SUPPORTING STATEMENT – PART B
U.S. Department of Commerce
U.S. Census Bureau
Survey of Industrial Research and Development
(Forms RD-1 and RD-1A)
OMB Control No. 0607-0912

B. Collection of Information Employing Statistical Methods

1. Description of Universe and Respondent Selection

Approximately 32,000 companies will be selected annually during the clearance period to represent the approximately 2 million companies with 5 or more paid employees in all private non-farm sectors of the economy. The sampling frame of 2 million companies will be created by aggregating establishment information from the Census Bureau's Business Register (BR) into a file of enterprises.

Based on the response rates for the 2005 cycle of the Survey and as summarized in the table below, the expected response rate for respondents that receive the Form RD-1 (3,426 companies) will be approximately 79 percent. The response rate for respondents that receive the Form RD-1A (28,574 companies) is expected to be approximately 78 percent, yielding an overall response rate of 78%. Expected response rates are summarized in the following chart:

Category	Number of companies in the universe	Number of companies in the sample	Expected response rate in percent
All companies	2,060,803	32,000	78
Top 300 R&D performing companies[1]	300	300 (298 RD-1 2 RD-1A)	90
Remaining Form RD-1 companies	3,128	3,128	77
Form RD-1A companies	2,057,375	28,572	78

2. Procedures for Collection of Information

Statistical Methodology for Stratification and Sample Selection

The sample design for the Survey is complex. The design takes advantage of several pieces of information concerning the conduct of R&D. They are (1) current year information from the Business Register, (2) information about R&D reporting from the Survey over the past 5 years, (3) information from the Bureau of Economic Analysis about R&D

[1] The response rate for the largest 300 industrial R&D performers for the 2003, 2004, and 2005 cycles of the Survey was 91%, 88%, and 90%, respectively.

performance, (4) information from various trade associations about R&D performance, and (5) information from the Company Organization Survey (COS) about R&D performance. It is estimated that about 2 percent of all U.S. companies perform R&D so additional information about R&D reporting is used to make the sample more efficient.

Partitioning of the frame – The sampling frame of approximately two million companies is first partitioned into two groups. Group 1 consists of companies that have responded to the Survey at least once in the past five years, or companies that have responded to the COS. That is, Group 1 consists of companies where the status (yes/no) of R&D is known from past surveys. Group 2 consists of the remaining companies; Group 2 consists of companies where the status of R&D is unknown. See table 1 for details.

Partitioning the groups into broad categories **–** The two groups are next partitioned into categories. Group 1 is partitioned into three categories: (1) companies who performed $3 million or more R&D in the past year, (2) companies who reported positive R&D at least once in the past five years (but less than $3 million in the past year), and (3) remaining companies who reported $0 R&D at least once in the past five years. Group 2 is partitioned into two categories: (1) the top 50 largest companies (based on payroll) in a state or industry and (2) all remaining companies.

Table 1 - 2006 Groups and Categories

Group/Category	**Population Size**
Group 1 – Known R&D status	*99,194*
G1C1 - ≥ $3 million (most recent positive R&D)	3,610
G1C2 - < $3 million (most recent positive R&D)	11,366
G1C3 - $0 R&D (2001 – 2005) (certainties) (noncertainties)	84,218 566 83,652
Group 2 – Unknown R&D status	*1,750,174*
G2C1 - Top 50 by state or industry	4,259
G2C2 - remaining	1,745,915
TOTAL	*1,849,368*

Stratifying the categories – One of the major goals of the survey is to provide estimates of R&D performance at the industry level. To provide enough sample by industry group, the categories are stratified. Group 1 category 1 (G1C1) is a certainty stratum. Group 1 category 2 (G1C2) is stratified into 50 industry groups. Group 1 category 3 (G1C3) is stratified into two strata: certainties and noncertainties. Group 2 category 1 (G2C1) is a certainty stratum. Group 2 category 2 (G2C2) is stratified into the same 50 industry groups as G1C2. Table 2 summarizes these categories.

Table 2 - 2006 Strata

Categories	**Strata Definitions**
Group 1 – Known R&D status	
G1C1 - ≥ $3 million (most recent positive R&D)	Certainty strata – take all
G1C2 - < $3 million (most recent positive R&D)	Stratify into 50 industry groups
G1C3 - $0 R&D (2001 – 2005)	Stratify into 2 broad strata – certainties (companies with active establishments in NAICS 5417 or from BEA file) and noncertainties
Group 2 – Unknown R&D status	
G2C1 - Top 50 by state or industry	Certainty strata – take all
G2C2 - remaining	Stratify into 50 industry groups

Industry classification - Each company in the frame is assigned one single 6-digit North American Industry Classification System (NAICS) code regardless of the number of business activities the company conducts. The NAICS code is assigned using a hierarchical 4-step procedure.

Step 1 - Determine the company’s economic sector (2-digit NAICS or combination of 2-digit NAICS) that accounts for the highest percentage of its aggregated payroll.

Step 2 - Determine the company’s economic subsector (3-digit NAICS) that accounts for the highest percentage of its payroll within the assigned economic sector.

Step 3 - Determine the company’s 4-digit industry code that accounts for the highest percentage of its payroll within the assigned economic subsector.

Step 4 - Determine the company’s 6-digit industry code that accounts for the highest percentage of its payroll within the assigned 4-digit industry code.

The industry stratification is based on the 4-digit NAICS code. In 2006, the 50 industry groupings consisted of 28 manufacturing groups, such as food, chemicals, computers or aerospace products and 22 nonmanufacturing groups, such as trade, utilities, or professional services. For a complete list of the 2006 industry groups, see table 3 below.

Table 3 – 2006 Industry Groups

Survey Industry Group	NAICS
01	311
02	312
03	313-316
04	321
05	322-323
06	324
07	3251
08	3252
09	3254
10	other 325
11	326
12	327
13	331
14	332
15	333
16	3341
17	3342
18	3344
19	3345
20	other 334
21	335
22	3361-3363
23	3364
24	other 336
25	337
26	3391
27	other 339
28	unclassified manufacturing, i.e. NAICS = 3100
30	21
31	22
32	23
33	42
34	44, 45
35	48, 49
36	5111
37	5112
38	5171, 5172
39	5174
40	other 517
41	5181
42	5182
43	other 51
44	52, 53
45	5413
46	5415
47	5417
48	other 54

49	621-623
50	55, 56, 61, 624, 71, 72, 81
00	unclassified, i.e. NAICS = 0000

Sampling methodology – To be as efficient as possible in the sampling, companies with positive known R&D are over sampled at sufficiently higher sampling rates to provide enough sample to produce industry level estimates. In addition, companies within certain states are over sampled at higher sampling rates to provide enough sample to produce state level estimates. This strategy resulted in enough sample to produce state by industry level estimates.

The sample selection methodology differed in the different categories.

(1) In the three certainty strata (G1C1, G2C1 and part of G1C3), all companies are selected for the sample. These companies are important to the sample because they are either known to conduct a large amount of R&D or are known to have large payrolls in a state or industry group and may be more likely to conduct R&D or have R&D labs.

(2) Probability proportionate to size (PPS) sampling is used to select companies within the 50 industry strata in G1C2 and G2C2. In G1C2, size is based on prior year reported R&D. In G2C2, size is based on payroll. The probabilities of selection are determined such that a company that is large relative to other companies in a given state or in a given industry has a higher probability of selection than a ‘smaller’ company.

(3) Simple random sampling (SRS) is used to select companies in the noncertainty portion of G1C3. It is possible, but not highly likely, that these companies are conducting R&D in the current year. Companies in the manufacturing stratum are selected with the same probability as those in nonmanufacturing. This sampling probability was 0.01. Table 4 summarizes the various sampling methodologies for each category.

Table 4 – Sampling Methodologies

Categories	**Strata Definitions**	**Sampling methodology**
Group 1 – Known R&D status		
G1C1 - ≥ $3 million	Certainty strata	Take all
G1C2 - < $3 million	Stratify into 50 industry groups	PPS with state & industry constraints
G1C3 - $0	Stratify into 2 broad strata – certainties and noncertainties	Take all for certainty portion and SRS for noncertainty portion
Group 2 – Unknown R&D status		
G2C1 – Top 50 by state or industry	Certainty strata	Take all
G2C2 – Remaining	Stratify into 50 industry groupings	PPS with state & industry constraints

Sample size – The overall sample size of roughly 32,000 is based on a combination of desired degree of precision at the industry or state level and on staffing and budget resources. Some initial overall constraints were set. (1) A minimum probability of selection was set at .01 for companies in manufacturing industries so that the maximum weight of any of these companies would be 100. (2) A minimum probability of selection was set at .004 for companies in nonmanufacturing industries so that the maximum weight of any of these companies would be 250. (3) Roughly ¾ of the sample was to be selected from the known positive R&D performers. (4) The sampling fraction for the noncertainty portion of the known zero group was set at 1 in 100. All of these constraints, along with the specified CV constraints, were set so as to meet the desired total sample size of roughly 32,000 companies. The final 2006 sample sizes by strata are shown in table 5 below.

Table 5 – 2006 Industry R&D sample sizes

Categories	**Population size** **N**	**Sample size** **n**
Group 1 – Known R&D status	*99,194*	*12,656*
G1C1 - ≥ $3 million	3,610	3,610
G1C2 - < $3 million	11,366	7,643
G1C3 - $0		
certainties	566	566
noncertainties	83,652	837
Group 2 – Unknown R&D status	*1,750,174*	*19,428*
G2C1 – Top 50 by state or industry	4,259	4,259
G2C2 – Remaining	1,745,915	15,169
Total	*1,849,368*	*32,084*

Estimation

Roughly 54 detailed statistical tables are produced each year from the Survey, including point estimates and coefficients of variation. For a majority of the estimates, the Horvitz-Thompson (H-T) estimator and variance is computed.

The H-T estimator can be expressed as[2]:

$$\hat{Y}_{HT} = \sum_{i}^{n} \frac{y_i}{\pi_i} = \sum_{i}^{n} w_i y_i$$

where

y_i is the measurement for the *i*th unit

π_i is the probability that the *i*th unit is in the sample > 0 (*i* = 1, 2, … N)

$w_i = \frac{1}{\pi_i}$ is the weight associated with the *i*th unit

The variance of the H-T estimator is given by:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1 - \pi_i)}{\pi_i} y_{,}^{2} + 2\sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j$$

where

π_{ij} is the probability that the *i*th and *j*th units are both in the sample

This is the true variance if the entire population is known.

The H-T estimator is design-unbiased and preserves desired additive properties within and across published tables. That is, the sum of the estimated R&D across all industries or across all states adds to the estimated U.S. total.

For state level estimates, presented in only three tables, a modified synthetic estimator is used[3]. This estimator preserves the desired additive properties yet provides smoother estimates over time for rare event populations, such as R&D by state.

[2] *Sampling Techniques 3rd Edition*, William G. Cochran, John Wiley & Sons, 1977.

[3] A Hybrid Estimation Approach to State Level Estimates in the Survey of Industrial Research & Development, Slanta and Mulrow, presented at the 2004 Joint Statistical Meetings in Toronto.

The new estimator used to produce state estimates from the R&D survey has the following form:

$$\hat{Y}_S = \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk} y_{Shk} + \sum_{h=1}^{L}\sum_{k=1}^{N_h} a_{hk}\left(\sum_{I=1}^{N_I} R_{IS}\left(w_{hk} - 1\right) y_{Ihk}\right)$$

where

$$R_{IS} = \frac{\sum_{k=1}^{N}\left(1 - \pi_k\right) X_{ISk}}{\sum_{k=1}^{N}\left(1 - \pi_k\right) X_{Ik}}$$

and

N = population size
N_h = population size of stratum h
N_I = number of independent non-aggregate industry publication tabulations
L = Number of sampling strata

$$N = \sum_{h=1}^{L} N_h$$

y_{Shk} = reported or imputed R&D in state S of k^{th} company in stratum h
y_{Ihk} = reported or imputed R&D in industry I of k^{th} company in stratum h
w_{hk} = weight of k^{th} company in stratum h, = reciprocal of probability of selection
a_{hk} = one (1) if k^{th} sampling unit in stratum h is selected and zero (0) otherwise
X_{ISk} = payroll in industry I and state S of k^{th} company, available from the frame
X_{Ii} = payroll in industry I of k^{th} company, available from the frame
π_k = probability of selection of k^{th} company

Payroll by industry and state is first obtained at the establishment level then rolled up to a company level. It should be noted that a company can have payroll in more than one industry or state. The numerator of R_{IS} is the expected value of the payroll of any given state within a given industry from companies that are not selected. The denominator of R_{IS} is the expected value of the payroll of a given industry from companies that are not selected. Companies selected with certainty do not figure in the calculation of R_{IS}.

The estimator itself can be decomposed into two major parts. The first part is the unweighted sum of the reported or imputed R&D in the state of interest. This value is the lower bound of all possible values of the true value given the selected sample. The second part is the portion of the difference between the weighted and unweighted R&D that is allocated to the state.

To obtain the variance of the modified synthetic estimator, the sample variance for the H-T estimator can be modified by replacing y_{Shk} with $\tilde{y}_{Shk}$, where

$$\tilde{y}_{Shk} = \frac{\left(y_{Shk} + \sum_{I=1}^{N_I} R_{IS} \left(w_{hk} - 1 \right) y_{Ihk} \right)}{w_{hk}} .$$

And the modified state estimator can be re-expressed as

$$\hat{Y}_S = \sum_{h=1}^{L} \sum_{k=1}^{N_h} a_{hk} w_{hk} \frac{\left(y_{Shk} + \left[\sum_{I=1}^{N_I} R_{IS} \left(w_{hk} - 1 \right) y_{Ihk} \right] \right)}{w_{hk}} = \sum_{h=1}^{L} \sum_{k=1}^{N_h} a_{hk} w_{hk} \tilde{y}_{Shk} .$$

Degree of Accuracy

The design coefficients of variation used to determine sample sizes by strata vary from 0.35% on overall totals to 6.8% for industry level estimates. Achieved relative standard errors (RSE) for the 2005 estimates differed from the pre-specified design levels. The RSEs on overall totals ranged from 0.7% on budgeted R&D to 11.6% on the total number of nanotechnology companies. The RSEs on subtotals at the manufacturing or nonmanufacturing level ranged from 0.3% on total manufacturing foreign R&D to 19.7% on the total number of nanotechnology companies. RSEs vary greatly at the individual industry and state level with the majority under 3.0% for manufacturing industries and under 7.0% for nonmanufacturing industries.

3. **Methods to Maximize Response and Account for Nonresponse**

Follow-up procedures - Form RD-1 companies will continue to have 60 days to report. Reminder letters will be sent to companies that have not responded by mid-May, unless they have been granted extensions. Follow-up letters will be sent in late May, June, and July. The first follow-up package sent in late May will include a duplicate form. In addition, Census Bureau staff will telephone companies among the largest 500 R&D performers that have not returned a survey form or requested a filing time extension rather than send the second and third notice. These companies account for as much as 85 percent of the value of the data and their responses are critical for the completeness of the estimates. Form

RD-1A companies that do not respond within 30 days are sent follow-up letters in April, May, June, and a phone follow-up in July (or until a response is received). Each mail follow-up package includes a duplicate form.[4]

Estimating for missing data - Estimates for Form RD-1 nonrespondent companies are made for total R&D spending based on the company's previous years' data and the change from the prior to the current year for responding companies in the same industry. The distributions of expenditures for nonrespondent companies or for partial respondent companies are based on the distribution for the company in the prior year. If the company has no previous distribution of expenditures, an attempt is made to impute data for only selected items based on the distribution of data for companies in the same industry. For Form RD-1A companies that were not selected in the prior year sample, total R&D spending is imputed based on the average expenditure in the current year for responding companies. R&D expenditures for selected data items are imputed for nonrespondent or partial respondent companies. This imputation is based on the average distribution of expenditures for responding companies.

Survey form redesign for improved response – Both survey questionnaires and the accompanying instructions were previously redesigned as a result of the conversion to the Census Bureau's Generalized Instrument Design System (GIDS), extensive review by NSF and Census Bureau staff, cognitive testing, and review by a noted survey methodologist who specializes in survey questionnaire design[5] (see Attachments 2 and 3 for copies of the current survey instruments and Attachments 4 and 5 for the proposed survey instruments for 2007). NSF and the Census Bureau will continue redesign efforts, which will include extensive testing and evaluation of proposed changes, in an effort to make response to the survey less burdensome. During the next clearance period, we plan to continue sending the Form RD-1 to companies that perform $3 million or more of industrial R&D.[6] The Form RD-1 will continue to include questions that probe company ownership, sales, total employment, R&D employment, R&D expenditures, R&D costs budgeted for the next year, R&D performed by outside organizations, R&D performed abroad, R&D funded by Federal government agencies, the types of costs incurred for R&D, R&D costs distributed by state, energy-related R&D the cost of fringe benefits for R&D personnel; the cost of R&D performed in three technology areas (biotechnology, software development, and materials synthesis and processing) and how much of the cost could be attributed to nanotechnology; and the cost of R&D performed by others outside of the company by type of organization.

[5] Professor Donald A. Dillman, University of Washington, is under contract with both NSF and the Census Bureau to give expert advice on survey issues. Dr. Dillman recently used the Form RD-1 as a teaching example in a seminar sponsored by the Census Bureau entitled "Questionnaire Design: Issues for Business Surveys."

[6] All other companies will be sent Form RD-1A, the abbreviated survey form that collects basic information about R&D expenditures and character of work (ie, basic and applied research and development).

4. Tests of Procedures or Methods

NSF and the Census Bureau plan further redesign of the survey questionnaires and instructions, and other methodological research and improvements to the Survey. To research and implement survey improvements, especially those resulting from the CNSTAT recommendations, a significant amount of cognitive and usability testing is planned.

5. Contacts for Statistical Aspects and Data Collection

Persons responsible for sample design and selection:

Paul L. Hsen, Assistant Division Chief
Research and Methodology
Manufacturing and Construction Division
U.S. Census Bureau
(301) 763-4586
(301) 763-7783 (FAX)
paul.l.hsen@census.gov

Stacey J. Cole, Chief
Manufacturing Programs Methodology Branch
Manufacturing and Construction Division
U.S. Census Bureau
(301) 763-4771
(301) 763-4718 (FAX)
stacey.j.cole@census.gov

Jock Black, Senior Mathematical Statistician
Division of Science Resources Statistics
National Science Foundation
(703) 292-7802
(703) 292-9092 (FAX)
jblack@nsf.gov

Person responsible for data collection:

Julius Smith, Jr., Chief
Special Studies Branch
Manufacturing and Construction Division
U.S. Census Bureau
(301) 763-7662
(301) 763-4718 (FAX)
julius.smith.jr@census.gov

Person responsible for analysis of the statistics and publication:

Raymond M. Wolfe, Economist
Research and Development Statistics Program
Division of Science Resources Statistics
National Science Foundation
(703) 292-7789
(703) 292-9091 (FAX)
rwolfe@nsf.gov

Attachments

1. Cover Letter

2. Survey Questionnaire (Draft 2007 Form RD-1)

2A. Draft Instructions for the 2007 Form RD-1

3. Survey Questionnaire (Draft 2007 Form RD-1A)

3A. Draft Instructions for the 2007 Form RD-1A

[4] In the past all respondents were asked to return the survey form within 60 days. Beginning with the 1998 survey, respondents sent Form RD-1A are asked to return the form within 30 days. Since nearly all companies sent Form RD-1A do not conduct R&D, and since the Census Bureau has made it much easier for them to report through the introduction of TDE, this has improved response without adding burden.