4,490 establishments were ultimately surveyed in the NOES. Substitutions were made for establishments which fell outside the scope of the survey, and inspection warrants were obtained and enforced where necessary. The effective refusal rate among establishments selected for inclusion in the survey was 0.3 percent.

Two stages of ratio estimation were used in the process of projecting survey data to national statistics. Variances of the estimates were calculated using the method of balanced repeated replications.

CONTENTS

CONTENTS (Cont.)

FIGURES

TABLES

II. ACKNOWLEDGEMENTS

The National Occupational Exposure Survey was conducted by Westat, Incorporated, 1650 Research Boulevard, Rockville, Maryland, under NIOSH contract 210-80-0057.

I would like to thank the following people for their critical review of this document:

Mr. Wallace Carr
Project Officer, National Occupational Hazard Survey-Mining
Division of Respiratory Disease Studies
National Institute for Occupational Safety and Health

Mr. Robert Hanson
Mr. John Edmonds
Ms. Diane Ward
Westat, Incorporated

Dr. Kathryn R. Mahaffey
Chief, Priorities and Research Analysis Branch
Division of Standards Development and Technology Transfer
National Institute for Occupational Safety and Health

Mr. Stanley Shulman
Mathematical Statistician
Division of Physical Sciences and Engineering
National Institute for Occupational Safety and Health

Mr. Gary Shapiro
Assistant Division Chief
Income, Longitudinal and Expenditure Survey Design
Statistical Methods Division
Bureau of the Census

I am grateful to Mr. Todd Frazier, Mr. David Sundin, and Mr. David Pedersen who each provided extensive comments as to the organization and content of this report. Mr. Randy Young provided information on computer operations during the National Occupational Exposure Survey. Mrs. Kathy Mitchell showed special patience in preparing the several drafts of this report.

III. INTRODUCTION

The National Institute for Occupational Safety and Health (NIOSH) is charged with developing information on the types and extent of exposures to occupational health hazards (1). To develop data of this type, NIOSH has carried out two on-site observational surveys of a sample of facilities representative of selected segments of American workplaces. The first, the National Occupational Hazard Survey (NOHS), was conducted by NIOSH from 1972 to 1974. The second was the National Occupational Exposure Survey (NOES) conducted by NIOSH between 1981 and 1983. To a great extent, the NOES was designed to provide results which could be compared to those obtained from the NOHS.

The NOES is a response by NIOSH to the continuing need for information on nationwide patterns of occupational exposure to health hazards. This report, second in a series of reports based on the NOES, details the development of the sample design, selection of sample establishments, and the statistical methodology developed to make national projections from data obtained by surveying a probability sample of worksites and potential workplace hazards. Volume I, National Occupational Exposure Survey - Survey Manual, detailed field guidelines and the actual questionnaire used in the NOES (2).

In summary, the objectives of the NOES were:

1. For selected industrial sectors, to develop estimates of the number of workers potentially exposed to chemical, physical, and biological agents;

2. To develop data that describe the nature and extent of these potential exposures to health hazards and the degree to which businesses have implemented programs to reduce occupational health problems; and

3. To compile data such that analysis of industrial hazard exposure trends would be possible by comparison with similar data collected in NOHS.

The target population was defined as employees working in establishments or job sites located in the United States reporting eight or more employees at the time of the survey, and with a primary activity or line of business on a list of target Standard Industrial Classification codes (SICs) (3). An establishment was defined as an economic unit, generally at a single location, where business, service, or industrial activities were performed.[1]

Development and implementation of the NOES sample design was done under NIOSH contract no. 210-80-0057 to Westat, Incorporated, Rockville, Maryland.

[1] The terms establishment, facility, firm, and worksite are used interchangeably in this report.

A. Development of the 1972-1974 Survey: The NOHS

Sampling methodology in the NOES was generally based on a design used for the NOHS. The NOHS involved a two-stage selection procedure with stratification. In that survey, primary sampling units (PSUs) were defined from the 247 Standard Metropolitan Statistical Areas (SMSAs) in 1970 and certain urban areas. Each PSU defined a geographic cluster of business and industrial establishments. The sample consisted of 67 PSUs selected with probability proportional to size, i.e., proportional to the number of establishments in defined strata.

Establishments within the 67 PSUs were stratified by probability of selection of the PSU, number of employees, and SIC code, and individual establishments were selected for field interview using systematic selection. A sample size of 4,636 facilities was determined in this manner.

Further details concerning sample selection and analysis for the NOHS is described in Volume II of the NOHS series, 'Data Editing and Data Base Development' (4).

B. 1981-1983 NOES Sampling Strategy

The NOES also used a two-stage sampling strategy for most of the sample. The first stage of sampling involved selection of a sample of 98 PSUs. PSUs in the NOES were defined across all 50 states rather than only as SMSAs as was done in the NOHS. This is one reason why more PSUs were selected in the NOES than in the NOHS. With the exception of samples of very large establishments drawn irrespective of geographic location, the interviewed sample was confined to these 98 PSUs. Stratification of PSUs was based on geography, number of employees, and concentration of establishments operating in select industries. The second stage (within PSU) selection for establishments was done using a systematic selection procedure. Very large establishments (2,500 or more employees) were treated separately in order to maintain more nearly equal probabilities of selection across establishments.

The SIC codes of firms eligible for this survey are shown in Appendix A. Establishments with eight or more employees and conducting business within this specific set of SICs (called the "target SICs") were considered to be in-scope in the NOES. Establishments with eight or more employees only were considered for comparability with the NOHS and because accurately surveying establishments with less than eight employees would have been difficult. Coverage of construction and manufacturing establishments was emphasized in the NOES by defining these SIC categories to include a broad range of SICs, while finance establishments as well as mining and mineral processing establishments were excluded from it.

The interviewed sample was designated in two steps: (1) a sample of 7,392 establishments was contacted by telephone to identify those establishments that were in the scope of the study; and (2) those establishments identified in (1) were visited and surveyed. A total of 4,490 establishments had complete field interviews in the NOES.

Figure 1 is an outline of the sampling strategy followed in the NOES. A total of 604 PSUs were defined for the sampling process. PSUs were defined geographically with the county as the primary unit. Some PSUs consisted of a single county, e.g., Orange County, California. Other PSUs were made up of counties that constituted a SMSA in 1980, which in a few cases crossed state boundaries: e.g., Cincinnati SMSA consisting of Dearborn County, Indiana; Boone, Campbell, and Kenton Counties in Kentucky; and Brown, Clermont, Hamilton, and Warren counties in Ohio. The 604 PSUs included 446,125 establishments eligible for the survey.

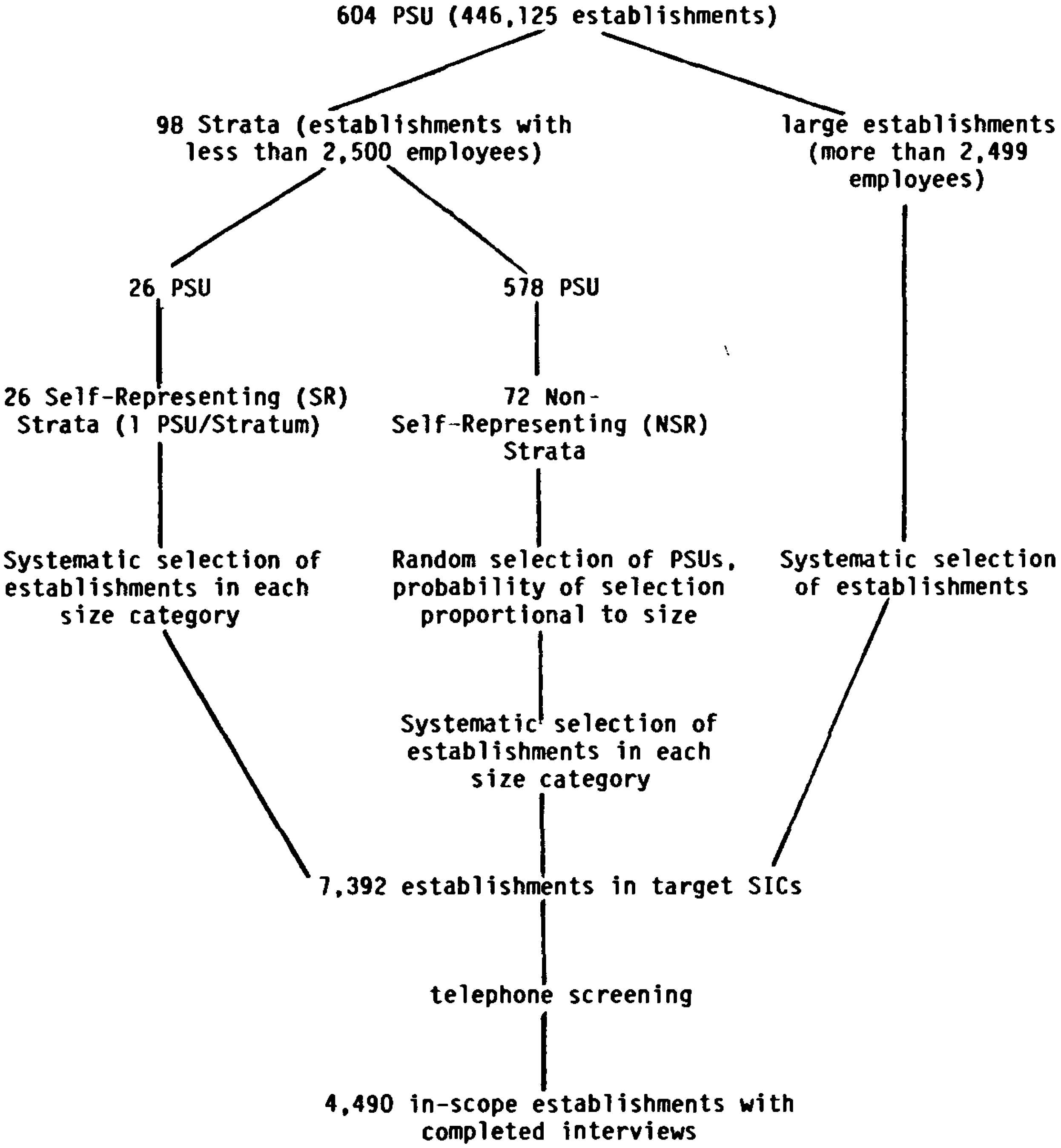
The 604 PSUs were stratified into 98 strata. The purpose of stratification was to obtain groups of PSUs which were of equal size and were homogeneous with respect to variables of interest in the NOES. Some of the criteria used for designing strata included: proportion of employees in firms where high potential exposure to health hazards might be found (e.g., chemical, rubber, or leather industries), geography (census region), and SMSA or non-SMSA. The 98 strata consisted of 26 self-representing (SR) strata, made up of 1 large PSU each, and 72 non-self-representing (NSR) strata made up of the remaining 578 PSUs.

The selection of establishments with less than 2,500 employees was done from 98 PSUs, one from each of the 26 SR and 72 NSR strata. These 98 PSUs are listed in Appendix B. A systematic sample of establishments in each of the 26 PSUs making up the SR strata was designated to be interviewed. Samples were selected independently across establishment size classes, where size was defined as the size of the workforce at that work site. PSUs in the 72 NSR strata from which establishments were to be selected were chosen as a random sample with probability of selection proportional to the number of establishments contributed by that PSU to that stratum, i.e., the measure of size of that PSU. One PSU was chosen from each NSR stratum. Systematic selection of establishments in each NSR PSU was then done using methods identical to those for selecting establishments from SR PSUs.

The sample of establishments employing 2,500 or more employees was designated without regard to sampling from PSUs. Samples in each of the size categories with these employee levels were determined using systematic selection of all firms nationwide with 2,500 or more employees.

Sample establishments were contacted by telephone to confirm that those establishments had enough employees and operated in an appropriate SIC to be included in the survey, and would participate in it. This sample was known as the 'screening' sample and consisted of 7,392 establishments. After screening, 4,504

FIGURE 1. OUTLINE OF SAMPLING STRATEGY
NOES 1981-1983


604 PSU (446,125 establishments)
98 Strata (establishments with less than 2,500 employees)
large establishments (more than 2,499 employees)
26 PSU
578 PSU
26 Self-Representing (SR) Strata (1 PSU/Stratum)
72 Non-Self-Representing (NSR) Strata
Systematic selection of establishments in each size category
Random selection of PSUs, probability of selection proportional to size
Systematic selection of establishments
Systematic selection of establishments in each size category
7,392 establishments in target SICs
telephone screening
4,490 in-scope establishments with completed interviews

establishments were designated for field interview of which all but 125 of these establishments were interviewed. Substitutes were found for 111 of the 125, making the total number of completed interviews 4,490. The effective refusal rate of establishments for participation in the NOES was .3 percent.

Two stages of ratio estimation were used in the estimation process. Variances of estimates were calculated using the method of balanced repeated replications.

Much of the sample selection was carried out as a computer operation. National estimates were also determined using a computer software package.

IV. SAMPLE DESIGN

Listings from the Bureau of the Census publication County Business Patterns - 1978 (CBP) provided data needed to establish sampling rates, while listings from the 1980 Dun and Bradstreet Market Inventory (DMI) were used to select establishments. Supplementing these lists for completeness was considered, but was not done because of the costs involved and extensive coverage of the DMI. The initial screening operation was done to select a sample of establishments employing eight or more employees and operating in one of the Standard Industrial Classification (SIC) (3) codes covered by the NOES. This screening was carried out as a telephone survey which identified establishments still in business and eligible for the survey during the 1981-1983 data collection period. The sampling plan attempted to produce minimum variance for a fixed cost by considering strata determined by number of employees at the worksite.

The design of NOES made use of prior experience from the NOHS. The NOHS data provided guidance as to the method of stratification and most efficient sampling rates in strata.

A. Sources of Data for the Sampling Frame

The design of NOES was based on information from the Bureau of Census publication County Business Patterns, 1978 (CBP) (5). The CBP was used to estimate the number of establishments and size of the workforce in establishments in each PSU. Information on individual establishments' size and location was supplied by the Dun and Bradstreet Market Inventory (DMI) (6). The DMI is a well-known and widely used industrial directory service. Historically businesses were listed in the file so as to establish credit ratings. Thus the list does not represent all U.S. industries. A special effort has been made by Dun and Bradstreet to expand the DMI file in order to have more complete listings of establishments, however, and the DMI is considered close to complete.

An examination of the completeness of the DMI was made before deciding on its use in the NOES. Establishments in the following SIC groups were found to have DMI to CBP employee ratios of less than 0.9; i.e., presumably ten percent under-representation was found in the DMI file:

- 451 & 452 - Air transportation
- 481 - Telephone communication
- 491 - Electronic services
- 493 - Combination electric, gas and other services combined
- 5541 - Gasoline service stations
- 7231 - Beauty shops
- 7241 - Barber shops
- 7299 - Miscellaneous personal services

Since supplementing the DMI list to cover these SICs was considered beyond available resources and the DMI already was quite extensive in coverage of most SIC groups included in the survey, the coverage provided by the DMI was considered without supplementation. See Appendix C for more discussion on this point.

B. Defining the Target Population

The target population was defined as those establishments or job sites located in the 50 states reporting eight or more employees and having as a primary activity one of the target SICs listed in Appendix A.

As is the case with any sample survey, inferences from the sample data are restricted to the target population. The following points provide a description of the target population.

Establishments included:

- Those located in metropolitan and other urbanized areas of the United States in 1980 and which were still worksites during the 1981 to 1983 period of data collection.
- Those reporting eight or more employees in the 1978 CBP and 1980 DMI files, provided that these establishments were still in business and operating during the period of data collection.

Establishments excluded:

- Establishments engaged in agricultural production, any mining activity except oil and gas extraction, railroad transportation, private households, finance institutions, and all Federal, State, and municipal government facilities.

Within each PSU, establishments were classified by number of employees. Eleven size classes were defined as follows: 8-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1000-1499, 1500-2499, 2500-4999, and 5000+ employees, and those for which employment totals were not available from the DMI. The two largest categories were treated separately from the others, since they represented a substantial expenditure of time for the surveyor teams and would affect calculation of variances of the survey results.

C. Derivation of the Sample Design

Methods of optimizing the sample design for a survey typically involve establishing a cost function for the study, expressing the sampling variance, and solving the equation which will produce the minimum variance for a fixed cost (7). This approach was an oversimplification of the needs for the NOES because it assumed there was a single statistic whose variance is to be minimized. There were several different types of statistics for which estimates were needed from this survey and quite different sample designs could have been chosen depending on which statistic was considered to be of greatest importance.

Much of the analysis in the 1972-1974 NOHS referred to industry-by-industry breakdowns. For these kinds of analyses, the samples of industries should have approximately equal reliability. This would lead to a sample design with roughly equal sample sizes by industry. On the other hand, an efficient sample for analysis of statistics for all industries combined would require that the sample size in each industrial sector be proportional to that sector's contribution to the total number of establishments eligible for the survey.

A second problem arises from the interest in data on the distributions of both establishments and employees. An efficient sample design for statistics on employees would use higher sampling rates for larger establishments than for smaller ones. For statistics on establishments, however, the number of plants, rather than their size, would be important.

The sample design developed for NOES maximized the reliability of estimates of numbers of employees. Although estimates of facilities are available using the methodology developed in NOES, breakdown by industry or data on the number of firms with specific characteristics was assigned lower priority in developing the sample design.

1. The Cost Function

A cost function expressing the total cost as the sum of costs over employee size strata was first determined. The cost within a size stratum was equal to the product of the number of sample establishments and the average cost of interviewing the establishments. Average costs were expressed as number of person-hours required for that size group.

This cost function recognizes only the unit costs and the total cost for those aspects of the survey that are directly affected by the sample size. The number of PSUs does not enter the cost function. There are several reasons for this. First, the cost of designating the sample of establishments, a major portion of which would involve the use of the telephone, would be directly related to the number of sample establishments and would have little relationship to the number or location of the PSUs. Second, the cost of surveyor travel between PSUs was assumed to be relatively small and concentrated during weekends and would not influence the total cost based on person hours during the regular week. Third, the time a team would be assigned to a PSU was restricted to below a 30-day maximum because of Federal government restrictions on per diem reimbursement. These conditions meant the number of PSUs would have little impact on the total survey cost, and so the number of sample PSUs was set as large as administratively feasible. A large number of PSUs also reduces the component of variance arising from the sampling of PSUs.

The cost function was defined as:

$$C = \sum_{a} n_a C_a \tag{1}$$

where

C = The total cost.

C_a = The cost per sample establishment in the ath employer size class.

a = Employee size class, 1 to 10, see Table 1.

n_a = The number of sample establishments selected in the ath size class.

The term C_a in the cost function is the total number of person hours of surveyor time per establishment in the survey in the ath size class. These unit costs varied according to the size of the establishment and were taken from a tabulation of average surveyor hours per establishment by size class experienced in the NOHS. They are listed by employee size class in Table 1. It was assumed the amount of time required to survey a sample firm would be similar to the experience in the NOHS. The total of all costs of the survey also included a number of more-or-less fixed charges that did not vary directly with moderate changes in the sample size; for example, writing specifications and computer programs for data processing, overhead costs, the cost of hiring and training surveyors, etc.

The total of all directly related costs of the sample was taken as the total number of paid person hours to support a proposed number of surveyors working for an expected survey period of two years. This ignored the cost of travel and all fixed costs. It also meant that other costs expected to vary with the sample size were assumed to be small and not important in determining the sample size. For example, the total cost of telephone screening was assumed not to be importantly affected by variations in the allocations of the sample among size classes, or by minor changes in the number of sample cases.

An average number of 21 surveyors were expected to be available for the survey period. The surveyors were to be assigned to five teams, and each team was to have a team leader. Because of time spent in team supervision, each leader was assumed to produce 80 percent of the production of the other team members. Each team member was to work a 40 hour week for 48 weeks of the year; the remaining four weeks were to be taken up by annual and

TABLE 1. SAMPLING RATES AND EXPECTED DISTRIBUTION OF SAMPLE OF ESTABLISHMENTS
NOES 1981-1983

Employee size class	Number of employees	Number of establishments[1] N_a	Survey cost (hrs.)[2] C_a	Average number of employees per facility	Relvariance factors	Sampling interval k_a	oversampling ratio f_a	Facilitie in sample n_a
1	8-19	237,445	3.18	12.4	2	199.530	1.0	1,190
2	20-49	114,508	4.28	32.4	1	125.250	1.593	914
3	50-99	44,567	5.82	71.7	1	66.030	3.022	675
4	100-249	30,601	8.66	158.1	1	36.520	5.464	838
5	250-499	10,887	14.36	349.9	1	21.260	9.384	512
6	500-999	5,055	26.79	690.1	1	14.700	13.576	344
7	1000-1499	1,424	49.78	1,200	1	11.580	17.235	123
8	1500-2499	906	66.52	1,900	1	8.389	23.785	108
9	2500-4999	520	86.16	3,500	1	5.545	35.984	94
10	5000+	212	189.25	9,250	2	2.190	91.110	97
11	Not Available	Unknown	Unknown	Unknown	1	199.530	1.000	Unknown
Total		446,125						4,895

[1] Based on tabulation of CBP county summary records for 1978.

[2] Person hours per establishment required to investigate facilities in NOHS.

sick leave and holiday time. For a 24 month data collection period, the hours contributed by the five team leaders would be:

(5 leaders) x (40 hours) x (48 weeks) x (2 years) x (.8) = 15,360 hours,

and from the remaining 16 team members:

(16 members) x (40) x (48) x (2) = 61,440 hours.

The total for all 21 surveyors = 76,800 person hours.

These assumptions turned out to be only a rough approximation of the actual survey conditions. The period of data collection was about 32 months rather than the predicted 24. Also, the survey force began initially with only 11 surveyors, rose to 15 after 5 months, and then fluctuated between 10 and 22 for most of the remaining survey period. The size of the field staff meant that surveyor teams did not function as expected. Costs in terms of hours to survey plants for NOES were also found to differ from the NOHS experience. The actual costs per establishment size class are detailed in Volume I of this series (2).

2. The Variance Function

The variance for the estimated total number of establishments was taken as:

$$\sigma^2(Y') = \sum_a N_a^2 \, (1/n_a - 1/N_a) \, S^2(\overline{Y}_a) \qquad (2)$$

where

Y' = a total, estimated from the survey.

N_a = The total number of establishments in the universe of study in the a th establishment size class.

n_a = The number of establishments in the sample from the a th size class.

$S^2(\overline{Y}_a)$ = The estimated population variance of the number of establishments with the characteristic y in the a th size class.

The estimated variance $S^2(\overline{Y}_a)$ for the a th size class is given by:

$$S^2(\overline{Y}_a) = \sum_{i=1}^{N_a} \frac{(Y_{ai} - \overline{Y}_a)^2}{(N_a - 1)} \qquad (3)$$

where

Y_{ai} = The number of employees having the characteristic Y in the i^{th} establishment in the a^{th} size class.

$\bar{Y}_a = (\sum_{i=1}^{N_a} Y_{ai})/N_a$ is the average number of employees with the characteristic per establishment in the a^{th} class.

Values of $S^2(\bar{Y}_a)$ were not available when the sample design was developed. Although data from NOHS could have been used to estimate the values of $S^2(\bar{Y}_a)$ for a selected set of characteristics, the time schedule prevented waiting for these variances to be prepared. Instead, an approximation in which the relvariances of desired characteristics were assumed to be constant within most size classes was employed (8, 9). This approximation was based on experience in other surveys.

With this assumption:

$$\text{relvariance} = \frac{S^2(\bar{Y}_a)}{(\bar{Y}_a)^2} = \text{constant}$$

$$S^2(\bar{Y}_a) = \text{constant} \times (\bar{Y}_a)^2 \qquad (4)$$

If the value of the constant and mean number of employees in size class a with characteristic Y are known, an approximation to the variance $S^2(\bar{Y}_a)$ for the a^{th} size class can be made. The assumption of a constant relvariance is weakest in the largest and in the smallest size classes, and so the constant was doubled for these classes. Values of the constant are also shown in Table 1.

The variance expression does not include the contribution to the variance that arises because most of the sample was restricted to the 98 sample PSUs. The between PSU variance did not need to appear in calculations for optimum sample size since, because the cost function did not account for total number of PSUs, it had been decided to have as many sample PSUs as possible, and so minimize that component of variance resulting between PSUs.

Optimum allocation of facilities selected in the a^{th} size class is that sample size which would produce minimum variance at the fixed cost. The equations involved and methods of solution are outlined in Appendix D. The optimum sample size to be selected from the a^{th} size class is given by:

$$n_a = \frac{N_a S(\bar{Y}_a)}{[\sum_a N_a S(\bar{Y}_a) \sqrt{c_a}\,]} \times \frac{C}{\sqrt{c_a}} \qquad (5)$$

where all quantities are as defined above.

If values of $S(\overline{Y}_a)$ from expression (4) are substituted in equation (5), the variance constant terms cancel out. Optimum sample sizes for size class a then involve only the relative sizes of the variance constants for size class a, the mean number of employees with characteristic Y, the number of establishments, and unit costs.

Since the CBP provided the most precise estimates of the current number of establishments and employees in the target SICs, it was used to determine N_a and $\overline{Y}_a$. In some cases adjustment of $\overline{Y}_a$ was necessary, however. The values of C_a were estimated from NOHS. The size classes used in the CBP records did not permit the size classifications defined earlier, so approximations of the CBP counts for the correct size classes were obtained by using the proportions of the establishments that appeared in those size classes in the DMI file.

Since N_a was based on total numbers of CBP establishments, the values of n_a that define optimum sample size were given in terms of CBP establishments. However, the important result of the optimization computations was to find the optimum sampling rates n_a/N_a for establishments in the a^{th} size class. These rates could then be applied to the DMI file. Parameters used in selecting the sample, and the expected number of CBP establishment selections resulting from the optimum sampling rates are given in Table 1. Numbers of establishments in each size class expected from both the CBP and DMI files are shown in Table 3 in Chapter VI.

The actual samples from the DMI were expected to differ somewhat from the expected sample totals derived from the CBP universe (see Table 3). A number of other factors also affected sample sizes in the DMI. When the computed sampling rates were applied to DMI universe files having duplicate records for establishments or having records for establishments no longer in business, the usual result was a larger sample than expected; however, the telephone screening operation eliminated the out-of-business sample cases. Similarly, when the incomplete file was sampled with these rates, inadequate coverage was reflected by a corresponding shortage in the number of cases selected. The DMI file was expected to have both under-coverage and multiple listing problems.

V. PRIMARY SAMPLING UNITS (PSUs)

Geographical and surveyor workload restrictions resulted in defining 604 Primary Sampling Units (PSUs). PSUs were made up of contiguous counties, parishes in Louisiana, census divisions in Alaska, and in metropolitan areas were composed of counties that made up Standard Metropolitan Statistical Areas (SMSAs). The 604 PSUs were stratified into 98 strata. Of these strata, 26 were grouped into self-representing strata with one large PSU per stratum. The remaining 578 PSUs were grouped into 72 strata, called non-self-representing strata. One PSU from each non-self-representing stratum was selected to represent all other PSUs in the stratum. This selection was done with probability proportional to size. A total of 98 PSUs were selected for analysis in the NOES. Sample establishments with less than 2,500 employees were selected from these 98 PSUs, while establishments with 2,500 or more employees were selected using systematic selection across all 604 PSUs.

A. Definition of Primary Sampling Units

The county was the basic building block for PSUs. This was done to enable the telephone interviewer and the surveyor to use a familiar boundary for a PSU. The DMI file used for sampling establishments also records the county location for establishments as part of the address information.

The system for defining individual PSUs was also heavily influenced by the expected organization of the surveyor staff and the number of surveyors expected to be available for conducting interviews at sample establishments. Originally, a staff of 21 trained surveyors working in five teams was expected to conduct field interviews over a period of two years.

All counties in the 50 States and the District of Columbia were combined into 604 PSUs for this survey. Several conditions were important in defining the PSUs:

1. PSUs as Combinations of Counties

PSUs were made up of contiguous counties. In Louisiana and Alaska, parishes and census divisions, respectively, took the place of counties. Independent cities were combined with neighboring counties.

2. Metropolitan PSUs

PSUs in metropolitan areas were made up of the counties that composed SMSAs at the time of the 1980 census. In some smaller SMSAs, additional non-metropolitan counties were added to provide sufficient interviewing workloads.

3. Non-metropolitan (Non-metro) PSUs

PSUs in non-metro areas were made up of groups of counties that had common boundaries. These counties or groups of contiguous counties were large enough so that a self-weighting sample of the size planned and which would provide a sufficient surveyor workload could be selected.

4. State Boundaries and PSUs

Although the intent was to construct PSUs from counties within the same state, multi-state PSUs did occur either because some SMSAs included areas in more than one state, or because the algorithm used to assign non-metro counties to PSUs occasionally included parts from more than one state.

5. Surveyor Workloads

Each PSU was constructed to provide enough sample establishments to keep a four-person surveyor team busy for a period of two to four weeks.

B. Establishing the Size of the PSUs

The sample was designed to incorporate a self-weighting sample within employee size classes. A self-weighting sample was determined by considering the probability of selecting a specific establishment. The overall probability of selecting an establishment is equal to (the probability of selecting the PSU containing the establishment) times (the probability of selecting the establishment from the selected PSU). For the size class having the lowest sampling rate (i.e., the size class with the greatest numbers of establishments), the self-weighting sample was defined by the following condition:

$$\frac{1}{k} = \left(\frac{M_{hj}}{M_h}\right) \times \left(\frac{M_h}{M_{hj}} \times \frac{1}{k}\right) \tag{6}$$

where

M_{hj} = The total number of establishments for the survey in the j^{th} PSU and h^{th} stratum, i.e., $\sum_a N_{hja} f_a$, the measure of size of the j^{th} PSU in the h^{th} stratum.

M_h = The total number of establishments in the h^{th} stratum, i.e., $\Sigma_j M_{hj}$, the measure of size of all PSUs in the h^{th} stratum.

N_{hja} = The number of establishments in the U.S. in the a^{th} employee size class (according to CBP) in the j^{th} PSU in the h^{th} stratum.

f_a = The oversampling ratio for establishments in the a^{th} size class (see below).

k = Sampling interval, $1/(n_a/N_a)$.

This expression is derived in Appendix E. The term (M_{hj}/M_h) on the right of expression (6) is the probability of selecting the sample PSU from among all PSUs in its stratum. The remaining term on the right defines the probability of selecting sample establishments from the sample PSU. For the a^th^ size class, the following general expression defines the sampling system:

$$\frac{f_a}{k} = \left(\frac{M_{hj}}{M_h}\right) x \left(\frac{M_h}{M_{hj}} \quad x \quad \frac{f_a}{k}\right) \tag{7}$$

where f_a is the oversampling ratio for establishments in the a^th^ size class. The oversampling ratio is the ratio of the largest sampling fraction to the sampling fraction in the a^th^ size class (see Chapter VI).

The terms on the right of expression (7) have the same meaning as in (6); the probability of selection of the PSU is the same but the probability of selection of establishments within the PSU reflects the larger overall sampling fraction f_a/k that applies to the a^th^ class.

These conditions gave rise to two restrictions which can be expressed statistically. The within PSU selection probability given in (7) is the basis of one condition:

1. The probability of selection of establishments within PSUs should not exceed 1; that is:

$$\left(\frac{M_h}{M_{hj}}\right) x \left(\frac{f_a}{k}\right) \leq 1$$

It follows that the PSU measure of size must satisfy:

$$M_{hj} \geq \frac{M_h \ f_a}{k} \tag{8}$$

This restriction was imposed so that a self-weighting sample could be obtained. Writing $\tilde{f}_{ahj}$ as the value of f_a for the largest class with an establishment in the hj^th^ PSU enabled a lower bound to be placed on the measure of size for the hj^th^ PSU:

$$M_{hj} \geq \frac{M_h \tilde{f}_{ahj}}{k} \tag{9}$$

2. At least two team weeks of effort should be required to survey the sample expected from the PSU. Expressed algebraically, this condition becomes:

$$2(139.7) \leq \left(\frac{M_h}{M_{hj}}\right) \sum_{a=1}^{8} \left(\frac{f_a}{k}\right) x \; (N_{ahj}C_a) \tag{10}$$

where C_a is the per firm surveyor hours for the a^th class and 2(139.7) hours of productive surveying per two week period were expected from each four person surveying team. The right side of this expression shows the total surveyor hours in the hj^th PSU as the sum of the products of the number of sample firms in the classes and the per firm survey hours needed. The number of hours of productive surveying per week was derived as follows:

		Hours Per Week
Supervision (40 hours x .2)	=	8.0
Leave (4 persons x 40 hours x 4/52 fraction of weeks in leave status)	=	12.3
Investigation (remaining hours of week)	=	139.7
Total (4 persons x 40 hours)	=	160.0

Condition 2 was used to define an upper limit on the PSU measure of size as:

$$M_{hj} \leq \left(\frac{M_h}{2(139.7)k}\right) \sum_{a=1}^{8} f_a N_{ahj} C_a \tag{11}$$

Although conditions 1 and 2 could be stated explicitly, it was not always practical to adhere to them rigidly. For example, the PSU measure of size M_{hj} for some PSUs could be made large enough to satisfy condition 1 only by defining PSUs covering excessively large areas, and, some PSUs had to be defined with measures that did not meet this condition. This problem occurred for employee size classes 3 through 8 in these PSUs and was dealt with by assigning special weights in the estimation procedure (see Chapter VIII).

C. Location and Stratification of PSUs

The grouping of U.S. counties into 604 PSUs for NOES was done in a series of manual and computer assisted steps following the conditions specified in Section A and conditions 1 and 2 of Section B.

All counties, parishes, and independent cities within the United States were listed in a contiguous sequence. The list was prepared by manually assigning sequence numbers to the counties on a series of maps. The ordering tried to minimize "cross-overs" from one side of a significant geographical feature to the other and from a state to its neighbors. Particular care was taken to minimize cross-overs from one Census Region to another (i.e., Northeast, North Central, South and West, as designated by the hundreds digit of the PSU number, 2, 4, 6, 8 respectively as shown in Appendix B).

SMSAs were defined as PSUs. In a few instances, one or more adjacent non-metro counties were added to smaller SMSAs to obtain minimum PSU sizes. This was done in a way to minimize 'cross-overs'

Although each of the very large SMSAs was treated as a single PSU, field interviewing was occasionally apportioned to more than one team to be done at different times. For example, one-half of the Chicago SMSA was surveyed by all of the interviewers available at the start of the survey and the remaining portion of the Chicago SMSA was interviewed later as a separate assignment.

Non-metro counties were combined into PSUs following the two conditions for size discussed in Section B. This step was done using a computer. The computer results were visually inspected to look for awkward geographical combinations that would make them inappropriate assignment areas. A few PSUs of very large area were generated in the Western states. Counties in these states were resequenced and a revised set of PSUs were generated. Later, when one of these large PSUs (in Alaska) was identified as a sample PSU, a subsample of the PSU area was selected to permit manageable travel patterns.

Stratification of the PSUs was imposed so that data from the many exposure groups included in the Survey could be handled easily. PSUs with similar characteristics such as number of employees or proportions of employees in certain industries were grouped and treated as a unit in the process of stratification. The 604 PSUs defined in the NOES were grouped into 98 strata. Selection of establishments was then done from 98 PSUs within the 98 strata, rather than from all 604 PSUs. Stratification also reduces the variance between PSUs within each stratum. The efficiency of a stratified design as measured by the variance is improved by defining strata of approximately equal size such that the PSUs within the strata are as homogeneous as possible with respect to the important statistics to be estimated from the survey. Homogeneity of PSUs within strata can sometimes be improved by using groups of PSUs with similar economic structure as strata.

The requirement that PSUs should provide an interviewing assignment of two to four weeks for a four person surveyor team was an important consideration in determining the size of the strata. The number of sample establishments and the average survey cost per establishment shown in Table 1 in Chapter IV indicate that the expected number of surveyor hours for establishments with 2,500 or more employees should have been about 35 percent of the total surveyor workload. As these large establishments were to be surveyed without regard to their location, they did not influence the number of sample PSUs. Strata sizes were therefore based on apportioning the remaining 65 percent of the 75,200 hours (about 48,400 person hours) required to survey establishments with less than 2,500 employees.

Given 139.7 hours of productive surveying per team, per week, the total number of team weeks for surveying establishments of less than 2,500 employees approximates:

48,400/139.7 = 345 team weeks

Assuming two to four weeks of surveying time per PSU, the average workload over all PSUs should be three team weeks. Then an approximate duration (in terms of hours spent surveying) for each stratum would be 3 weeks out of 345 (or about 1 in 115) of the total survey workload for establishments of less than 2,500 employees.

The disparity in size of the PSUs interfered with establishing strata of equal sizes because some PSUs were larger than the desired average stratum size. The largest of these PSUs were defined as separate strata (self-representing strata) and the remaining PSUs were grouped into strata of approximately equal size.

Sample establishments with less than 2,500 employees were to be designated from PSUs within each strata. Since very large establishments with 2,500 or more employees were to be selected without regard to their PSU location, that did not influence the stratification process.

PSUs in the strata should also be relatively homogeneous with respect to statistics of interest for the survey. Groups of PSUs with significant concentrations of employees in certain key target industries that were likely to have serious and common health hazards were identified. This worked fairly well for most of the small PSUs. However, for PSUs which contained a wide range of target industries, it was not always possible to produce strata that were homogeneous in this regard. This was particularly true in the larger employee size classes. Additional stratification criteria, in addition to employee concentration by SIC, were therefore used. The computer was used to group PSUs and display the distribution of PSUs with respect to the following variables:

1. Proportion of employees in establishments working in manufacturing SICs.

2. Proportion of employees in establishments within the PSU falling in the largest size classes.

3. Concentration of employees in the petroleum and/or chemical, rubber, leather industries.

4. Geography - Census region.

5. SMSA or Non-SMSA.

This listing also reflects the order of importance of each variable in the formation of strata. As a first step, large groups were formed comprising PSUs that were similar with respect to numbers of employees in manufacturing and large establishments. If possible, employees in industries thought to have high potential exposures, e.g., petroleum, chemical, rubber, and leather industries, were also similarly concentrated. The measure of size in each large group determined the number of strata that should be produced from the group. If two or more strata were to be constructed, the PSUs were sorted by the five variables in the order listed above and then divided into strata, based upon total measure of size and the similarity across PSUs for each variable above.

The process produced a total of 98 strata. Twenty-six of these strata contained only one large PSU; these strata are called self-representing (SR) because the single PSU represents itself in the sample. The remaining 578 PSUs (604 minus 26) were grouped into 72 non-self-representing (NSR) strata having about equal measures of size; the term NSR was applied to these strata because one PSU was selected to represent all other PSUs in its stratum. The final groupings of PSUs into strata were done by the contractor, and are not available.

D. Selection of Sample PSUs

Once the strata were defined, all PSUs were listed by stratum showing M_{hj}, the PSU measure of size. Prior to sampling, the strata were compared to locate pairs of strata that were composed of roughly similar PSUs. The pairing of strata was significant, since the computation of variances described in Chapter VIII employed a paired stratum method.

One PSU was selected at random from each stratum with the probability of selection for each PSU proportional to the measure of size contributed by that PSU. The composition of the 98 PSUs selected for the NOES is shown in Appendix B. Parts of 40 States and the District of Columbia appear among the sample PSUs.