



## Current and Alternative Sources of Data on the Science and Engineering Workforce

Working Paper | SRS 07-202 | June 2007

by Ronald S. Fecso (formerly Division of Science Resources Statistics, National Science Foundation), G. Hussain Choudhry, Graham Kalton, Adam Chu, and Richard Phelps (Westat, Inc.)

### Disclaimer

Working papers are intended to report exploratory results of research and analysis undertaken by the Division of Science Resources Statistics. Any opinions, findings, conclusions or recommendations expressed in this working paper do not necessarily reflect the views of the National Science Foundation. This working paper has been released to inform interested parties of ongoing research or activities and to encourage further discussion of the topic.

### Introduction

Knowledge about the size and characteristics of the science and engineering (S&E) workforce is extremely important in understanding and assessing the place of the United States in this increasingly technological world. For more than 50 years, the National Science Foundation (NSF) has attempted to meet the need for current information on this important segment of our workforce. Although representing only about 4.6% of the civilian noninstitutionalized population in 1997, or about 12.5 million people, the impact of this group on our society far exceeds its number because of the contributions that scientists and engineers make to technological innovation and economic growth. It is no exaggeration to anticipate that scientists and engineers will prove critical to the continued social and economic development of the nation as it confronts the challenges of the future.

The most comprehensive current source of data on the S&E workforce in the United States is the integrated Scientists and Engineers Statistical Data System (SESTAT), a revised system that NSF introduced in 1993. SESTAT draws upon three surveys, with data collection carried out in 1993 and then at 2-year intervals for the rest of the decade. The three surveys are as follows:

- *National Survey of College Graduates (NSCG)*. In 1993, the NSCG sample was drawn from all people age 75 or younger who had a bachelor's or higher degree at the time of the 1990 decennial census. Certain occupations and groups were oversampled because a main purpose of the survey was to screen for individuals age 75 or younger with a bachelor's degree or higher in an S&E field at the time of the 1990 decennial census (whether employed in S&E or not in 1993), and those with a bachelor's degree or higher in non-S&E disciplines in 1990 who were employed in an S&E occupation in 1993. These two groups comprise what is termed the *NSCG S&E panel* that was followed in 1995 and later rounds.

- *National Survey of Recent College Graduates (NSRCG)*. The NSRCG covers those individuals who received an S&E bachelor's or master's degree from a U.S. educational institution within the academic years 1 or 2 years prior to the survey reference date. A sample from each cohort of the NSRCG is resurveyed every 2 years thereafter, using the same questionnaire as the NSCG.
- *Survey of Doctorate Recipients (SDR)*. The SDR covers individuals age 75 years or younger who received a doctorate degree in an S&E field from a U.S. educational institution since June 1942. The SDR sample of S&E doctorate recipients is subsampled over time, with samples of new S&E doctorate recipients added every 2 years. The subsampling rates for earlier doctorate recipients and the sampling rates for recent doctorate recipients are chosen to maintain a sufficient total sample for analysis purposes. The panel design produces longitudinal data that may be used to analyze the professional careers of S&E doctorate recipients.

Thus, the SESTAT target population includes residents of the United States who were noninstitutionalized and age 75 or younger with at least a bachelor's degree who were either educated as or working as a scientist or engineer.

Although SESTAT provides substantial employment-related data for those in the target population, users note some limitations in the data. The current definition of the SESTAT target population is restricted to those with at least a bachelor's degree. Thus, individuals with a science or engineering occupation but who do not have a bachelor's or higher degree (in any field) are excluded from the target population.

In addition, the current NSF system has some problems covering the S&E target population. Although those residing in the United States at the time of the 1990 decennial census with only non-U.S. degrees are covered by SESTAT through the NSCG, those with degrees who currently reside in the United States but came here after the census, and who did not receive a degree from a U.S. institution since 1990, are not covered. Also, individuals holding non-S&E degrees in 1990 and working in non-S&E occupations in 1993 who changed to S&E occupations after that time are not covered. Furthermore, individuals working in an S&E occupation who received non-S&E degrees since 1990 are not covered.

This report addresses issues of target population and coverage and examines the feasibility and desirability of other survey approaches for obtaining the required data with the desired sample sizes. Other existing federal statistical data collection efforts are considered for use as possible sources of samples for the SESTAT surveys. Thus, the report considers alternative methods for addressing two issues: (1) defining the target population to meet user data needs, and (2) achieving good coverage of the target population.

As noted above, being a small proportion of the total population, scientists and engineers are a "rare" population, i.e., they are difficult to study in a cost-effective manner using general population survey techniques. For example, to locate individuals in the general population who are in scope for the S&E population, and to oversample particular subgroups of the S&E population, would require a large volume of household screenings. It is essential, therefore, to find a cost-effective approach to screen the general population. In this context, it is natural to look at a variety of alternative sample designs, as well as to examine existing large-scale, ongoing governmental household surveys that might lend themselves effectively to such a screening requirement. To that end, this report explores and examines the

feasibility of using several federal data-gathering efforts for screening purposes, including the planned American Community Survey, the Current Population Survey, the National Health Interview Survey, the National Immunization Survey, and other selected large household or population surveys. Another very different approach is to sample members of the S&E population at their places of work, using an establishment-based sample design. This report also explores this option as a source of data on the S&E workforce.

## Background

---

SESTAT is a database created to enable study of scientists and engineers in the United States. The SESTAT database for the 1990s was designed in response to recommendations of a panel of the National Research Council's Committee on National Statistics. The panel's analyses and recommendations for the data system were presented in the 1989 report *Surveying the Nation's Scientists and Engineers: A Data System for the 1990s*. The report includes the following statement from the panel:

We strongly urge that the NSF personnel data system for the 1990s strive to provide information that will permit users to apply their own definitions of the science and engineering population to suit their particular research and analysis purposes within a framework that facilitates cross-comparison with other widely used data sources. Specifically, we believe that the system should support analysis of the science and engineering community from each of the two major perspectives... from the perspective of occupational employment or jobs and from the perspective of academic training or careers. (Citro and Kalton 1989:55–56)

The SESTAT target population includes individuals who, as of the survey reference period, had the following characteristics:

- at least a bachelor's degree
- either a degree in S&E or working in an S&E occupation
- in the civilian noninstitutionalized population
- 75 years of age or younger

The degree fields considered to be S&E include computer and mathematical sciences, life sciences, physical sciences, social sciences (including psychology), and engineering. Occupational categories considered to be S&E include computer and mathematical scientists, life scientists, physical scientists, social scientists (including psychologists), and engineers.

The SESTAT database is created by integrating the NSCG, the NSRCG, and the SDR. The 1990 decennial census long form sample was used as a screening device for obtaining a sample of scientists and engineers for the NSCG. The NSCG was first administered in 1993 to a nationally representative sample of all college degree holders who were identified through the 1990 decennial census. Because information on degree field was not available on the census long form, the sample included individuals in the United States with a bachelor's degree or higher in any field, not just in science or engineering, as of April 1990. The sample members of the 1993 NSCG who had S&E degrees in April 1990 and/or who were working in S&E occupations in 1993 constituted the NSCG S&E panel, which was followed in subsequent rounds of the NSCG in 1995, 1997, and 1999.

The NSRCG has been administered biennially since the early 1970s to recent U.S. bachelor's and master's degree recipients in S&E disciplines. The NSRCG employs a two-stage sample design. First, a sample of institutions that grant S&E degrees was selected and asked to provide lists of their relevant graduates in the 2 academic years before the survey (except in the case of the 1993

NSRCG, which covered graduates in spring 1990 as well as in the 1991 and 1992 academic years). Second, samples of graduates with bachelor's and master's degrees in S&E fields were selected from the lists for inclusion in the NSRCG. The samples of graduates selected in each round of the NSRCG are eligible to be sampled and then become part of the NSCG in the next round of SESTAT.

The SDR has been sponsored by NSF with some financial contributions from other federal agencies since the early 1970s. This survey follows a sample of holders of S&E doctorates earned at U.S. institutions throughout their careers from year of doctorate degree award through age 75. Every 2 years, a sample of new S&E doctorate degree earners is added to the SDR from another NSF-sponsored survey, the Survey of Earned Doctorates (SED), which is an annual census of all recipients of research doctorates from U.S. institutions. The overall sample size is maintained by subsampling of the older cohorts to make room for the sample of new graduates.

The current SESTAT target population excludes individuals who do not hold bachelor's or higher degrees but are currently working in science or engineering fields. There is, however, a growing interest in such individuals, particularly in fields such as information technology. It may therefore be desirable to expand the target population to include individuals without bachelor's degrees who are working in at least some science or engineering fields.

In addition to possible extension of the definition of the SESTAT population, a second issue relates to coverage gaps with the current definition. The SESTAT system misses some individuals with S&E degrees and many individuals with non-S&E degrees who are working in S&E occupations at the time of a given SESTAT round. For example, the system does not include foreign citizens who received only non-U.S. degrees and entered the country after the 1990 decennial census, because an adequate sampling frame for such individuals has not been identified. Also, U.S. citizens who did not have at least a bachelor's degree at the time of the 1990 decennial census, but who subsequently obtained an S&E degree from abroad, are missed. In addition, those individuals with only non-science or non-engineering degrees at the time of the census who were working in non-science or non-engineering occupations or not working at the time of the 1993 NSCG, but who subsequently entered a science or engineering occupation, are not covered in SESTAT. Furthermore, SESTAT does not cover graduates receiving their first bachelor's degrees in non-S&E fields after April 1990 who entered S&E occupations.

A third issue is the timeliness of the data, or the elapsed time between the reference date in each survey and the date on which survey data are released. In recent years, this time period has been reduced by about 10%. Still, some SESTAT data users feel that the time period remains too long. Timeliness is not addressed directly here, however improvements in frames and/or interagency cooperation may lead to further improvements in timeliness.

Many of the users of S&E personnel data focus their attention on unique subgroups of the S&E population, e.g., the employed or the unemployed, or women and minorities in specific fields. Because the SESTAT surveys are required to produce estimates for various subgroups of interest with specified precision levels, this interest implies certain subgroups have to be oversampled to meet the specified reliability criteria. An indication of the sizes of various

subgroups of interest is provided by the following brief description of selected characteristics of the S&E workforce.

The total number of employed scientists and engineers in the United States in 1997 was 10.6 million according to the SESTAT integrated database.[1] Of these, the vast majority (10.1 million) held at least one 4-year degree in a science or engineering field. Only about 30% (3.1 million) of the 10.1 million S&E degree holders in the workforce were employed in S&E occupations. Almost 57% of the individuals employed in S&E jobs reported their highest degree type as a bachelor's degree, whereas 29% listed a master's degree and 14% a doctorate.

The private for-profit sector is by far the largest employer of individuals who are members of the SESTAT population. In 1997, 73% of scientists and engineers in the workforce whose highest degree was a bachelor's degree and 60% of those whose highest degree was a master's degree were employed in a private, for-profit company. It should be noted that they were not necessarily employed as scientists or engineers. The academic sector was the largest sector of employment for those with doctorates (49%).

Although women made up close to half (46%) of the U.S. labor force in 1997, they accounted for only slightly more than one-fifth (23%) of the S&E labor force. With the exception of Asians and Pacific Islanders, minorities compose a much smaller proportion of scientists and engineers in the United States than they do of the total U.S. population. Asians and Pacific Islanders were 10% of scientists and engineers in the United States in 1997, although they were only 4% of the total U.S. population. Blacks comprised 12%, Hispanics 11%, and American Indians and Alaskan Natives 1% of the U.S. population in 1997, whereas Blacks and Hispanics each comprised only about 3%, and American Indians and Alaskan Natives about 0.5%, of scientists and engineers.

A draft version of the material presented in the main body of this report was a key topic of discussion at an expert panel meeting about SESTAT sampling design, which was held at NSF on 5 December 2000. The agenda of the expert panel meeting appears in appendix A, and a list of meeting participants is provided in appendix B. Appendix C contains a brief summary of the meeting and the recommendations made by the panel. Appendix D lists desired and acceptable coefficients of variation (CVs) for S&E workforce estimates. Appendix E contains a discussion of the characteristics of some establishment data collections.

---

## Footnotes

[1] Note that measurement of the number of scientists and engineers is dependent on the definition used. At the time of this redesign research, 1997 data were the latest available. More recent data are now available at the NSF/SRS website, <http://www.nsf.gov/statistics/>. Various definitions of the number of scientists and engineers may be found there as well.

## **Current Design of the SESTAT Surveys**

---

As described in the previous sections, the SESTAT data are collected from three surveys that have been conducted every 2 years since 1993. A brief description of the three SESTAT component surveys follows. See the SESTAT website <http://sestat.nsf.gov/> for further details.

### **National Survey of College Graduates**

The NSCG is a panel survey that started with a sample of individuals with at least a bachelor's degree at the time of the 1990 census and then added samples of recent S&E graduates at subsequent SESTAT rounds. The 1993 NSCG, which primarily covered the experienced S&E population, was conducted by the U.S. Census Bureau, using a subsample of the 1990 decennial census long form sample. A sample of eligible respondents from the 1993 NSCG was followed in subsequent SESTAT rounds and supplemented by samples from the NSRCG.

The 1993 NSCG was a special baseline survey of a sample of all those who had earned a bachelor's degree or higher (in any field) before 1 April 1990, the date of the 1990 decennial census, and were age 72 or younger at that time. The sample design was a two-phase stratified random sample of individuals with at least a bachelor's degree. Phase 1 consisted of the procedure used by the Census Bureau for sampling households for the census long form. That procedure was a stratified systematic sample, with differing sampling rates for administrative areas of different sizes. Phase 2 consisted of subsampling individuals with at least a bachelor's degree and age 72 or younger from the long form records, within strata defined according to demographic characteristics (race/ethnicity, citizenship, and disability status), highest degree achieved, occupation, and sex. Within each stratum, individuals were selected using probability-proportional-to-size (PPS) systematic sampling. The long form sampling weight was used as the size measure for selection to compensate as much as possible for the differing long form sampling rates, and hence to come as close as possible to an overall self-weighting sample within each phase 2 stratum. The maximum sampling rate was 3.00%, but most strata were sampled at rates of between 2.03% and 2.82%. The unweighted response rate for the 1993 NSCG was 78%, yielding more than 148,000 individuals who had at least a bachelor's degree, and identifying an additional 19,000 people not eligible to receive the survey (e.g., those who did not have a bachelor's degree, were deceased, were over 75, or were no longer living in the United States). Survey responses were then used to determine whether the respondents fit into SESTAT's target population of scientists and engineers by virtue of having an S&E degree at the time of the census and/or working in an S&E occupation at the time of the 1993 NSCG. More than 74,000 survey respondents matched the SESTAT definition.

The 1995 NSCG sample was selected from 1993 NSCG eligible respondents and the 1993 NSRCG respondents (described in the next section), using a unique identification rule to avoid two chances of selection for individuals eligible for both surveys. The 1995 NSCG frame was stratified by factors such as highest S&E degree level, highest S&E major study field, demographic group, and sex. A sample of 62,004 individuals was selected for the survey using PPS sampling within these strata, with the 1993 analysis weight being used as the size measure for PPS sampling. Sampled individuals were contacted initially by mail. A total of 41,522 eligible sample members responded to the

mail component of the survey. Nonrespondents were then subsampled for computer-assisted telephone interview (CATI) or computer-assisted personal interview (CAPI) follow-up. Across all data collection modes, a total of 53,448 eligible scientists and engineers responded to the 1995 NSCG. The conditional unweighted response rate (conditional on having responded in 1993) was 95%. The unconditional response rate (taking into account nonresponse in 1993) was approximately 74%.

The 1997 NSCG was selected from eligible respondents to the 1995 NSCG (itself derived from 1993 NSCG and 1993 NSRCG respondents) and augmented by a sample of the 1995 NSRCG respondents. The 1995 NSRCG respondents were oversampled to support more detailed analyses of recent S&E college graduates, and a few extra questions were added to the NSCG questionnaire for these individuals. The cases originally sampled in the 1993 NSRCG and 1995 NSRCG were referred to in 1997 and subsequent years as the 1993 NSRCG panel and the 1995 NSRCG panel.[2] The unweighted conditional response rate for the 1997 NSCG was 94% and the unconditional response rate is estimated to be about 70%.

The 1999 NSCG sample was drawn from a frame consisting of eligible respondents from the 1997 NSCG (itself including 1993 NSCG and 1993 and 1995 NSRCG panels) and the 1997 NSRCG. The conditional unweighted response rates were 90% for the 1993 NSCG and 1993 NSRCG panel components combined, and 81% for the 1995 and 1997 NSRCG panel components combined. The unconditional response rate was in the region of 60%.

### **National Survey of Recent College Graduates**

During the 1990s, the NSRCG covered those individuals who received an S&E degree from a U.S. educational institution in the 2 academic years before the SESTAT survey reference dates. Specifically, the 1993 NSRCG covered individuals who received bachelor's or master's degrees in an S&E field from a U.S. educational institution between 1 April 1990 and 30 June 1992. The 1995 NSRCG covered those who received bachelor's or master's degrees in an S&E field from a U.S. educational institution in the period from 1 July 1992 to 30 June 1994; the 1997 NSRCG covered the period from 1 July 1994 to 30 June 1996; and the 1999 NSRCG covered the period from 1 July 1996 to 30 June 1998.

A two-stage sample design was used in each round of the NSRCG. Educational institutions were sampled at the first stage, and S&E bachelor's degree and master's degree graduates were sampled from within these institutions at the second stage. The Integrated Postsecondary Education Data System (IPEDS) was used to construct the sampling frame for educational institutions. For the 1993 NSRCG, 196 of the eligible institutions had such large numbers of S&E graduates that they were selected with certainty. From the remaining institutions, 79 were selected using systematic, PPS sampling from a file sorted by ethnicity, region, public/private status, and presence of agricultural courses. The measure of size was devised to account for the rareness of certain fields of study and for the incidence of Hispanic, black, and noncitizen students. Of the 275 sampled institutions, 273 provided lists of their students receiving bachelor's or master's degrees in S&E fields between 1 April 1990 and 30 June 1992. From the 273 responding institutions, 25,785 students were selected

using stratified systematic sampling, with markedly varying sampling rates by stratum field. Of the 25,785 selected students, a total of 19,426 eligible degree recipients responded to the 1993 NSRCG. The unconditional unweighted response rate over both stages of sampling was 85%.

The 1995 design for the NSRCG was similar to the 1993 design. In the two-stage sampling approach, educational institutions were again sampled in the first stage using PPS sampling. In 1995, a composite measure of size was introduced that was designed to facilitate oversampling of rare domains of interest (e.g., minority graduates). The 1991–92 IPEDS was used to construct the sampling frame for institutions. There were 102 institutions that were so large that they were selected with certainty. In addition, 173 institutions were sampled from the remaining portion of the frame after stratifying by region, control (public versus private), and percentage of S&E degrees. From the 266 responding institutions, 23,771 students were selected using stratified systematic sampling. Initial nonrespondents and those who had to be traced were subsampled for further follow-up, thus reducing the sample size to 21,000 graduates. A total of 16,338 eligible degree recipients responded to the 1995 NSRCG. The unconditional unweighted response rate was about 83%.

The 1997 NSRCG retained the same sample of institutions that was selected for the 1995 cycle, that is, the same 102 certainty selections and 173 noncertainty selections. All of the 275 sampled institutions responded. A total of 14,057 graduates were sampled. Of these, 1,032 were ineligible, and 10,452 eligible respondents completed the survey. The unconditional unweighted response rate was 82%.

The 1999 NSRCG retained the 275 institutions sampled in 1995 and surveyed again in 1997. Also, four additional institutions were included in the sample, all selected with certainty, to improve population coverage. Of the 279 sampled institutions, 1 turned out to be ineligible and 1 did not provide a graduate list. A total of 13,918 graduates were sampled, of whom 987 were ineligible and 9,984 were eligible and completed the survey. The unconditional unweighted response rate was 78%.

### **Survey of Doctorate Recipients**

The SDR covers individuals who received a doctorate degree in an S&E field from a U.S. educational institution after 1 January 1942. (Non-U.S. doctorates in S&E are covered by NSCG if the recipient entered the United States by 1 April 1990.) The sampling frame for SDR was constructed from the Doctorate Records File, which is a database of all U.S. research doctorate recipients since 1920. The 1993, 1995, 1997, and 1999 rounds of SDR covered those who received degrees up to 30 June 1992, 30 June 1994, 30 June 1996, and 30 June 1998, respectively.

The SDR is a panel survey of doctorate recipients. Samples of new cohorts are added to the base sample every 2 years, and some cuts are made to maintain overall sample size. The SDR frame is restricted to two groups of recipients of U.S. S&E doctorates under age 76: (1) U.S. citizens, and (2) non-U.S. citizens who plan to remain in the United States after receiving their doctoral degrees. A two-phase sample design was used in 1991 and 1993 for the SDR to facilitate oversampling of the disabled and certain minority groups and to facilitate overall sample size reduction for the survey.

The overall sampling rate for the 1993 SDR was 8.8%, with rates for individual sampling strata ranging from 4.5% to 66.7%. The sample size was 49,228 doctorate recipients, of whom 39,495 were eligible S&E doctorate respondents. The conditional unweighted response rate was 87%.

For the 1995 SDR, a sample of those earning doctoral degrees at U.S. institutions between 1 July 1992, and 30 June 1994 (the new cohort) was added, and the previous sample of doctorate recipients whose degrees were received between 1 January 1942, and 30 June 1992 (the old cohort) was subsampled to produce a combined sample of about the same size as the 1993 sample. The sampling rates for the new and old cohorts were similar within strata defined by demographic group, field of study, and sex. An initial sample of 49,829 cases was selected. Of these, 31,243 responded by mail. Nonrespondents were subsampled for CATI follow-up, again using stratified PPS sampling procedures. Across all modes of data collection, 35,370 eligible doctorate recipients responded to the survey. The conditional unweighted response rate for the 1995 SDR was 77%. The corresponding weighted rate, allowing for the nonrespondent subsampling for CATI follow-up, was 85%.

Following the same procedures, the 1997 SDR sample was composed of those earning doctoral degrees at U.S. institutions between 1 July 1994 and 30 June 1996, and a subsample of the previous sample of doctorate recipients (degrees received between 1 January 1942 and 30 June 1994). The new cohort cases were sampled at about twice the rate of the old cohort cases, but the proportion of cases within strata (defined by demographic group, field of study, and sex) was similar for the old and new cohort. An initial sample of 54,103 degree recipients was selected, 28,886 of whom responded by mail. Of these cases, 27,382 were deemed complete interviews, with the remainder being either permanently or temporarily out of scope. Nonrespondents were subsampled for CATI follow-up, again using stratified PPS sampling procedures. Across all modes of data collection, 35,667 eligible doctorate recipients responded to the survey. The conditional unweighted response rate for the 1997 SDR was 85%.

Like the previous rounds of the survey, the 1999 SDR added a new cohort. In this case the new cohort comprised those who earned S&E doctorate degrees between 1 July 1996 and 30 June 1998. The new cohort was oversampled such that out of a total sample size of 40,000 doctorates, 4,000 were allocated to the new cohort. The sample from the 1997 SDR was divided into two subgroups, termed the *old cohort* and the *nearly new cohort*, corresponding to those who earned their doctorates before 1 July 1992 and those who earned their doctorates between 1 July 1992 and 30 June 1996. The nearly new cohort was then sampled at a somewhat higher rate than the old cohort for the 1999 SDR. Of the initial sample of 40,000 individuals, 27,269 responded by mail, and of those 26,216 were deemed complete responses; the remainder were out of scope. Mail nonrespondents were followed up by CATI, which yielded 5,102 complete interviews. Thus overall 31,318 eligible doctorate recipients responded. The conditional unweighted response rate was 82%.

### **Weighting and Estimation**

Selection probabilities for the SESTAT surveys vary greatly. The sampling weight for each sampled individual is defined as the reciprocal of that individual's probability of selection. The sampling weights are then adjusted for nonresponse and poststratification using weighting class adjustment procedures. The final adjusted sampling weights become the analysis weights, which have

been added to each individual's record in the survey database.

In the 1993 NSCG, poststratification adjustment was used to adjust the weighted counts for survey respondents to the 1990 decennial census long form sample estimates. In the 1993 NSRCG and the 1993 SDR, the weights were adjusted only for nonresponse. Similarly, for the 1995, 1997, and 1999 surveys, weights were adjusted for nonresponse with no poststratification adjustment.

The SESTAT database was constructed for each survey round by combining the three component surveys, which meant addressing the potential for cross-survey multiplicity. Scientists and engineers in SESTAT could belong to the surveyed population of more than one component survey, depending on their degrees and when they received them. For instance, a person with a bachelor's degree at the time of the 1990 decennial census who went on to complete a master's degree in 1991 could be selected in the 1993 NSCG and the 1993 NSRCG. The following unique-linkage rule was devised to remove these multiple selection opportunities: each member of SESTAT's target population is uniquely linked to one and only one component survey, and that individual is included in the integrated SESTAT database only when he or she is selected for the linked survey. As a result, each person has only one chance of being selected into the combined SESTAT database. Individuals with multiple selection opportunities were first linked to the SDR, and then to the NSRCG if the individual was not linked to the SDR. In the NSCG, sampled individuals who also had a chance of being selected for the NSRCG or the SDR in that year were assigned zero as their SESTAT analysis weight. Similarly, sampled individuals in the NSRCG who also had a chance of being selected for the SDR in that year were assigned zero as their SESTAT analysis weight. The component survey's analysis weight for all other cases was used to develop the SESTAT analysis weight. Cases with a zero weight are not eligible to be sampled in future waves of the longitudinal samples.

### **Sampling Frame Issues**

As can be seen from the foregoing, the current NSF approach is complex and has sample selection and coverage problems. The two most significant problems are as follows:

- Responses to the census long form questionnaire are not a particularly efficient means of identifying those with S&E degrees. The long form only asks about the level of the degree attained, not degree field. Hence, it does not provide the means to identify just those with S&E degrees.
- Using a decennial census to identify the stock of engineers and scientists to be interviewed over the decade, together with new graduates in S&E fields, means that some population groups are missed. For example, apart from individuals with bachelor's or master's degrees at the time of the census who are working in S&E jobs at the time of the first NSCG in the decade, those with non-S&E degrees who move into S&E jobs during the decade are not covered in any of the surveys. The computer science field, for example, includes a significant number of workers who were not trained in a science or engineering discipline. Another population group of interest that the current approach misses is scientists and engineers trained outside the United States, who entered the United States after the decennial census and receive no further degrees in the United States.

It is important to recognize that a substantial number of members of the S&E population are not S&E graduates from U.S. educational institutions. Data from the 1993 SESTAT database substantiate this point. The data showed that there were 593,600 individuals in S&E occupations with non-S&E degrees only; this number excludes those who graduated in non-S&E fields between 1 April 1990 and 1993 who were working in S&E occupations in 1993. Additionally, there were approximately 428,000 individuals in the SESTAT database who had only foreign degrees. There was some overlap between these two populations, so in 1993 a total of approximately 1,020,000 individuals in the SESTAT database either (a) had an S&E occupation but no S&E degree, or (b) had only foreign degrees. These individuals comprised approximately 9% of the 1993 SESTAT universe of 11.6 million individuals.

NSF seeks to explore alternative survey approaches for developing a cost-effective S&E personnel data system that provides a more complete representation of the universe of scientists and engineers than the current approach. If possible, the approach should provide means to include individuals who are working in S&E occupations but do not hold a bachelor's degree or higher in an S&E field. Some alternative approaches are discussed in the next section, as is the possibility of using establishment survey frames for this purpose.

### **Sample Design Issues**

The type of sample design chosen should derive from user data needs and the types of analysis to be conducted. One important issue is whether analysts are interested in longitudinal analyses of data from these surveys. There would appear to be great potential for such analyses, but there has been limited use of the longitudinal data to date. The current design provides longitudinal data as a byproduct of sample generation, which is not the case with the designs based on the alternative surveys discussed in this report, such as the American Community Survey and the National Immunization Survey. Another consideration that would affect the choice of sample design is survey frequency. Currently, the SESTAT surveys are conducted biennially. The assumption for this review is that this will continue to be the case, although the Division of Science Resources Statistics (SRS) has considered other cycle lengths.

We consider four types of sample design and briefly discuss the data analysis implications and the response burden issues associated with each.

- *Panel Design.* Respondents are interviewed for several waves over the life of the panel. Panel designs provide a very rich source of data for longitudinal analyses, specifically providing insight into how changes occur over time. Tracking the individual respondents over time carries implications for survey cost. The respondent sample size tends to decrease if some sampled individuals cannot be tracked. Moreover, the survey response rates start to decline because of response burden as the panel ages.
- *Rotating Panel Design.* The sample is divided into a number of rotation groups (or panels), and one of the panels is rotated out at each survey period. The main reason for choosing a rotating panel design is to reduce response burden and limit nonresponse. A rotating panel design implies tracking the individuals in the overlapping rotation groups over time.

- *Repeated Cross-Sectional Sample Design.* Independent cross-sectional samples are selected for each survey period. For this design, there is no need to track individuals over time, and response burden is minimized. The survey data can produce cross-sectional estimates but not longitudinal analyses.
- *Split Panel Design.* This design is a compromise between a panel design, as above, and a cross-sectional design with a new sample of the population. At each survey round in a split panel design, one part of the sample is a fresh selection, and the other part is a panel component.

---

### **Footnotes**

[2] Although conceptually the 1995 NSRCG panel is a component of the NSCG, it has been conducted as part of the NSRCG data collection and not as part of the NSCG data collection.

## **Alternative Approaches**

---

As has been noted, the S&E population is rare with respect to a general population sampling frame and there is no separate, complete, and up-to-date frame of the S&E population for use throughout the decade. Lacking a separate frame, a large-scale screening exercise is needed to identify a sample of the S&E population. In this situation, it is natural to seek another large-scale data collection to provide the screening. Currently, the decennial census long form serves this purpose, but it is not an ideal vehicle. Because the long form does not include field of degree, further screening is needed to distinguish individuals with S&E degrees from non-S&E degree holders. Also, since the census is conducted only every 10 years, supplementary sampling is needed to update it. This is the role of the NSRCG and SED but, as noted earlier, the update does not include new non-S&E degree earners working in S&E jobs and new immigrants with only foreign degrees.

This section examines the possible use of some large-scale ongoing federal surveys to provide the screening phase of surveys of the S&E population. In considering alternative surveys for this purpose, many factors need to be addressed including the coverage of the target S&E population they provide, the ability to screen out nonmembers of the S&E population, the timeliness and frequency of the surveys, the response rates to the surveys and the likely response rates to the follow-up surveys of the S&E population, issues of operational feasibility, and cost implications. However, as indicated below, an overriding factor that rules out most potential surveys is that their sample sizes are too small for screening purposes; the screening sample size must be large enough to generate the sample sizes needed for the S&E population and for specified subpopulations.

The required sample sizes are, of course, a direct function of the reliability standards established for key survey estimates for the total S&E population and for subpopulations. Appendix D gives the desired and acceptable coefficients of variation (CVs) for estimates of the sizes of various S&E subpopulations of interest that were developed by NSF (e.g., the desired CV for the total S&E population is 0.10% but 0.20% is acceptable). The discussion below explores the screening sample sizes needed to achieve those CVs.

### **Using a General Household Survey to Screen for a Population of Scientists and Engineers**

Variance formulas for simple random sampling are employed to provide rough guidance on the size of screener sample needed to satisfy the reliability specifications given in appendix D; in practice, these formulas will likely underestimate the actual variances (i.e., the design effects will exceed 1). Design effects will exceed 1 as a result of unequal selection probabilities, clustering, and weighting adjustments for nonresponse. Also, no allowance is made for the sample loss from nonresponse. Thus, the screener sample sizes presented are underestimates of those actually required.

To achieve the desired CV of 0.10% for the total S&E population, a screening sample of the general population would need to be large enough to yield a sample of about 893,000 scientists and engineers. If the frame consists of about 101.1 million households, approximately 7.25 million households would have to be screened to obtain a sample of 893,000 members of an S&E population of about 12.5 million people. Moreover, this very large sample size is calculated

without any allowance for a design effect of greater than 1 or for survey nonresponse.[3] This size of sample, already larger than operationally feasible, still would not satisfy the CV requirements for some rare subgroups of the S&E population, such as race/ethnicity groups. These rare population subgroups would have to be oversampled substantially to achieve the required levels of CVs for them, which would further increase the screening cost. To obtain the acceptable CV of 0.20% for the estimate of the size of the total S&E population, more than 1.8 million households would need to be screened, again assuming no clustering, equal selection probabilities, and complete response.

We consider the following three options of screener sample size in terms of number of households for a general household population survey:

- *Option A.* Screener sample of 1.0 million households
- *Option B.* Screener sample of 1.5 million households
- *Option C.* Screener sample of 2.0 million households

Tables 1–3 give the approximate CVs of the survey estimates for various sized subgroups of the S&E population for the above three options assuming equal probabilities of selection, no clustering of households, and a total of 101.1 million households in the United States. The achieved CVs actually would be higher, depending on the design effect.

TABLE 1. Option A: Screener sample size of 1.0 million households

| S&E subpopulation (millions) | S&E subpopulation members in sample | Coefficient of variation (%) |
|------------------------------|-------------------------------------|------------------------------|
| 12.50                        | 123,621                             | 0.28                         |
| 10.00                        | 98,897                              | 0.31                         |
| 7.50                         | 74,173                              | 0.36                         |
| 5.00                         | 49,448                              | 0.44                         |
| 4.00                         | 39,559                              | 0.50                         |
| 3.00                         | 29,669                              | 0.57                         |
| 2.00                         | 19,779                              | 0.70                         |
| 1.00                         | 9,890                               | 1.00                         |
| 0.75                         | 7,417                               | 1.15                         |
| 0.50                         | 4,945                               | 1.41                         |
| 0.25                         | 2,472                               | 2.00                         |

S&E = science and engineering.

TABLE 2. Option B: Screener sample size of 1.5 million households

| S&E subpopulation (millions) | S&E subpopulation members in sample | Coefficient of variation (%) |
|------------------------------|-------------------------------------|------------------------------|
| 12.50                        | 185,432                             | 0.23                         |
| 10.00                        | 148,346                             | 0.25                         |
| 7.50                         | 111,259                             | 0.29                         |
| 5.00                         | 74,173                              | 0.36                         |
| 4.00                         | 59,338                              | 0.40                         |
| 3.00                         | 44,504                              | 0.47                         |
| 2.00                         | 29,669                              | 0.57                         |
| 1.00                         | 14,835                              | 0.81                         |
| 0.75                         | 11,126                              | 0.94                         |
| 0.50                         | 7,417                               | 1.15                         |
| 0.25                         | 3,709                               | 1.63                         |

S&E = science and engineering.

TABLE 3. Option C: Screener sample size of 2.0 million households

| S&E subpopulation (millions) | S&E subpopulation members in sample | Coefficient of variation (%) |
|------------------------------|-------------------------------------|------------------------------|
| 12.50                        | 247,242                             | 0.19                         |
| 10.00                        | 197,794                             | 0.22                         |
| 7.50                         | 148,345                             | 0.25                         |
| 5.00                         | 98,897                              | 0.31                         |
| 4.00                         | 79,117                              | 0.35                         |
| 3.00                         | 59,338                              | 0.40                         |
| 2.00                         | 39,559                              | 0.50                         |
| 1.00                         | 19,779                              | 0.70                         |
| 0.75                         | 14,835                              | 0.81                         |
| 0.50                         | 9,890                               | 0.99                         |
| 0.25                         | 4,945                               | 1.41                         |

S&E = science and engineering.

The CVs for selected subgroups of the S&E population known to be of interest to user groups are presented in table 4. These CVs are computed under the three options, and assume no oversampling and no clustering effect.

TABLE 4. Achieved coefficients of variation for various S&amp;E population subgroups under three options

| S&E population subgroup       | Coefficient of variation (%) |          |          |
|-------------------------------|------------------------------|----------|----------|
|                               | Option A                     | Option B | Option C |
| Female                        | 0.59                         | 0.48     | 0.41     |
| Hispanic                      | 1.63                         | 1.33     | 1.15     |
| Black                         | 1.63                         | 1.33     | 1.15     |
| Asian/Pacific Islander        | 0.89                         | 0.73     | 0.63     |
| American Indian/Alaska Native | 4.00                         | 3.26     | 2.82     |
| Unemployed                    | 2.25                         | 1.84     | 1.59     |

S&E = science and engineering.

As table 4 shows, none of the options can satisfy the precision requirement for estimating the number of female scientists and engineers given in appendix D. The acceptable CV is 0.3%, whereas even option C, with a screener sample size of 2.0 million households, achieves a CV only as low as 0.4%. The screener sample size would have to be increased to about 3.6 million households to achieve the acceptable CV. Similarly, the required (or acceptable) CVs for estimating the numbers of Hispanic and black scientists and engineers cannot be achieved even under option C. The screener sample size would have to be increased to about 2.6 million households to achieve the required CV of 1.0% for Hispanics and blacks. On the other hand, the required CV of 1.0% for the estimate of the size of the Asian S&E population can be met even under option A, using a screener sample size of 1.0 million households. It should be noted that the above CVs are obtained under the assumption that the survey design effect is equal to 1 (i.e., simple random sampling of households and no clustering of S&E within households). In practice, however, the design effect would be somewhat larger because of clustering of the sample and unequal selection probabilities.

The above discussion indicates that a very large screener sample size is needed to generate sample sizes that meet the CV requirements for some small S&E population subgroups. Note, however, that not all of the rest of the screened sample of the S&E population is needed for the survey. A subsample of sufficient size can be selected to satisfy the reliability requirements for other subgroups and for combinations of them.

With the very large number of households that would need to be screened to

meet the various reliability requirements, conducting a new independent survey of scientists and engineers with its own screening would be extremely costly. For this reason, the approach adopted here is to explore the effectiveness of using an existing large-scale household survey for the screening.

### **Combination of Sampling Frames**

Use of multiple frames is beneficial in sample surveys when the combined frames offer a more efficient means of accomplishing the survey objectives. The current SESTAT is in fact an example of a multiple-frame approach. A multiple-frame approach would also likely be needed with the use of one of the alternative surveys under investigation for screening for the S&E population. These surveys have, in theory, almost complete coverage, and they could be used to screen for a representative sample of the entire S&E population. However, given their current sample sizes, they would yield very few individuals with S&E doctoral degrees; similarly, a general screening of the population for scientists and engineers with doctoral degrees would be very expensive. To address this problem, the SDR could be continued to provide required sample sizes for doctorate recipients, with new cohorts being sampled from the frame provided by the SED as at present. The estimates for the S&E population with doctoral degrees could then be constructed using one of the following options:

- Retain the sample individuals with doctorates from the alternative (complete) sampling frame and combine this sample with the SDR sample to produce composite estimates.
- Screen out the sample individuals with U.S. doctorates from the alternative (complete) sampling frame and base the estimates solely on the SDR sample.

Similarly, a dual-frame design could be considered for the S&E population with graduate degrees from abroad. For example, one could explore the feasibility of using lists from the Immigration and Naturalization Service (INS) to identify immigrant scientists and engineers who obtained their degrees abroad. Such a design would be difficult to implement, but might be worth considering if the alternative survey frame lacked coverage of this portion of the S&E population.

### **Alternative Screening Surveys**

As the preceding discussion has shown, if another household survey is to be used as a first-phase screening survey for locating members of the S&E population, the survey would need to have a very large sample size. The main advantages of linking the S&E workforce survey with an existing ongoing federal survey would be decreased cost and improved efficiency by consolidating field data collection operations that are common to the different surveys. The remainder of this section examines the possible use of the American Community Survey and other large-scale federal household surveys, including the Current Population Survey, for screening for the S&E population. The possibility of using an existing large-scale telephone survey, such as the National Immunization Survey, to screen for members of the S&E population is also examined. With a household telephone survey, consideration needs to be given to the increasingly serious problem of nonresponse that such surveys often face.

Another very different approach is to sample members of the S&E population at their places of work, using an establishment-based sample design. The use of a list of establishments as the frame to survey employed scientists and engineers is examined in the section "Establishment-Based Sample Design."

### **American Community Survey**

The American Community Survey (ACS) is a rolling sample survey (Kish 1990) that is being developed by the U.S. Census Bureau as an intended replacement for the census long form. The ACS will include the same detailed socioeconomic subject areas as the census long form questionnaire. Instead of collecting data from about 17 million housing units at one time (as is done during the decennial census), the ACS will sample about 250,000 addresses each month, or some 3 million each year throughout the decade (Alexander, Dahl, and Weidman 1997). The ACS is currently in the development and testing stage. The sampling frame will be the Census Bureau's Master Address File, which will be updated throughout the decade to keep it fully current. The sample will be distributed throughout the country with no clustering, and with higher sampling fractions in small governmental units. Each address could be sampled at most once in a 5-year period.

The ACS will be conducted using a mail-out, mail-return, self-response approach, combined with initial CATI follow-up, supplemented by a CAPI follow-up of a subsample of the remaining nonrespondents. As an ongoing survey, ACS is a flexible vehicle capable of adapting to changing user needs. Once fully implemented, there is the potential to add supplemental questions on subjects of current interest or to help identify special population groups. It must be noted, however, that it is not necessarily easy to add questions to the ACS and in no case will changes be made before 2008. It is likely that a legal reason will be required for the inclusion of new questions, in addition to several years of lead time for testing.

Once the ACS is fully implemented, estimates will be available annually for areas and population groups of 65,000 or more people. Estimates for smaller areas will be provided on a multiple-year average basis. Estimates for the smallest areas or for small population groups will be available on a 5-year average cycle, with reliability consistent with that provided for these groups in recent decennial censuses.

The ACS could offer a range of sampling options, flexibility in design and content, and more current data for analysis than the census long form. It also could be used to efficiently identify households or individuals with unique characteristics. Once identified, follow-up interviews could be conducted by mail or telephone, or in person, if required, to meet the needs for more detailed information about these households. Furthermore, the ACS could not only provide more timely data for use in designing the S&E workforce surveys, it could also improve survey coverage. If used as a screener for the S&E population, the problems associated with pooling S&E cases across multiple years to achieve sufficient samples of scientists and engineers would need to be addressed.

Because ACS is planned to be a continuous survey, each round of the NSF data collection could be based on a fresh sample selected from the ACS, thus covering the full current S&E population (a repeated cross-sectional survey approach). An approach using recent ACS samples could lead to changes in

content, frequency, or sample design of the NSCG. The ACS might be able to provide the coverage updating function of the NSRCG samples, although the small size of this subpopulation may present problems in having enough cases in the ACS samples. Because the ACS would not provide a large enough sample of doctorate recipients at the small domain levels necessary (e.g., field by race/ethnicity by sex), it would seem desirable to continue with a separate SDR survey. However, the ACS could provide representation of scientists and engineers graduating abroad and non-S&E graduates working in S&E occupations. If the census long form is replaced by the ACS, the use of the ACS as a vehicle for conducting the NSF S&E workforce surveys needs to be explored. Furthermore, NSF needs to determine whether using the ACS could result in significant cost savings and provide improved coverage of the S&E population.

If implemented, the ACS will be an attractive option to consider as a possible venue for collaboration on an existing survey. Given the ACS annual sample size of 3 million housing units and assuming a completion rate of 75%, data will be collected from some 2.25 million housing units, or more than 6 million people, which will yield an annual S&E sample of more than 275,000 individuals.

**Issues to Consider.** Issues include the following:

- Because smaller governmental units are to be oversampled, the final survey weights will vary somewhat. Also, the CAPI sample is a subsample of the previous nonrespondents, which also will lead to variation in the weights. The possibility of counteracting the differential weighting in subsampling should be explored in developing the optimum design for an S&E workforce survey.
- The present ACS questionnaire does not include a question about degree field. Such a question could not be added before 2008, and whether it could be added after then is uncertain. Therefore, the design would require screening from the ACS for the desired S&E population, much like what is done with the census.
- The existing ACS questions meet the current screening needs for an S&E activity to the same extent as does the census long form. Accepting the limitation of lack of degree field, the survey procedures could then follow those used in the NSCG. In this case, there would be no addition to the length of the ACS questionnaire or other impact on the ACS. Thus an S&E effort should not have an adverse effect on ACS response rates.
- The ACS is a mandatory data collection activity, conducted under the Census Bureau's enabling legislation (Title 13, USC). As such, current law precludes the Census Bureau from providing names and/or addresses of respondents to a third party. Discussions should be initiated with the Census Bureau to explore conditions under which any subsequent data collection activity could be carried out that would differ from those currently employed in NSCG's use of the census data.
- Although an S&E activity using the ACS may appear feasible at one point in time, ACS availability for this use throughout the decade is necessary

for the approach to be practical.

- Careful consideration would need to be given to many operational aspects involved in using the ACS to provide screening samples for the S&E workforce surveys. For example, the ACS sample is spread across all the months of each year. If the S&E sample is accumulated for data collection at a single point in time, the problem of tracing movers will arise.
- The costs associated with adopting the ACS for S&E workforce surveys need to be examined to ensure that they can be met.
- It is unlikely that the Census Bureau would consider adding an S&E follow-up survey until the ACS is fully implemented and its operations are well established.

### **National Immunization Survey**

The ongoing National Immunization Survey (NIS) that is being conducted by the National Center for Health Statistics (NCHS) and the National Immunization Program of the Centers for Disease Control and Prevention could offer an alternative opportunity for screening for a sample of the S&E population. The NIS is a large-scale random digit dialing (RDD) survey that uses CATI methods to screen more than 900,000 households each year. The survey was initiated in April, 1994 to monitor vaccination coverage levels of children 19–35 months of age on an ongoing basis. NIS covers all 50 states and the District of Columbia. The sample is allocated to produce national estimates and separate estimates for 78 Immunization Action Plan (IAP) areas.

The use of a list-assisted RDD approach for sampling households in the NIS results in noncoverage of households without telephones and households with unlisted numbers in the 100 banks containing no listed numbers. Although approximately 5% of the households in the United States do not have a telephone, this is probably a minor concern for households containing individuals in the S&E population.[4] Noncoverage as a result of residential unlisted telephone numbers belonging to the 100 banks containing no listed numbers is less than 2%, which is negligible.

For purposes of identifying the eligible S&E population, it would be necessary to modify the NIS screening questionnaire by adding special questions. If the NIS sample is not large enough to meet all the sample size requirements for the S&E data system, the set of special questions could also be made into a separate instrument to use in screening an additional sample of households for members of the S&E population. The cost of administering such a screener, which might be substantial, is of real concern and would need to be explored completely and carefully. The main advantage of an independent screener is that the screening instrument would be developed to suit the needs of S&E data user requirements. In view of the rarity of those with S&E doctorates, it would almost certainly still be necessary to augment the screened sample with a sample of S&E doctoral degree holders from the SDR.

**Sample Design.** The NIS uses a two-phase sample design. For the first phase, a quarterly sample of telephone numbers is drawn for each IAP area, and a screening questionnaire is administered to locate households with one or more children 19–35 months of age. When an eligible child is found, the person most

knowledgeable about the child's vaccinations is identified. If that person is available, the full interview is administered at that time; otherwise the interviewer arranges a time to call back. Both the screening and the immunization interviews are conducted by CATI.

The NIS seeks to attain a coefficient of variation no larger than 5.0% for the annual vaccination coverage estimates in each IAP area. To satisfy this precision requirement, the sample size target for NIS is set at 440 eligible children per IAP area for each four-quarter period, or 110 per quarter. To achieve a sample of 110 children per quarter for each IAP area, the sampling rates vary across IAP areas according to the sizes of the areas. For example, the larger IAP areas are sampled at lower sampling rates.

The NIS target sample size is 8,580 completed household interviews per quarter for all 78 IAPs (110 per IAP area). For 1994, an average of 512,800 sample telephone numbers were drawn per quarter, and 420,500 sample telephone numbers were actually dialed per quarter by interviewers after the prescreening for business and nonworking numbers had removed 18% of the initial sample. Assuming that approximately 60% of the numbers are residential numbers and there is no nonresponse, the sample yield would be 252,300 residential telephone numbers per quarter. By accumulating the sample over a period of 2 years, the sample size would be more than 2 million households, or 5.4 million people. The expected sample of individuals in scope for the S&E population would be in excess of 240,000. However, because the NIS sample is designed to produce estimates for each of the 78 IAPs, and the IAP areas vary considerably in population size, the NIS file contains highly differential weights that lead to a sizable loss in precision for national estimates. Also, allowance needs to be made for nonresponse to the S&E screener and follow-up data collection.

**Issues to Consider.** Issues include the following:

- Three options present themselves in considering the NIS as a vehicle for an S&E workforce survey.

*Option A.* Use the entire NIS sample over a 2-year period for screening. To evaluate this possibility, the design effect due to variation in the NIS sample weights should be examined.

*Option B.* Subsample from the NIS sample so that the sampling rates are similar for all the areas. The overall sampling rate could be that of the area with the smallest sampling rate (or largest population) or somewhat higher. Whether this approach would yield a sample adequate for producing S&E estimates with the required reliability would need to be examined.

*Option C.* Determine the sample size for the design using uniform sampling rates so that the desired CVs for the S&E estimates can be achieved. Under this scenario, the sampling rates would increase for areas with larger population sizes (i.e., more screening would be needed), and sampling rates would decrease for smaller areas (i.e., a subsample of the NIS sample would be used). The cost and operational implications of this approach are yet to be determined.

- Some households would be eligible for both the NIS and S&E workforce surveys. The increased response burden for such households could adversely affect the response rates. However, because different individuals in the household could be involved for each of the surveys, this may not present a major problem. In any case, research should be undertaken to examine the possible extent of overlap and any effects on NIS response.
- For screening for members of the S&E population, a supplementary telephone sample could be added to the NIS to increase sample sizes for rare domains of analytic interest. More information could be collected in this screener to identify the specified domains for which an extra sample is needed. NCHS currently uses NIS as a vehicle to conduct other health surveys. Would it be operationally feasible to add one more study to the NIS?
- The S&E sample would be built up over a period of time. A careful assessment of the effect of this feature on the survey estimates would need to be made.

### Other Household Surveys

Several other federally sponsored household surveys are conducted on a continuous basis. For example, the Current Population Survey (CPS) has been conducted monthly by the Census Bureau since 1942. Other examples of household survey vehicles that could be explored for studying the S&E workforce are the Census Bureau's Survey of Income and Program Participation (SIPP) and the National Health Interview Survey (NHIS) of NCHS.

Two options would be available for using these surveys to screen for the S&E population: (1) accumulate samples over time and conduct the S&E workforce survey at the required intervals; or (2) collect data on a continuing basis and accumulate data over time to produce estimates at the required intervals. Under option 1, the costs associated with tracing and locating respondents could be very high. On the other hand, the estimates produced under option 2 would not reflect the S&E population for a single point in time, but rather average values over the data collection time period (as is also the case with the ACS).

**Current Population Survey.** The CPS is a monthly survey of about 50,000 households conducted jointly by the Census Bureau and the Bureau of Labor Statistics (BLS). The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutionalized population. Respondents are interviewed to obtain information about the employment status of each member of the household who is 15 years of age and older.

Estimates obtained from the CPS include employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, marital status, and educational attainment. They are also available by occupation, industry, and class of worker.

The CPS employs a stratified multistage clustered sample design. Since the inception of the survey, there have been various changes in the design of the CPS sample. The survey is traditionally redesigned and a new sample selected after each decennial census. The current sample design, introduced in January

1996, includes about 59,000 housing units from 754 sample areas. The number of eligible households is about 50,000, and the number actually interviewed is about 46,800 every month, i.e., about 94% of eligible households respond to the survey. The CPS monthly sample of 46,800 households cannot support the analytic needs of SRS. The CPS sample is designed to support the measurement of the U.S. labor force as a whole rather than that of specialized populations such as the S&E workforce. The CPS sample would need to be aggregated over time to be a possible option for studying the S&E workforce.

The CPS employs a rotating panel design in which only one-eighth of the sample is changed each month. Each monthly sample comprises eight representative subsamples or rotation groups. A given rotation group is interviewed for a total of 8 months, divided into two equal periods. The rotation group is in the sample for 4 consecutive months, leaves the sample during the following 8 months, and then returns for another 4 consecutive months. In each monthly sample, one of the rotation groups is in the first month of enumeration, another rotation group is in the second month, and so on. Because of the rotating panel design of the survey, only about 158,000 unique households are interviewed over a 12-month period. Similarly, over a 24-month period, the accumulated sample size would be 228,000 households, and over a 3-year period, 298,000 households. Even after accumulating the sample over 3 years, the screened sample of individuals who are in scope for the S&E workforce surveys would be approximately 36,000 individuals. The effective sample size would be even smaller for two reasons: design effects attributable to disproportional allocation of the sample, and intracluster correlation, because the unique households that are being sampled are sampled from the same primary sampling units and segments (clustering effect). Thus, the sample sizes would not be large enough to produce S&E estimates with the required reliability.

In response to a legislative mandate under the State Children's Health Insurance Program (SCHIP), the Census Bureau expanded the monthly sample for the CPS in 2000. This expansion was introduced over a 3-month period, beginning with the September 2000 survey, and occurred in 31 states and the District of Columbia. In all, the total number of households eligible for the monthly survey increased from about 50,000 to about 60,000 households.

The SCHIP legislation requires that the Census Bureau improve state estimates of the number of children who live in low-income families and lack health insurance. The expansion of the monthly CPS sample is one part of the Census Bureau's plan for improving the SCHIP estimates. Other parts of the plan include an increase in the number of households that are asked the questions from the annual March supplement to the CPS, the source of information on income and access to health insurance.

The increased sample yields roughly 357,600 interviewed unique households over a 3-year period. Therefore, even the increased CPS sample will not be sufficient for S&E sample size requirements. Moreover, the increase is aimed at improving the state-level estimates. Thus, most of the sample increase is allocated to smaller states, which means larger design effects.

**Survey of Income and Program Participation.** The SIPP, conducted by the Census Bureau, provides information on the economic situation of households and individuals in the United States. SIPP began in late 1983 with a design that

attempted a compromise between the twin goals of collecting accurate cross-sectional and longitudinal data on income and program participation by using a multiple-panel overlapping design. A revised design was introduced in April 1996 to focus primarily on providing accurate and useful longitudinal data by using abutting 4-year panels.

There are three basic elements contained in the overall design of the survey content. The first is a control card, which is used to record basic social and demographic characteristics for each person in the household at the time of the initial interview. The second major element of the survey content is the core portion of the questionnaire. The core questions are repeated at each interview and cover labor force activity, the types and amounts of income received during the 4-month reference period, and participation status in various programs. The third major element is the various supplements or topical modules that are embedded during selected household visits.

The sample for the first SIPP panel in 1983 consisted of about 20,000 households selected to represent the noninstitutionalized population of the United States. The 1996 panel has a sample size of approximately 36,800 households. Households in this SIPP panel were interviewed at 4-month intervals over a period of 4 years. The reference period for the questions is the 4-month period preceding the interview. The sample households within a given panel are divided into four samples of nearly equal size. These subsamples are called rotation groups, and one rotation group is interviewed each month. In general, one cycle of four rotation groups covering the entire sample using the same questionnaire is called a *wave*. The rotation group design was chosen because it provides a steady workload for data collection and processing.

Data collection operations are managed through the Census Bureau's 12 permanent regional offices. A staff of interviewers assigned to SIPP conducts interviews during monthly personal visits, with most interviewing completed during the first 2 weeks of that month. Completed questionnaires are transmitted to the regional offices where they undergo an extensive clerical edit before being entered into the Bureau's SIPP data processing system.

The Census Bureau's current working plan for future panels is shown in table 5. Each of these planned panels will be interviewed every 4 months over 3-year periods. A large panel is started every third year, with smaller panels starting in other years. It should be noted that the Census Bureau had delayed the beginning of the second large panel from 2000 to 2001 because of operational considerations associated with the 2000 decennial census. Moreover, the 2004 (and subsequent) panels will be state-representative, but they will not produce reliable state-level estimates unless some additional sample can be included. The additional sample would not only improve the reliability of the state-level estimates, but it would also compensate for the loss of efficiency for national estimates resulting from differential weights.

TABLE 5. Description of proposed Survey of Income and Program Participation panels

| Panel | Start/end dates   | Households            |
|-------|-------------------|-----------------------|
|       |                   | interviewed in wave 1 |
| 2000  | Feb 2000/Jan 2003 | 11,400                |
| 2001  | Feb 2001/Jan 2004 | 36,700                |
| 2002  | Feb 2002/Jan 2005 | 11,400                |
| 2003  | Feb 2003/Jan 2006 | 11,400                |
| 2004  | Feb 2004/Jan 2007 | 36,700                |

As can be seen from table 5, even after accumulating data over all the proposed panels from 2000 through 2004 the sample size will only be 107,600 households. Because that number is not sufficient for the S&E workforce survey, SIPP is not a viable option for studying the S&E workforce.

**National Health Interview Survey.** The NHIS, which has been in continuous operation since 1957, is designed to produce national and selected subnational estimates of health indicators, health care utilization and access, and health-related behaviors for the U.S. resident civilian noninstitutionalized population. The NHIS is conducted by NCHS, a component of the Centers for Disease Control and Prevention, U.S. Public Health Service, in the Department of Health and Human Services. In accordance with specifications established by NCHS, the Census Bureau participates in the planning of the NHIS and in the data collection. The NHIS sample has been redesigned after each decennial census, and the specific parameters of the design have changed over time. For example, the 1973–84 NHIS design was based on a sample of 386 primary sampling units (PSUs), the 1985–94 NHIS design was based on a sample of 198 PSUs, and the current 1995–2004 NHIS design is based on a sample of 358 sample PSUs. The NHIS sample for the data collection years 1995 to 2004 was designed to improve the precision for various domains defined by race and ethnicity and to enhance the survey's ability to provide state estimates.

The estimates from NHIS can be produced for various population subgroups, including those defined by age, sex, race, family income, geographic region, and place of residence. The number of interviewed households per year is about 41,500. The total number of screened housing units is much larger, about 71,500 per year including nonrespondent and vacant housing units. Compared with other households, the NHIS oversamples Hispanic households at a relative ratio of 2.1:1 and black households at a relative ratio of 1.4:1 to improve the estimates for the Hispanic and black populations. The approximately 30,000 housing units screened but not interviewed are white and "other race" housing units, as well as vacant and nonrespondent units.

The NHIS is based on a stratified multistage area sampling design with clustering of housing units. The PSU is a metropolitan statistical area or a group of one or more counties. The PSUs are stratified by region and state, with some oversampling occurring in small states. An area segment sample of housing units is selected at the second stage of selection. Most housing units constructed since 1990 are separately sampled from new construction permit records.

The NHIS sample is randomly partitioned into four nationally representative subsamples of approximately equal size and conceptually similar statistical features. These partitions are referred to as "panels," each of which contains about 104 PSUs. The largest self-representing PSUs are included in all panels, and no non-self-representing (NSR) PSU is included in more than one panel. The sample is also divided into a number of temporal subdesigns. The sample is first divided into subdesigns that are assigned for data collection for each year in the period from 1995 to 2004. Annual NHIS samples are then divided into 52 weekly interviewer assignment samples, with each weekly sample constituting a national probability sample of housing units. An average NSR PSU has five weekly assignments during a year. Large self-representing PSUs have assignments in many more weeks per year. NCHS processes groups of 13 weekly samples that correspond to quarters of the calendar year to produce

national estimates for each quarter.

Given the very small NHIS sample size, which in fact would have to be accumulated for more than 10 years to obtain an adequate S&E sample, the option of using the NHIS to screen for the S&E population is not feasible.

### **Establishment-Based Sample Design**

Another approach is to sample the S&E population by place of work, using an establishment-based sample design. As part of the ongoing SRS program, one survey is already being conducted using lists of companies as a sampling frame. The annual survey of Research and Development in Industry collects information on R&D expenditures and employment of scientists and engineers from a nationally representative sample of about 23,000 companies. The survey was started in 1992 and includes data from both manufacturing and nonmanufacturing companies. The frame for the annual Survey of Research and Development in Industry is a potential establishment-based frame for finding members of the S&E workforce.

Alternatively, the sampling frame for the S&E data collection could be constructed using establishment-level primary and secondary Standard Industrial Classification (SIC) codes to identify establishments likely to employ members of the S&E population. A sample of employees who are in scope for the S&E workforce could then be selected from the sampled establishments. There are several business frames and databases that could be used as frame sources. Some business frames, such as Dun's Market Identifiers (DMI) of the Dun and Bradstreet Information Services, are fairly comprehensive. The DMI contains more than 10 million records on establishments small and large. It is an establishment-based frame but also has data on corporate structure. The DMI database was used as a sampling frame for the 1994 National Employee Health Insurance Survey (NEHS) sponsored by NCHS. According to an assessment study conducted by Marker and Edwards (1997), the database has about 99% coverage of the universe. Coverage of family farms and the self-employed, however, is somewhat weak. Although family farms are not especially important for an S&E workforce study, the absence of the self-employed may prove much more significant. Recently established small businesses also are likely to be missed disproportionately. Coverage of all employees is probably higher than the coverage of establishments, because it is much more likely for large establishments to be included in the frame.

A weakness of the DMI file is its continued inclusion of many small establishments that are no longer in business. This weakness does not cause bias, but would entail increased costs to identify and eliminate sampled establishments that are no longer in operation and that result in increased variances of the survey estimates. However, the reliability of the survey estimates can be improved through poststratification.

The DMI file contains information about business establishments (individual business locations) in the private sector as well as government entities. Individual records in the private-sector portion of the file represent a business establishment, and include basic information such as company name, address, telephone number, and names of corporate officers. Also, the file provides information about total sales and the number of employees. To reduce cost, a full DMI abstract file containing a limited number of design-related variables could be acquired for survey design and sample selection purposes. The full

range of data items could then be obtained for sampled records only. The number of employees in the business establishment is missing for about 13% of the records, but these could be treated as a separate category for sample design purposes. However, such an option would be very costly.

Both the U.S. Census Bureau and BLS maintain business registers for use as sampling frames for their business surveys (see appendix E). Because of confidentiality regulations, these registers are not available to other government agencies. They can be used only if the Census Bureau or BLS conducts the survey and controls the data. Otherwise, the DMI file is the only establishment list that can be used as a sampling frame for the NSF workforce data. The population control totals can, however, be tabulated from the BLS list of establishments for ratio adjustment (poststratification). Wallace et al. (1995) describe how the respondent weights in the private-sector portion of the NEHIS were ratio-adjusted (poststratified) to align with independent estimates of the number of employees provided by BLS. This weight ratio adjustment method reduces the sampling variability in estimates that are correlated with the number of employees, and it also provides a mechanism for adjusting for sampling frame undercoverage. It should be noted that there would still be a risk of potential bias for subgroups for which there is systematic undercoverage on the frame that is not accounted for in the poststratification adjustment.

The DMI file could be used as a sampling frame to study only the employed portion of the S&E workforce population. This frame does not cover those who are unemployed or not in the labor force, and some who are self-employed. These groups are often important subjects for SRS workforce information efforts. A further concern, even for studying the employed, would be whether businesses (or government or academic institutions for that matter) would provide NSF or its contractor with either access to the employees or with names and addresses, which could then be used to contact the employees. Perhaps establishment surveys should be considered as a collection vehicle only for special modules for which the data would be collected directly from the establishments.

---

### Footnotes

[3] With a simple random sample, the estimated number of elements in the population with a given characteristic is given by  $Np$ , where  $N$  is the population size and  $p$  is the sample estimate of the proportion of the population with that characteristic. The CV of this estimated number is approximately  $\sqrt{(1-P)/nP}$ , where  $P = R/N$ ,  $R$  is the number of elements in the population with the characteristic, and  $n$  is the sample size (see, for example, Cochran 1977, section 3.2). These formulas have been used in the calculations in this section under the simplifying assumption that each household contains at most one member of the S&E population. With this assumption, the simple random sampling formulas can be applied with the household as the unit of analysis. The total number of households in the United States has been taken to be  $N = 101.1$  million, and  $R$  denotes the number of members of the S&E population or subpopulation of interest.

[4] However, because only landline phones can be surveyed with RDD methods, the increasing number of households with only cell phones is an emerging problem.

## Summary

---

For more than 50 years, the NSF has attempted to meet the need for current information on the important S&E segment of the population. Although SESTAT provides substantial employment-related data for those with S&E degrees from U.S. educational institutions, some users have noted gaps in the data. For example, some users want more details concerning working environments and career paths. Also, the current definition of the S&E population excludes individuals with degrees below the bachelor's level or who lack degrees but have S&E occupations, but such individuals are of interest to some users. Another limitation of the current NSF system is that it has coverage gaps. For example, most individuals currently living in the United States who received their S&E degrees outside the United States after the 1990 decennial census are not included in SESTAT. Similarly, individuals with non-S&E degrees who were not working in S&E occupations at the time of the 1993 NSCG but who entered S&E occupations after that time are not included.

This report considers the feasibility and desirability of using other ongoing federal surveys, such as the ACS, the NIS, and the CPS as possible alternative sources of a sample of the S&E population. The use of an "establishment" approach is also examined.

On review, based on sample size considerations, the conclusion is that only the ACS and the NIS could be reasonably considered in their current form as alternative screening vehicles for an S&E workforce survey. The sample sizes from all the other federally sponsored household surveys are far too small to satisfy even much less stringent precision requirements than those currently specified for the S&E workforce estimates.

It is possible that the ACS or, less likely, the NIS, might offer the potential for NSF to develop a cost-effective S&E personnel data system that responds more fully to user needs. Even given the large sample sizes of these surveys, however, the minimally acceptable CVs established for the various S&E workforce estimates (appendix D) cannot be met from one year's data. Serious consideration should be given to relaxing the levels of some of the target CVs and to accumulating data over 2 or more years. Accumulation may be of particular relevance for producing estimates of required precision for some "rare" domains. It is important to note that both the ACS and NIS may have limitations on their availability imposed by the sponsoring agency. It should also be noted that the use of one of these surveys as the basis of an S&E workforce survey would result in the loss of the longitudinal analysis capabilities of the current design. There is much further work to be done before establishing that either survey could serve the needs of an S&E data system.

Establishment surveys do not seem to provide a feasible means of reaching the full S&E universe, although an establishment survey approach could be considered for the collection of limited information directly from employers, without requiring contact with the employees. However, it is not clear whether the existing government lists of establishments maintained by the Census Bureau and by the Bureau of Labor Statistics could be utilized easily, if at all. Existing private establishment lists, as noted, contain some limitations, and extensive research and investigation would be required before this approach could be deemed acceptable.

The limitations found in the alternative approaches examined in this report

make their use for the 2003 SESTAT impractical. The sampling frames of the 1990s are the most cost effective and practical approach for the near term. The possible use of the ACS later in the decade after it has become fully operational deserves serious preparatory research in the immediate future.

## Appendix A. Workshop Agenda

---

Workshop on Sampling Frames for the Scientists and Engineers Statistical Data System (SESTAT).

5 December 2000

National Science Foundation, Room 375

- 8:30–9:00 Continental Breakfast
- 9:00–9:15 Introductions  
Moderator: Denise Glover  
Lynda Carlson, Director, Division of Science Resources Studies
- 9:15–9:35 Opening Statements  
Ronald Fecso, Chief Statistician, Division of Science Resources Studies  
Graham Kalton, Senior Vice President, Westat
- 9:35–10:30 NSF Data Collections in S&E Personnel: Historical Overview  
Mary J. Golladay, Director, Human Resources Statistics Program  
The Personnel Data System for the 1990s: SESTAT  
Nirmala (Nimmi) Kannankutty, Human Resources Statistics Program  
Content Revisions and New Populations for the 2000 Decade  
Lawrence Burton, Human Resources Statistics Program
- 10:30–10:45 Break
- 10:45–12:30 Frame by Frame Overview of Other Agency Sampling Frames  
Hussain Choudhry, Senior Statistician, Westat
- How each frame could be relevant to current SESTAT collection
  - How each frame could be relevant to proposed SESTAT collection
  - Comments from agencies sponsoring each frame are encouraged
- 12:30–1:15 Lunch
- 1:15–2:00 Frame by Frame Overview of Other Agency Sampling Frames (cont.)
- 2:00–2:15 Technical Issues Resulting from Integration of Multiple Surveys

Brenda Cox, Mathematica Policy Research

2:15–2:30 Break

2:30–3:15 Discussion 1: Which Proposals Can Work; Which Ones Cannot

Ronald Fecso, Graham Kalton, discussion leaders

3:15–4:00 Discussion 2: Can We Combine Frames? Combine Sampling Efforts?

Ronald Fecso, Graham Kalton, discussion leaders

4:00 Adjourn

## Appendix B. List of Participants

Workshop on Sampling Frames for the Scientists and Engineers Statistical Data System (SESTAT).

5 December 2000

National Science Foundation, Room 375

### National Science Foundation Hosts, Invited Experts, and Speakers

|                               |  |
|-------------------------------|--|
| Charles (Chip) Alexander, Jr. | Census Bureau, Assistant Division Chief, Demographic Statistical Methods Division  |
| Chester E. Bowie              | Census Bureau, Chief, Demographic Surveys Division   |
| Norman Bradburn               | National Science Foundation, Assistant Director, Directorate for Social, Behavioral and Economic Sciences  |
| Lawrence Burton               | National Science Foundation, Senior Analyst, Human Resources Statistics Program, Division of Science Resources Studies, Directorate for Social, Behavioral and Economic Sciences |
| LaTerri D. Bynum              | Census Bureau, Demographic Surveys Division  |
| Lawrence S. Cahoon            | Census Bureau, Demographic Statistical Methods Division  |
| Lynda Carlson                 | National Science Foundation, Director, Division of Science Resources Studies, Directorate for Social, Behavioral and Economic Sciences   |
| Hussain Choudhry              | Westat, Senior Statistician  |
| Steve Cohen                   | Bureau of Labor Statistics, Office of Statistical Methods Research   |
| Brenda Cox                    | Mathematica Policy Research, Inc.  |
| Randy Curtin                  | National Center for Health Statistics  |
| Ron Fecso                     | National Science Foundation, Chief Statistician, Division of Science Resources Studies, Directorate for Social, Behavioral and Economic Sciences                                 |
| John M. Finamore              | Census Bureau, Demographic Statistical Methods Division  |
| Mary J. Golladay              | National Science Foundation, Program Director, Human Resources Statistics Program, Directorate for Social, Behavioral and Economic Sciences                                      |
| Kirk Hagemeyer                | Bureau of Labor Statistics   |
| Michael Hoefler               | Department of Justice, Immigration and Naturalization Service, Statistical Director  |
| Graham Kalton                 | Westat, Senior Vice President  |

|                             |  |
|-----------------------------|--|
| Nirmala (Nimmi) Kannankutty | National Science Foundation, Division of Science Resources Studies, Directorate for Social, Behavioral and Economic Sciences |
| Arthur Kennickell           | Federal Reserve  |
| Nancy Kirkendall            | Department of Energy, Energy Information Agency  |
| Paula Knepper               | National Center for Education Statistics, Postsecondary Studies Division, Senior Technical Advisor                           |
| Enrique Lamas               | Census Bureau, Demographic Surveys Division  |
| Charles P. Pautler, Jr.     | Census Bureau, Chief, Economic Statistical Methods & Programming Division  |
| Susan Schechter             | Office of Management and Budget, Administration Office   |
| George Stamas               | Bureau of Labor Statistics, Employment and Unemployment Statistics   |
| Jean-Louis Tambay           | Statistics Canada  |
| Alan R. Tupek               | Census Bureau, Chief, Demographic Statistical Methods Division   |

### **Invited Observers**

|                  |   |
|------------------|---|
| J. Connor        | University of Michigan, Survey Research Center, Institute for Social Research |
| David Edson      | Mathematica Policy Research, Inc.   |
| Dhiren Ghosh     | Synectics for Management Decisions, Inc.                                      |
| Gail Henry       | QRC   |
| Donsig Jang      | Mathematica Policy Research, Inc.   |
| Max Larsen       | The Gallup Organization   |
| B. Pennell       | University of Michigan, Survey Research Center, Institute for Social Research |
| Michael Pergamit | National Opinion Research Center  |
| Linda Piccinino  | Abt Associates  |
| Sameena Salvucci | Synectics for Management Decisions, Inc.                                      |
| Carolyn Shettle  | Temple University, Institute for Survey Research                              |
| Lori Thurgood    | SRI International, Science and Technology Policy Program                      |
| Roy Whitmore     | Research Triangle Institute   |

### **Contractor Support Staff**

|               |                           |
|---------------|---------------------------|
| Denise Glover | Westat, Education Studies |
|---------------|---------------------------|

|                |                           |
|----------------|---------------------------|
| Joan Michie    | Westat, Education Studies |
| Richard Phelps | Westat, Education Studies |
| Faith Sproul   | Westat, Education Studies |

## Appendix C. Expert Panel Meeting Summary

---

An expert panel meeting was held at NSF on 5 December 2000, to discuss sampling design issues for SESTAT. (An agenda of the meeting is given in appendix A.) Most of the experts invited were from other federal agencies and were either familiar with the potential data sources or were experts in sampling methods.

Panel participants consisted of eight senior staff from the U.S. Census Bureau, three staff members from BLS, and one each from the National Center for Education Statistics, the Office of Management and Budget, the Energy Information Agency, NCHS, INS, and the Federal Reserve Board. An expert on similar survey efforts at Statistics Canada also attended. Presentations were made to this group by some of the many individuals at NSF or who are under contract to NSF who have studied SESTAT sampling issues. Among the speakers were Westat representatives, who presented the content of this report in some detail.

Expert panelists were asked to comment on the analysis and conclusions made by the researchers at NSF and in this report and to offer suggestions for future actions regarding NSF sampling activities. Some experts responded at the meeting or sent a short response after the meeting. The experts gave the issues substantial consideration and asked many questions, but in the end, no one suggested any better solution or had any better recommendations for approaching these challenges than those already proposed.

At the conclusion of the expert panel meeting, the co-chairs, Ron Fecso and Graham Kalton, offered the following summary of the expert panel's deliberations and recommendations for action:

- How well could other, existing large-scale surveys work for NSF's scientist, engineer, and technician population sampling needs?

(1) Establishment surveys probably would not satisfy the SESTAT data needs for several reasons. First, the problem of the employer as an intermediary screener may be difficult to overcome. Second, there is the non-representativeness of the population that is employed, and particularly the underrepresentation of the population employed in establishments above a minimum size, if the establishment survey has a minimum size cut-off.

(2) The sample size for the Census Bureau's SIPP is too small. Also, the SIPP would be difficult to use because of its complicated, nested rotating panel structure and the dedicated thematic foci of each wave.

(3) The small sample sizes of the CPS and NHIS are also severe limitations on their use as a screener survey. There are some possibilities of accumulating cases over time, but that would introduce complications.

(4) The large sample size of the NIS is a positive feature in considering its use as a screener survey. However, there would be problems using it as a screener survey given its focus on families with children and its disproportionate allocation to the IAP areas.

(5) The size, breadth, and currency of the Census Bureau's planned ACS make it a promising possibility to serve as a screener survey. However, at the time of the workshop, its funding situation was uncertain; whether it will be funded, the level of funding, and the continuation of funding into the future were unclear. Ongoing use of the ACS, particularly if a field of degree question is added, would provide opportunities for changes in the design and scope of the NSCG and, to a lesser extent, the NSRCG.

(6) Some possibilities for sampling new immigrants in S&E occupations may present themselves in the future at the INS, and lines of communication should be kept open. Currently, occupation is recorded on an INS form upon arrival, but it can be ambiguous or changed. Efforts to improve the currency and reliability of this information may be made in the near future. A frame with current and reliable information on degree and/or occupation of immigrants could help fill an important coverage gap in the current SESTAT structure.

- The panel felt that it would be unwise for NSF to rely solely on the ACS as its screener survey. The two main alternative strategies are as follows:

(1) Continue with the current panel, used throughout the 1990s, with freshening of the panel for new entrants to the S&E population. This option would retain all the past problems, such as response rate declines over the decade and the resultant degradation in the quality of the panel.

(2) Repeat the process of the last decade, starting with a sample drawn from the long form of the 2000 census to produce the first wave of a new NSCG in 2003.

With either of these strategies, there would still be a need to conduct the NSRCG to freshen the sample. NSF could switch to the ACS later in the decade, as soon as it is practicable. Alternatively, NSF could continue with one of the strategies throughout the rest of the decade, perhaps conducting pilot surveys with the ACS and then switching to the ACS at the beginning of the 2010 decade.

- If NSF adopts one of these strategies, it will probably need to continue its research on cost-effective quality improvements for the SESTAT system, which might include the following:

(1) Research on nonsampling error issues

(2) New technology options regarding cell phones and web-based data collection

(3) Improved ways to maintain current panels (e.g., better tracing and searching)

The main SESTAT policy decisions facing NSF at the time of the meeting were as follows:

- Sample quality considerations

(1) NSF is not conducting the NSCG in 2001. Thus, if the current panels are maintained and these panel members are next contacted in 2003, it will have been 4 years since the panel was last contacted.

(2) If a sample from the 2000 census long form is not selected and contacted soon, it will become difficult to find the individuals, and that difficulty will increase as time goes on.

- Budget considerations

(1) If NSF decides to wait for the ACS option to work out, and it then fails, budget limitations may make it impossible to restart the current survey later.

(2) The NSCG response rate is declining as the panel ages. It now stands at about 65%, and further declines could threaten the quality reputation of the survey. Maintaining the current panels for a while is the least expensive option.

Other important SESTAT policy decisions facing NSF follow.

- Subgroup considerations

(1) Any insistence on producing precise estimates for very small domains, e.g., for very small population groups, such as subdomains within American Indians and the disabled, can substantially increase costs. Resources spent on accumulating adequate sizes of rare subpopulations from the already rare population of scientists, engineers, and technicians are resources unavailable for other survey needs. An assessment needs to be made of the necessity for all the current survey categories. Some may be less cost effective to sample than others.

(2) It may be adequate to combine some small groups for detailed categories of responses and to report on them individually only for major categories of responses. Perhaps NSF could establish a cost-effective threshold population size (which some small groups may not meet).

(3) The population of the SDR is covered by other surveys but not at high enough sampling rates. It is therefore highly likely that the SDR will continue in its current form. Besides, the SDR has a dedicated user group that will likely not want to risk major changes.

## Appendix D. Desired and Acceptable Coefficients of Variation (CVs) for Estimates of the S&E Population

| Group of interest                    | Desired CV (%)                    | Acceptable CV (%)                 | 1997 SESTAT integrated database CV (%) |
|--------------------------------------|-----------------------------------|-----------------------------------|--|
| Total population                     | 0.1                               | 0.2                               | 0.2                                    |
| Highest degree level                 |                                   |                                   |  |
| Associate                            | 1.0                               | 1.5                               | N/A                                    |
| Bachelor                             | 0.4                               | 0.4                               | 0.4                                    |
| Master                               | 0.4                               | 0.6                               | 0.8                                    |
| Doctorate                            | 0.4                               | 0.4                               | 0.8                                    |
| Professional                         | 1.0                               | 1.5                               | 1.6                                    |
| Race/ethnicity                       |                                   |                                   |  |
| Hispanic                             | 0.8                               | 1.0                               | 1.6                                    |
| White                                | 0.8                               | 1.0                               | 0.2                                    |
| Black                                | 0.8                               | 1.0                               | 1.3                                    |
| Asian/Pacific Islander               | 0.8                               | 1.0                               | 1.0                                    |
| Other                                | 3.0                               | As it falls out based on sampling | 5.7 for American Indian/Alaska Native  |
| Citizenship                          |                                   |                                   |  |
| U.S.citizen – native                 | 0.3                               | 0.3                               | 0.3                                    |
| U.S. citizen – naturalized           | 1.5                               | 1.5                               | 1.6                                    |
| Non-U.S. citizen, permanent resident | 1.5                               | 1.5                               | 2.8                                    |
| Non-U.S. citizen, temporary resident | 3.0                               | As it falls out based on sampling | 5.6                                    |
| Sex                                  |                                   |                                   |  |
| Female                               | 0.1                               | 0.2                               | 0.2                                    |
| Male                                 | 0.2                               | 0.3                               | 0.3                                    |
| Age cohort                           |                                   |                                   |  |
| <30                                  | 1.0                               | 1.0                               | 1.2                                    |
| 30–39                                | 1.0                               | 1.0                               | 0.8                                    |
| 40–49                                | 1.0                               | 1.0                               | 0.7                                    |
| 50–59                                | 1.0                               | 1.0                               | 1.0                                    |
| 60–75                                | 1.0                               | 1.0                               | 1.2                                    |
| Disability status                    |                                   |                                   |  |
| Nondisabled                          | 1.0                               | 1.5                               | 0.2                                    |
| Disabled                             | As it falls out based on sampling | As it falls out based on sampling | 1.6                                    |

|  |                                   |                                   |                  |
|--|-----------------------------------|-----------------------------------|------------------|
| Employer sector                        |                                   |                                   |                  |
| Education                              | 1.0                               | 1.0                               | 1.1              |
| Government                             | 1.0                               | 1.0                               | 1.3              |
| Business/industry                      | As it falls out based on sampling | As it falls out based on sampling | 0.4              |
| Broad field of highest degree          |                                   |                                   |                  |
| Mathematics and computer sciences      | 0.3                               | 0.5                               | 1.0              |
| Life sciences                          | 0.3                               | 0.5                               | 0.8              |
| Physical sciences                      | 0.3                               | 0.5                               | 1.0              |
| Social sciences                        | 0.3                               | 0.5                               | 0.6              |
| Engineering                            | 0.3                               | 0.5                               | 0.6              |
| Non-S&E fields                         | —                                 | —                                 | 0.9              |
| S&E related                            | 2.5                               | 3.5                               | —                |
| Broad category of occupation           |                                   |                                   |                  |
| Mathematicians and computer scientists | 1.5                               | 2.5                               | 1.2              |
| Life scientists                        | 1.5                               | 2.5                               | 2.2              |
| Physical scientists                    | 1.5                               | 2.5                               | 2.2              |
| Social scientists                      | 1.5                               | 2.5                               | 2.4              |
| Engineers                              | 1.5                               | 2.5                               | 0.9              |
| Non-S&E occupations                    | —                                 | —                                 | 0.5              |
| S&E managers                           | 2.0                               | 3.0                               | a                |
| Other managers                         | 2.5                               | 3.5                               | a                |
| Health-related occupations             | 1.5                               | 2.5                               | 1.9 <sup>b</sup> |
| Precollege teachers                    | 2.5                               | 3.5                               | 2.1 <sup>b</sup> |
| Non-S&E postsecondary teachers         | 2.5                               | 3.5                               | 5.1 <sup>b</sup> |
| Technicians                            | 2.0                               | 3.0                               | 2.9 <sup>b</sup> |
| Other occupations                      | 3.0                               | 4.0                               | —                |

— = not a design objective.

<sup>a</sup> This population cannot be specifically identified in current database.

<sup>b</sup> This CV is limited to the current SESTAT population definition.

SOURCE: National Science Foundation, Division of Science Resources Statistics, staff recommendations 2006.

## **Appendix E. Characteristics of Establishment-Based Collections**

The section "Establishment-Based Sample Design" discusses the possible use of an establishment-based data collection for SESTAT and possible sampling frames for establishments. This appendix provides more detail on the sampling frames developed at the Census Bureau and at the Bureau of Labor Statistics. It also includes a description of the Census Bureau's National Employer Survey.

### **Available Establishment Lists**

As described, Dun's Market Identifiers represent the only nationally comprehensive list of establishments available to surveyors outside the federal government. Two other comprehensive lists exist at the Census Bureau and at the Bureau of Labor Statistics. The data can be accessed only by those agencies because of federal confidentiality laws.

The Census Bureau's Standard Statistical Establishment List (SSEL) is assembled from data provided by the Internal Revenue Service (IRS) and the Social Security Administration (SSA). The Census Bureau conducts the Company Organization Survey annually, which helps establish the necessary connections between the IRS and SSA data files.

The Bureau of Labor Statistics (BLS) conducts the Occupational Employment Statistics (OES) Survey. The OES includes annual samples of approximately 400,000 establishments, taking 3 years to fully collect data from a total sample of 1.2 million establishments.[1] The 1997 OES sample frame did not include establishments with fewer than five employees. (Sampling of establishments with one to four employees began in 1998.) According to BusinessUSA, about 70% of companies in the services sector (SIC codes 70–87) and 63% of companies in the computer services sector (SIC 737) have fewer than five employees. In calculating OES estimates, establishments with five to nine employees were weighted more heavily in an effort to compensate for the undercoverage. To the extent that establishments with fewer than five employees differ from those with five to nine employees, the problem of coverage bias may remain.

OES data collection is conducted primarily by mail and is managed by state employment security agencies using samples and procedures provided by BLS. For very large establishments, site visits are used to ensure that the data for the company are collected. Officials from the sampled establishments respond to the survey.

### **National Employer Survey**

The National Employer Survey (NES) is one example of many establishment-based surveys that are conducted in the United States. It is carried out by the Census Bureau, using the Bureau's own SSEL, for the U.S. Departments of Education and Labor. First conducted in 1994 and repeated in 1997, the NES was a telephone survey of employers of more than 4,000 private establishments. In the 2000 survey, employer-respondents received a batch of questionnaires to distribute randomly to their employees. Since the 2000 survey subsampled the sample of employers from the 1997 survey, it is not fully representative cross-sectionally for 2000, but it should be able to provide some longitudinal information based on a 1997 representative cross-section. Public-sector employers, not-for-profit institutions, establishments with less than 20 employees, and corporate headquarters were excluded from the sample.

The jointly funded Educational Quality of the Workforce (EQW) project based at the University of Pennsylvania determined the NES survey design and content and analyzed its resulting data. EQW is primarily interested in studying school-to-career paths and workplace training among workers, so survey items tend to focus on such topics. According to EQW, "...the EQW-NES provides an important baseline for understanding how employers recruit workers, how they organize work, which educational credentials and experiences they use in screening job applicants, and what role education and training play in providing a skilled workforce." The EQW-NES differs from other national surveys because it focuses on the interaction of establishment practice, organizational structure, and workforce proficiency; documents how employers satisfy their need for skilled employees (in particular cataloging employer attitudes toward schools as likely suppliers of skilled employees); and measures the outcomes of both formal and informal training.

As noted above, the 1997 NES sample size was more than 4,000 establishments. The 1997 survey did not sample employees. The 2000 NES, which did sample employees, was half the size of the 1997 survey. Another administration of the NES was planned for 2003. At its current sample size, however, the NES is too small to meet NSF's data needs.

---

#### **Footnotes**

[1] The survey includes private establishments classified in the agricultural services; mining; construction; manufacturing; transportation; public utilities; wholesale and retail trade; and finance, insurance, real estate, and services industries. All state and local government establishments as well as private and government establishments in selected Standard Industry Classification (SIC) codes are included. See [http://www.bls.gov/oes/oes\\_emp.htm#scope](http://www.bls.gov/oes/oes_emp.htm#scope) for a complete list of the NAICS codes corresponding to the industries that are included in the survey. Note that only nonfarm employment is included and all self-employed individuals are excluded from this survey.

## Bibliography

---

- Alexander CH, Dahl S, Weidman L. 1997. Making estimates from the American Community Survey. In: *Proceedings of the Section on Government Statistics and Section on Social Statistics*, p 88–97. Alexandria, VA: American Statistical Association.
- Battaglia MP, Starer A, Oberkofler J, Zell ER. 1995. Pre-identification of non-working and business telephone numbers in list-assisted random-digit-dialing samples. In: *Proceedings of the Section on Survey Research Methods*, p 957–962. Alexandria, VA: American Statistical Association.
- Citro CF, Kalton G, editors. 1989. *Surveying the Nation's Scientists and Engineers: A Data System for the 1990s*. Washington, DC: National Academy Press.
- Cochran WG. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.
- Ezzati-Rice TM, Zell ER, Battaglia MP, Ching LYH, Wright RA. 1995. The design of the National Immunization Survey. In: *Proceedings of the Section on Survey Research Methods*, p 668–672. Alexandria, VA: American Statistical Association.
- Judkins D, Marker D, Waksberg J. 1996. National Health Interview Survey: Research for the 1995 Redesign. Rockville, MD: Westat.
- Kish L. 1990. Rolling Samples and Censuses. *Survey Methodology* 16:63–78.
- Lepkowski JM. 1988. Telephone sampling methods in the United States. In: Groves RM, Biemer PP, Lyberg LE, Massey JT, Nicholls WL II, Waksberg J, editors. *Telephone Survey Methodology*, p 73–98. New York: John Wiley & Sons, Inc.
- Marker D, Edwards WS. 1997. Quality of the DMI file as a business sampling frame. *Proceedings of the Survey Research Methods Section, American Statistical Association*, p 21–30. Alexandria, VA: ASA.
- Wallace L, Bryant C, Chapman DW, Marker DA, Moriarity CL. 1995. Weighting and estimation procedures for the 1994 NEHIS. In: *Proceedings of the Survey Research Methods Section, American Statistical Association*, p 192–197. Alexandria, VA: ASA.

## Suggested Citation, Acknowledgments

---

Fecso R.S., Choudhry G.H., Kalton G., Chu A. and Phelps R. 2007. *Current and Alternate Sources of Data on the Science and Engineering Workforce*. Working Paper SRS 07-202 . Arlington, VA: Division of Science Resources Statistics, National Science Foundation.

---

The authors thank Norman Bradburn, Lynda Carlson, and Mary Frase (all of the National Science Foundation) for management support and editorial input; Joan Mitchie (Westat) for review and contract support; and Lawrence Burton, Mary Golladay, Kelly Kang, Richard Morrison, Mark Regets, Wendy Scholetzky, John Tsapogas, and Nirmala Kannankutty for helpful comments on early versions of the paper.

### Division of Science Resources Statistics

Lynda T. Carlson  
*Division Director*

Mary J. Frase  
*Deputy Director*



### Division of Science Resources Statistics (SRS)

The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA  
Tel: (703) 292-8780, FIRS: (800) 877-8339 | TDD: (800) 281-8749