Using the American Community Survey as the Sampling Frame for the National Survey of College Graduates

Working Paper | NCSES 12-201 | August 2012

by Ronald S. Fecso (formerly Division of Science Resources Statistics, National Science Foundation), Mary J. Frase (retired), and Nirmala Kannankutty (National Center for Science and Engineering Statistics, National Science Foundation)[1]

Disclaimer and Acknowledgements

Working papers are intended to report exploratory results of research and analysis undertaken by the National Center for Science and Engineering Statistics (NCSES). 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.

The background work for this study was done by the Division of Science Resources Statistics (SRS), the predecessor to NCSES. Section 505 of the America COMPETES Reauthorization Act of 2010 renamed SRS as the National Center for Science and Engineering Statistics. The ideas elaborated in this report were adapted from material prepared for the Panel on Assessing the Benefits of the American Community Survey for the Committee on National Statistics, National Research Council. The final report of the Panel, published in 2008, *Using the American Community Survey for the National Science Foundation's Science and Engineering Workforce Statistics Programs*, can be found at http://www.nap.edu/catalog.php?record_id=12244. The report documents the choice of a panel design for NCSES's National Survey of College Graduates (NSCG) for the 2010 decade.

The initial draft of this paper was prepared by Ron Fecso, previously Chief Statistician at NSF/Division of Science Resources Statistics, and currently Chief Statistician of the U. S. Government Accountability Office (GAO). He began work on this paper at NSF and, with the approval of GAO, further work was done after he left NSF. Further revisions to the paper were made by Mary Frase and Nirmala Kannankutty.

Suggested Citation

Fecso RS, Frase MJ, Kannankutty N. 2012. *Using the American Community Survey as the Sampling Frame for the National Survey of College Graduates*. Working Paper NCSES 12-201. Arlington, VA: National Science Foundation, National Center for Science and Engineering Statistics. Available at http://www.nsf.gov/statistics/ncses12201/.

Contact

Stephen Cohen,
Chief Statistician
National Center for Science and Engineering Statistics
National Science Foundation
4201 Wilson Blvd., Suite 965
Arlington, VA 22230
E-mail: scohen@nsf.gov
Phone: 703-292-7769

Contents

Introduction 4
Background on NSF's Workforce Surveys 5
NSCG Sample Design 8
The American Community Survey 14
Replacing the Decennial Long Form with the ACS as the NSCG Frame 15
    Sample Size 18
    Timing 19
    Quality Improvements 20
Adding a Field of Degree Question to the ACS 22
    NSCG Sampling with an ACS FOD Item 24
    Technical Issues for Sampling from ACS 26
Analytic Uses of ACS Data 31
Conclusion 32
References 33
Notes 34

Tables

Table 1. Degree field and occupational classification of respondents to the 2003 NSCG 9
Table 2. Yield of SESTAT-eligible cases from the 1993 and 2003 NSCG 10
Table 3. NSCG sample sizes, respondents, and unweighted response rates: 1993–2006 13
Table 4. Issues of concern regarding the two FOD question versions 29

Figure

Figure 1. SESTAT surveys in the 1990s and 2000s 7

Introduction

Congress has mandated the National Science Foundation (NSF) "to provide a central clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources and to provide a source of information for policy formulation by other agencies of the Federal Government" (NSF Act of 1950, as amended; 42 U.S.C. 1862). A critical component of this mission is information on the science and engineering (S&E) workforce. NSF thus reports information for the United States based on data collected on the number, characteristics, and employment of the S&E workforce. Additionally, the Science and Engineering Equal Opportunities Act (Public Law 96-516) gave a further mandate to NSF to ensure that obtaining information on women, minority group members, and persons with disabilities in the S&E workforce were important considerations in data collection and analysis.

NSF has obtained such information in the past in large part through its National Survey of College Graduates (NSCG), conducted for NSF by the U.S. Census Bureau. The NSCG is a critical component of NSF's Scientists and Engineers Statistical Data System (SESTAT), an integrated database derived from three surveys using multiple frames to collect data on scientists and engineers in the United States. Between 1960 and 2000, the long form of the decennial Census provided the sampling frame for the NSCG or its predecessors. Because the long-form component of the decennial Census was eliminated in 2010, an alternative sample frame for the NSCG was needed. The American Community Survey (ACS) replaced the long form and remains the only feasible alternative sample frame for the NSCG, and NSF has received authorization to utilize the ACS for this purpose.

This paper presents ideas on how to use the ACS data as the replacement for the decennial long-form sampling frame. The discussion goes further to present other quality enhancements possible when using the ACS for improving the NSCG, including survey design options that could lead to survey and data quality improvement. Additionally, there is discussion of the potential analytic uses that are possible with the ACS. It was prepared by the National Center for Science and Engineering Statistics (NCSES) as part of the background material for the Panel on Assessing the Benefits of the American Community Survey for the NSF Division of Science Resources Statistics, Committee on National Statistics, National Research Council. The final

report of the panel, published in 2008, *Using the American Community Survey for the National Science Foundation's Science and Engineering Workforce Statistics Programs*, can be found at http://www.nap.edu/catalog.php?record_id=12244.

Background on NSF's Workforce Surveys

The target population for the NSF/NCSES SESTAT system of surveys is noninstitutionalized individuals living in the United States who are 75 years of age or less and are considered to be scientists and engineers. NCSES currently defines scientists and engineers as those who hold a bachelor's degree or higher in an S&E or S&E-related field OR who have a bachelor's degree or higher in a non-S&E field but hold an S&E or S&E-related occupation.[2]

NSF uses a system of three surveys with separate frames to achieve coverage of the population of scientists and engineers. These surveys are the following: the NSCG, the National Survey of Recent College Graduates (RCG), and the Survey of Doctorate Recipients (SDR). The three surveys utilize nearly identical data collection instruments and data processing procedures and are fielded at the same time with the same reference date.[3] The three surveys have been designed to provide the advantages of a large sample size and greater coverage of the target population, with special emphasis given to relatively rare populations (e.g., doctorates, recent graduates, and minorities).

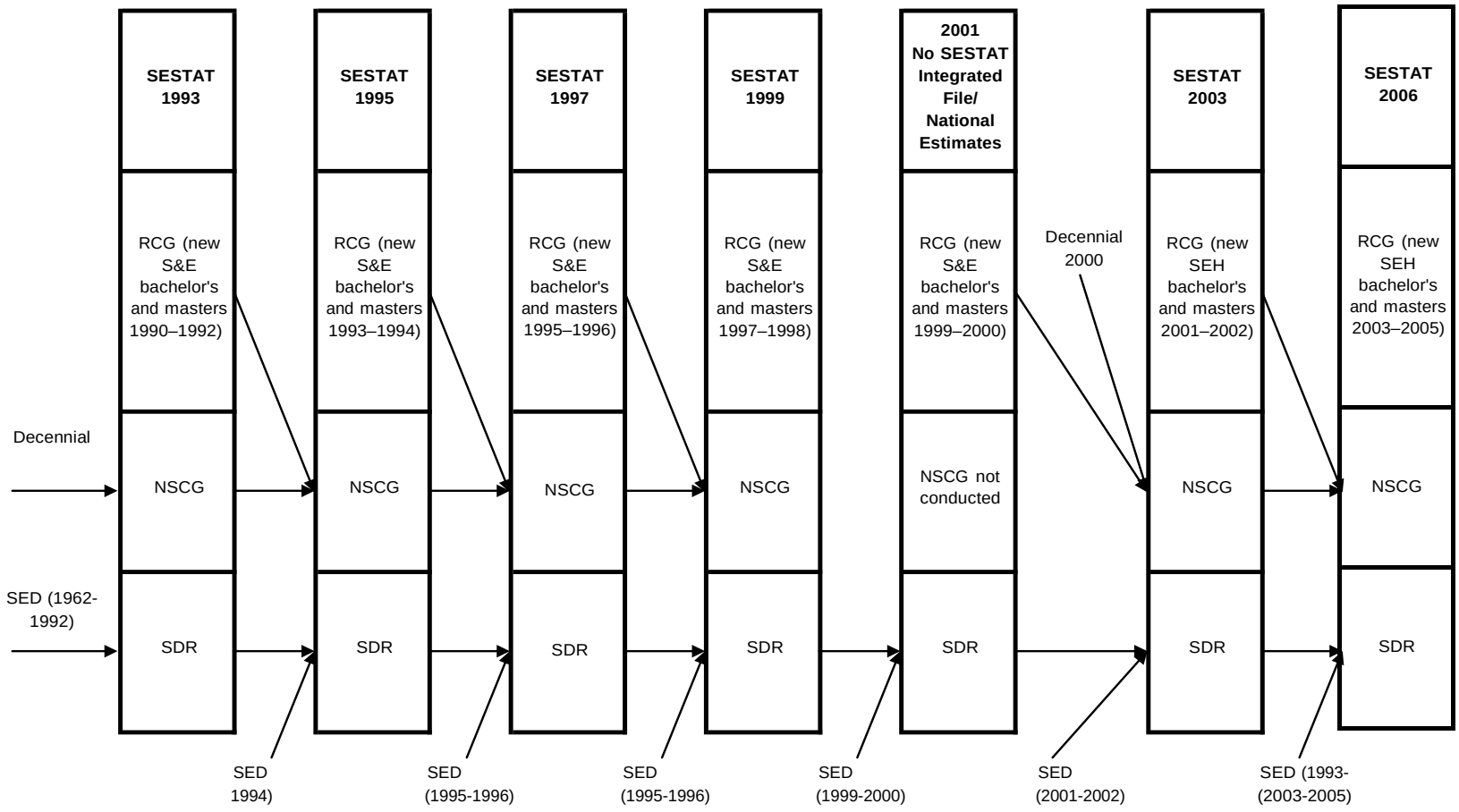
Data from the three surveys for all cases that qualify as scientists and engineers according to the SESTAT target population definition are integrated into a comprehensive database, the SESTAT integrated file, covering all college-educated scientists and engineers in the United States. In creating the integrated database, the issue of potential for eligibility for more than one of the surveys is addressed, as it is possible that cases identified in one survey might also be eligible for another survey.[4] The integrated file is used to produce national estimates of the number and characteristics of scientists and engineers in the United States.

The SESTAT surveys are conducted approximately every 2 to 3 years and provide cross-sectional time-series data; preliminary SESTAT longitudinal files have been prepared for the period covering 1993–99.[5] The NSCG provides the majority of cases in the SESTAT integrated

database and represents the “stock” of scientists and engineers at the beginning of the decade. The SDR provide the stock of experienced U.S. doctorates, as well as the “flow” of U.S. new science, engineering, and health (SEH) doctorates. The RCG captures the flow of new U.S. SEH bachelor’s and master’s graduates.

The NSCG is a panel survey with a new panel selected at the beginning of each decade. Respondents to the NSCG who are identified as eligible for the SESTAT target population are eligible for the NSCG follow-up surveys for the rest of the decade. The RCG is a cross-sectional survey of new bachelor’s- and master’s-degree recipients in SEH fields; after entering the SESTAT system in the RCG (a new flow), a subsample is followed in the NSCG (as part of the stock). The target population for the SDR is all SEH doctorates awarded at U.S. institutions. While the overall sample size of the SDR is held steady, for each new round a sample of new SEH doctorates is added to the SDR sample from its frame, the Survey of Earned Doctorates (SED). Figure 1 shows a conceptual diagram of the stocks and flows that make up the SESTAT system of surveys.

FIGURE 1. SESTAT surveys in the 1990s and 2000s



SESTAT = Scientists and Engineers Statistical Data System

RCG = National Survey of Recent College Graduates

NSCG = National Survey of College Graduates

SDR = Survey of Doctorate Recipients

SED = Survey of Earned Doctorates

NSCG Sample Design[6]

In 1962, NSF and other agencies sponsored the Postcensal Manpower Survey in order to obtain information on S&E personnel resources. This was a single, cross-sectional survey with the sample derived from the long form of the 1960 decennial Census (hence, "postcensal"). In 1972, NSF sponsored another postcensal survey (the Professional, Technical and Scientific Manpower Survey) for similar purposes, with smaller follow-ups through 1978. In the 1980s, NSF again conducted a postcensal survey with follow-ups through 1989. After a major redesign following the 1990 Census, the NSCG continued in the mode of a large postcensal (baseline) survey, with smaller follow-up surveys, during the remainder of the decade.[7] The decennial long form did not contain any information on educational background, aside from the level of educational attainment that would allow for the identification of individuals with S&E or S&E-related degrees. Therefore, the baseline NSCG served two purposes: (1) to provide a once-in-a-decade view of all college graduates in the United States, and (2) to act as a screening device (through detailed educational histories collected in the NSCG) for obtaining a sample of scientists and engineers for the SESTAT integrated file. The NSCG follow-up surveys, generally every 2 to 3 years, were limited to those meeting the SESTAT target population definition of a scientist or engineer.

At the beginning of each decade, the U.S. Census Bureau created a sampling frame for the NSCG based on the decennial Census, which was used to draw the baseline NSCG sample. All long-form respondents with a bachelor's degree or higher at the time of the Census had a chance of selection into the postcensal NSCG sample.[8] To capture the stock of scientists and engineers from the NSCG, it was necessary to sample all occupations from the decennial long form because a high proportion of individuals with S&E or S&E-related degrees do not work in S&E or S&E-related occupations (57% of this population in the 2003 NSCG either worked in a non-S&E occupation or were not working). Additionally, the NSCG is the only source of information for the SESTAT integrated database on individuals with non-S&E degrees working in S&E or S&E-related occupations. In 2003, the size of this population was 1,510,100 (see table 1).

Utilizing data from the educational history section of the 2003 NSCG,[9] it was possible to identify individuals sampled in non-S&E occupations, but who had S&E or S&E-related degrees,

as well as to identify individuals with non-S&E degrees working in S&E or S&E-related occupations. Especially at the bachelor's level (which is the large majority of all those with a bachelor's degree or higher), people with S&E degrees are likely to be in a non-S&E occupation. Individuals in S&E occupations and those in non-S&E occupations with a higher likelihood of being held by someone with an S&E degree are sampled at higher rates than other cases.

TABLE 1. Degree field and occupational classification of respondents to the 2003 NSCG

Degree status / Occupation status	S&E occupation	S&E-related occupation	Non-S&E occupation	Not working	Total
At least one S&E degree	3,648,286 (22,669)	2,024,479 (6,676)	6,484,052 (13,959)	2,703,312 (7,877)	14,860,128 (51,181)
No S&E degrees, but at least one S&E-related degree	219,317 (1,135)	2,341,453 (5,637)	1,054,030 (2,130)	746,094 (1,623)	4,360,894 (10,525)
No S&E or S&E-related degrees	720,632 (2,897)	789,427 (1,901)	15,292,588 (26,020)	4,596,750 (7,878)	21,399,398 (38,696)
Total	4,588,235 (26,701)	5,155,359 (14,214)	22,830,670 (42,109)	8,046,156 (17,378)	40,620,420 (100,402)

NSCG = National Survey of College Graduates; S&E = science and engineering.

NOTE: Top values in each cell are weighted estimates; values in parentheses are unweighted cases.

In 1993, NSF requested that the Census select about 215,000 individuals for the NSCG sample from the decennial long-form sample frame. The 1993 sample netted about 75,000 cases that met NSF's definition of a scientist or engineer and therefore were eligible for the SESTAT integrated database and the NSCG follow-up surveys. In 2003, about 171,000 individuals were selected from the 2000 Census long-form frame, which yielded approximately 67,000 cases with S&E and S&E-related degrees or occupations. Table 2 shows the yield of cases that were obtained from the NSCG in 1993 and 2003.

TABLE 2. Yield of SESTAT-eligible cases from the 1993 and 2003 NSCG

Characteristic	1993 NSCG	2003 NSCG
Sample size	214,643	170,800
Respondents	148,905	100,402
Individuals with at least one S&E degree or S&E occupation[a]	74,462	–
Individuals with at least one S&E/S&E-related degree or S&E/S&E-related occupation[a]	–	66,504
Yield A (Sample size to usable cases)	2.88:1	2.56:1
Yield B (Respondents to usable cases)	2.00:1	1.51:1

– Not applicable.

NSCG = National Survey of College Graduates. S&E = science and engineering. SESTAT = Scientists and Engineers Statistical Data System

[a] The National Science Foundation changed its definition for cases it would follow up as part of the S&E population from
1993 to 2003.

The Census Bureau conducts the NSCG for NSF because it is a subsample of the decennial Census long-form sample. Any record derived from decennial Census records is protected by Title 13 and must be used only under Census Bureau supervision; thus the NSCG sample must be drawn by Census, and the survey must be conducted by the Census Bureau. The postcensal NSCG survey has generally been fielded about 3 years after the decennial Census because of the time needed to process the decennial Census and make the data available for NSCG sampling.

The postcensal NSCG sample design has been a two-phase, stratified random sample of individuals with at least a bachelor's degree at the time of the Census. Phase 1 consisted of the sampling households for the Census long-form sample. That procedure utilized a stratified systematic sample, with differing sampling rates for administrative areas of different sizes (about a 1-in-12 to 1-in-16 sampling rate). Phase 2, the NSCG postcensal sample, consisted of subsampling persons with at least a bachelor's degree whose reported age from the long-form records would result in them being age 75 years of age or younger at the time of reference date for the postcensal NSCG. In 2003, the major sampling variables used to create the strata for the frame were the following: educational attainment (bachelor's degree or higher), by highest degree level achieved; occupation; demographic group (which combines citizenship,

race/ethnicity, and disability status); 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.

Survey responses to the postcensal NSCG baseline survey determined eligibility for the follow-up NSCG surveys, depending on whether a respondent met the definition of a scientist or engineer for SESTAT. Because the NSCG baseline survey collects information on educational background, the sampling for the follow-up survey includes the original sampling strata, as well as field of highest S&E degree.

A review of the NSCG sample design found that the long-form frame approach had sample selection and coverage problems.[10] Three significant problems were the following:

1. Responses to the Census long-form questionnaire were not an efficient means of identifying those with S&E degrees because there was only information about the highest level of degree attained and not of degree fields. It did not provide the means to identify those with SEH degrees.

   The past sample design using the postcensal survey as a screening mechanism made possible valuable comparisons of scientists and engineers with nonscientists and nonengineers once a decade. The 1993 survey created a large database about college-educated individuals that could be utilized for analysis by those interested in fields other than S&E. However, the postcensal sample included many people who were not SESTAT eligible. On average, almost 3 cases were surveyed to find one SESTAT-eligible respondent. Using the 1993 NSCG, it was possible to improve the efficiency of the 2003 NSCG sample design, to approximately 2 sample cases to yield one SESTAT-eligible respondent, while still maintaining a sufficient level of accuracy to make comparisons between S&E, S&E-related and non-S&E domains.

2. The approach of using a decennial Census to identify the stock of engineers and scientists to be interviewed over the decade, together with new graduates of U.S. institutions in S&E fields from the RCG and the SDR, meant that some population groups were missed. Some were missing from the beginning in the postcensal NSCG, while for other groups the coverage problems grew worse over the decade.

   a. Those with non-S&E degrees who entered S&E or S&E-related jobs after the postcensal NSCG were not covered in any of the surveys later in the decade. The computer occupations, for example, included a significant number of workers not educated in a science, engineering, or related discipline.

   b. The NSCG is the only SESTAT survey that includes scientists and engineers whose degrees were all earned abroad. However, this population was captured in the sample only once a decade in the baseline survey. Foreign-educated scientists and engineers entering the United States after the decennial Census and receiving no further degrees in the United States were not included in any SESTAT survey, and so the undercoverage of this group grew throughout the decade.

   c. A substantial number of scientists and engineers fall into both of the populations described above: non-S&E graduates in S&E and S&E-related occupations and foreign-educated scientists and engineers. Data from the 2003 NSCG and the 2003 SESTAT integrated file show that there were 720,632 individuals in S&E occupations and 789,428 individuals in S&E-related occupations with non-S&E degrees only.[11] Additionally, there were estimated to be 1,470,729 individuals in the SESTAT population who had only foreign degrees. Taking into account some overlap between these two populations, approximately 2.6 million individuals in 2003 in the SESTAT population (a) had an S&E occupation, but no S&E degree; or (b) had only foreign degrees. Such individuals represent approximately 12% of the 2003 SESTAT population of 21.6 million persons.

3. Another problem was increasing cumulative nonresponse through the decade. The postcensal surveys have had a response rate that historically has been near 80%; the 2003

postcensal NSCG had an unweighted response rate of 63%. Follow-on surveys later in the decade generally have had very good response rates (in the high 90s), but past nonrespondents are removed from these surveys. Thus, there is an increasing unconditional nonresponse rate as the decade progresses (see Table 3 for the NSCG sample sizes, respondents, and unweighted response rates).

TABLE 3. NSCG sample sizes, respondents, and unweighted response rates: 1993–2006

Year	1990 Decennial sample				2000 Decennial sample			
	Sample size	Respondents[a]	Conditional response rate	Unconditional response rate	Sample size	Respondents[a]	Conditional response rate	Unconditional response rate
1993	214,643	148,905	78	78	–	–	–	–
1995	53,315	46,622	95	74	–	–	–	–
1997	46,075	42,506	94	70	–	–	–	–
1999	35,714	31,928	91	63	–	–	–	–
2003	–	–	–	–	170,800	100,402	63	63
2006	–	–	–	–	59,349	51,694	88	55

– Not applicable.

[a] Because of cross-survey multiplicity, only a subset of NSCG respondents was integrated into the SESTAT file. The number in each year is: 1993—67,855; 1995—46,361; 1997—42,331; 1999—31,825; 2003—59,969; 2006—To be determined.

The NSCG sample from the 1980 decennial Census remaining at the end of the 1980s was discarded at the recommendation of a previous CNSTAT panel due to coverage, bias, and other problems. NSF considered doing the same with the 1990 postcensal remaining sample, primarily due to the low unconditional response rate (63%), as well as concerns about panel attrition, fatigue, and possible bias. However, problems with this approach include the complete loss of longitudinal continuity and a lack of information about how nonresponse adjustments during the decade might cause a shift in the time series. NSF addressed these issues by embedding an experiment in the design of the 2003 NSCG. In addition to drawing a new population from the 2000 decennial long-form sample, NSF also included the remaining 1999 NSCG respondent population (which included cases originally sampled in the 1993 NSCG, as well as the 1995–99 RCG surveys) to receive the 2003 survey. The evaluation of this experiment found some large differences in estimates of the scope of coverage between various nonresponse adjustment cells made from newly drawn 2000 postcensal samples versus retained longitudinal samples from the 1999 NSCG in 2003 (Finamore, Hall, and Fecso 2006). Further research is needed to determine all the factors that may have contributed to the differences.

NSF's potential sampling frames and designs have been reviewed several times by scientific bodies over the decades. Each time, the design based on the Census long form for the NSCG was found to be the best available strategy for the time.[12] More recently, in preparation for the NSCG surveys in the 2000s, NSF explored alternative sampling frames for the S&E workforce data system. NCSES looked for a frame that could provide a more complete representation of the universe of scientists and engineers than the long-form approach (Fecso, Baskin, et al. 2007). No suitable alternative to the long-form frame for the NSCG was identified, primarily because no other survey had sufficient sample size to include the number of scientist and engineers, a relatively rare population, to meet the needs of the NSCG and SESTAT. However, the ACS was identified as a future potential alternative. At the time, the ACS was in a developmental mode. The ACS is now fully operational, and NCSES has been given the authority to use the ACS as a sample frame for the NSCG in the future.

The American Community Survey

The ACS, conducted by the U.S. Census Bureau, has replaced the decennial long form.[13] It is conducted every month, with estimates from the monthly samples aggregated into annual national estimates and aggregated over longer periods for smaller area estimates. The ACS surveys about 3 million households annually. Questions on the ACS are generally identical to the questions that were on the decennial long form.[14] Therefore, the ACS can provide reasonably detailed information about households and individuals each year rather than once a decade from the decennial Census.

The ACS has a new sample each month. It samples about 250,000 addresses each month, or some 3 million each year.[15] Over a decade, the ACS will survey approximately 30 million addresses, compared with the long form's 17 million housing units at one time in the decennial Census. The sampling frame for the ACS is the U.S. Census Bureau's Master Address File, which is updated throughout the decade to keep it current. The samples are distributed throughout the country with no area or other cluster sampling, but there will be some higher sampling fractions in small governmental units.

The ACS is conducted using an initial mail-out, mail-return, self-response questionnaire. Mail nonresponse follow-up is first by computer-assisted telephone interviewing (CATI), followed

by computer-assisted personal interviewing (CAPI) of a subsample of the remaining nonrespondents. A key function of the ACS is to produce estimates for various levels of geography (from small areas to the nation). ACS provides estimates annually for areas (and population groups) of 65,000 or more persons.[16] Estimates for smaller areas/groups are provided based on data aggregated over multiple years. Estimates for the smallest areas or population groups are available based on data aggregated over 5 years, with reliability similar to that provided for these groups in recent decennial Censuses.[17]

Replacing the Decennial Long Form with the ACS as the NSCG Frame

The concept of a baseline postcensal NSCG, followed by subsequent panel follow-up surveys, has been used for the NSCG for a variety of reasons. Identifying and then locating the stock of scientists and engineers of interest are both difficult and expensive. Having identified them once through the initial baseline NSCG, it was most efficient to keep them in the NSCG throughout the decade. Additionally, this also provided some stability to the estimates being made. The other SESTAT surveys do provide some of the new flows of U.S.-educated scientists and engineers to the overall population (e.g., new bachelor's- and master's-degree SEH graduates from the RCG, and new SEH doctorates from SDR). The alternative to maintaining the NSCG postcensal sample was to draw a new sample every 2 to 3 years, but there was no benefit and a considerable cost to doing so. Additional screening surveys with large samples would be very expensive, and there would be no improvement in the coverage of the population because the sample frame (the decennial long form) would not change.

There are some populations covered by the NSCG that are covered by the other two SESTAT surveys. For example, the stock and flow of SEH doctorates are well covered by the SDR. Earlier sampling-frame research conducted by NCSES recommended that the SDR be maintained, given the small number of U.S. SEH doctorates in the NSCG and the great policy interest in SEH doctorates.[18]

The SDR also has great value as a stand-alone survey, enabling longitudinal analysis of the careers of U.S. SEH doctorate holders. Use of the ACS as a frame for the NSCG does not change the value of and need for the SDR survey. ACS does not provide a sufficient sample of doctorate recipients, unless multiple ACS years are combined and such a combining of multiple

years would nullify any of the quality-enhancing features of using the ACS. The desire for small domain estimates for these individuals (e.g., doctoral field by race/ethnicity by sex) and the readily available SED (a census of all U.S.-earned SEH doctorates) for a sampling frame makes continued use of a separate SDR survey a very efficient approach for the SEH doctorate domain.

As was the case with the long-form records, the ACS records can be stratified by households or persons with specific characteristics. Thus, the ACS can provide an efficient frame for follow-on surveys. The ACS provides a means to include in the NSCG frame scientists and engineers earning all their degrees abroad who then come to the United States and enter the labor force. Similarly, it can provide better coverage throughout the decade of non-S&E graduates working in S&E or S&E-related occupations, a shortcoming of the long-form sample design.

With no change in its survey content, the ACS can be used as a frame for the NSCG in several different ways. There is more flexibility in possible NSCG designs compared to the previous long-form frame, particularly given that current data throughout the decade will be available. The continuous survey approach of the ACS makes the following options (or some combination of them) possible for an NSCG sampling frame. Cost considerations are also important because the potential costs of pursuing the various options are likely to vary considerably.

1. Continue the current approach, refreshing the sample once a decade

ACS data could be used once a decade to draw a new panel for the NSCG. The existing ACS questions are nearly identical to those found on the past decennial long form, and they are suitable for a screening survey for NSCG, much like what was done using the Census long form. The survey procedures could then follow those previously used in the postcensal NSCG.

2. Update the entire sample more than once a decade

A fresh sample could be selected from the ACS for each round of the NSCG—or, if not every survey cycle, more frequently than once a decade. With a freshly drawn sample, the coverage of the full SEH population would be more current and would reduce or eliminate the coverage problems that develop over the decade in the once-a-decade

approach. This approach would involve screening for eligible scientists and engineers each time a new sample was drawn from the ACS and would require large sample sizes (and, thus, higher costs) each time a new sample is drawn.

Drawing new samples more frequently would also reduce (or eliminate) the longitudinal feature of the ACS. If the sample were redrawn every survey cycle, the NSCG would become a series of cross-sectional surveys. One result would be considerably more variation in the estimates from cycle to cycle than with the current longitudinal design. This phenomenon would be especially noticeable in important small-domain estimates, such as field by race/ethnicity estimates.

3. Rotating sample approach[19]

ACS could be used to convert the NSCG to a rotating sample design. For example, initially the 2003 NSCG sample could be divided into several equal-sized panels. A new panel would be drawn from the ACS to replace one of these NSCG panels. Each survey cycle, a new ACS panel could replace an old NSCG panel until the entire 2003 NSCG was rotated out. Then the oldest ACS sample panel could be replaced by a draw from the most recent ACS year(s) each NSCG survey cycle. This approach captures some of the coverage advantages of Option 2 but would involve more-complex survey operations (and screening for the new sample portion to identify scientists and engineers). Rotation designs are often recommended in longitudinal surveys, where there is a problem with sample attrition due to respondent fatigue. The NSCG has had declining response rates as each decade progressed and increasing refusal rates.

The duration of the transition process of phasing in ACS panels could be lengthened or shortened depending on the size of the NSCG panels to be replaced (or replacing multiple 2003 NSCG panels each survey cycle). The rotation schedule for the transition to the ACS need not be the same as that established for the longer term once the NSCG consists entirely of ACS panels.

4. Selective updates

The ACS could be used as a frame to update the sample for certain domains whose coverage becomes a problem as the decade progresses (e.g., recent immigrants) or for populations of special interest. Data items in the ACS could be used to subset ACS respondents into a frame for targeted groups(s). For example, the question on when a person came to the United States could be used to create a subset that contains recent immigrants. The ACS could be used as a frame to examine real-time events (e.g., the rise and fall of technology/information technology [IT] firms and the impact on IT occupational employment). Such supplemental frames for special domains could be sampled during any survey cycle rather than once a decade.

ACS data could be analyzed for indicators of meaningful change in categories of interest (e.g., large increases/decreases in a field; occupation or immigration patterns). Frame updates could be implemented whenever the ACS data signaled there had been a significant change that would warrant an update.

NSF sought advice about the strengths and weaknesses of each of these options, combinations of these options, the frequency of utilizing such options, and suggestions and reviews of any other options for using the ACS as a frame for the NSCG, both during a potential transition period as use of the ACS was being phased in, and on a longer-term basis. NSF welcomed suggestions or recommendations about any changes in the current design of using the coordinated SESTAT surveys to achieve coverage of the population of scientists and engineers that might be possible with the availability of the ACS as well as advice about any potential pitfalls or problems in using the ACS as a frame.

Sample Size

Given that the ACS surveys 250,000 addresses a month, most uses of the ACS for an NSCG frame will require aggregating multiple months of the ACS for a frame. The largest sample needed to be drawn from the ACS at any one time would be when the entire NSCG sample was to be redrawn at one time (Options 1 and 2 above)—sample sizes of 215,000 and 171,000 in 1993 and 2003, respectively. Much smaller sample sizes from the ACS might be needed at any one time under Options 3 and 4.

The ACS annual sample size is approximately 3 million housing units that include7.8 million people.[20] In 2005, the ACS had a completion rate of 66%, which means ACS data being collected about some 2 million housing units, including 5.2 million persons, annually. NSF estimated that from this population, approximately 18.8% have a bachelor's degree or higher and are aged 75 years of age or under (SESTAT target population definitions), so approximately 978,640 cases would be eligible for the NSCG. This compares to 6.4 million cases that were eligible from the long form for the 2003 NSCG.

Based on analysis of the full-year 2005 ACS data, NCSES determined that 1 year of ACS samples (January to December) would contain enough cases to equal or surpass the size of past NSCG postcensal samples for some populations, but it was unlikely to have enough samples to equal the previous NSCG cell size for the more-rare populations (e.g., minority groups). At least 2 years of monthly samples might be necessary to provide sufficient coverage of many of these small-population groups.[21] Because, under current procedures, the U.S. Census Bureau processes the ACS monthly samples on a calendar-year basis (12 months of sample are processed together after data collection has closed), sampling for the NSCG could require 2 years of ACS data if a completely new sample is drawn. If NCSES phased in the use of the ACS (e.g., by continuing to use some of the current 2000 decennial sample until the ACS provides sufficient sample for NSCG sampling), it would be possible to use 1 year of ACS samples initially.

Timing

The schedule for processing ACS data has implications for the reference date for the NSCG—and, thus, for the other two SESTAT surveys. A full calendar year (or years) of ACS data need to be available sufficiently in advance of the NSCG reference date to allow Census time to clean and weight the ACS data so as to be usable by the U.S. Census Bureau unit that does the sampling for the NSCG and to allow for sufficient time to select and prepare the NSCG sample for the field. To have ACS frame data that are as fresh as possible at the time the NSCG goes into the field, the ACS collection year must end about 8–10 months prior to the NSCG survey reference date. A fall NSCG reference date accomplishes this, and the reference date for the 2008 and 2010 SESTAT surveys was October 1. According to the U.S. Census Bureau, the 12-

month-calendar-year ACS data are ready for use in sampling before the end of June of the following year.[22] An October SESTAT reference date allows several months to process the files, stratify the frame, select the sample, and create the mailing records.

Such a time schedule has advantages in terms of the age of the data. Typically, there have been about 3 years between the reference dates for the long form and the NSCG postcensal survey. With an October reference date and a sample based on ACS monthly samples for the previous calendar year, some of the contact data would be less than 12 months old, and none would be older than 22 months. (If 2 years of ACS sample were used, only the oldest data would be similar in age to the long-form data.) Some sample cases will have moved between the time they were sampled in the ACS and the NSCG data collection, but there will be many fewer than in the postcensal surveys.

Pooling monthly ACS samples across multiple months creates some issues in estimation and determining NSCG/SESTAT eligibility. In the past, the postcensal NSCG eligibility was based on a sample with a single reference date (the date of the decennial Census). In the ACS, each monthly sample has a different reference date. This will require the NSCG to use a different strategy for determining eligibility. For example, degrees are conferred at many points during the year. For those newly earning a bachelor's degree during a particular ACS calendar year, their eligibility for the NSCG could depend on which month they were in the ACS sample, which could be before or after receiving their degree. This issue can be addressed using domain estimation techniques. The target population can be defined as those earning a bachelor's degree before the first month of the sequence of ACS sample months pooled to create the frame. A similar approach might be considered for immigrants where the target population could be defined as those in the United States at a defined cutoff date.

Using such a procedure would result in a very small proportion of sample members being screened out as ineligible during the NSCG. Using ACS data from a calendar year and a cutoff month of the preceding December, only a small number of sample cases would have received their first bachelor's degree after December but before the ACS sample cutoff month.

Quality Improvements

Using the ACS as the sampling frame for the NSCG also provides quality-enhancing opportunities over the previous long-form approach to improve the survey in several dimensions—timeliness, accuracy, relevance, and cost. Such opportunities reflect in part the availability of the frame—and, therefore, the fielding of the NSCG—much sooner after the frame data were collected than for the long form.

Being able to draw a sample and field the NSCG closer to the time the frame data were collected can reduce costs in several ways. A shorter time period between the frame and NSCG data collection reduces the likelihood of changes in eligibility status between the two dates, such as moving abroad or earning another degree, and should improve the ability to locate individuals for participation. With a shorter time gap for all or most of the sample between the ACS frame data and the NCSG reference date, a smaller fraction of the NSCG sample cases should have moved from where they were living at the time of the ACS compared to the long-form frame. Additionally, it should be easier to locate individuals who have moved within the United States when the time they have been gone from the previous address is shorter. Such factors should reduce the cost of locating, which should cut survey costs and possibility reduce time in the field.

As mentioned before, the NSCG historically has provided the stock of scientists and engineers near the beginning of the decade, while the RCG and SDR have captured the new flows of those receiving SEH degrees during the decade after the postcensal NSCG. A person who was sampled in the NSCG (or RCG) but subsequently earned another degree (bachelor's, master's, or doctorate) in an SEH field is eligible for inclusion in the RCG or SDR by virtue of that additional degree. To keep the frames for the three surveys mutually exclusive and to eliminate the possibility of double-counting these populations, all NSCG/RCG cases involving individuals earning another eligible degree after they were originally sampled in one of the surveys are considered out-of-scope cases for the integrated SESTAT data set. Reducing the number of such sample cases that are excluded from the integrated database increases the effective sample size and thus reduces variance slightly, improving accuracy. Alternatively, the actual sample size could be reduced while maintaining the effective sample size, which could reduce survey costs somewhat.

The annual availability of the ACS as a frame provides the opportunity to update the frame throughout the decade, which can reduce or eliminate coverage problems associated with the 10 years between updates (long-form frame, or Option 1 using the ACS). In addition to the possibilities of redrawing the entire sample (or a substantial portion of the sample) periodically (Options 2 and 3), there is the potential to refresh the sample for particular groups during the decade (Option 4). For example, additional samples could be drawn for groups that assume heightened interest during the decade (e.g., post-9/11 issues, end of the dot.com era). The frame also could be refreshed during the decade to capture aspects of the flow of scientists and engineers during the decade that are not captured by the RCG and SDR, such as foreign citizens entering the United States after the time period of the ACS frame (and who did not subsequently earn a U.S. SEH degree) or new graduates with non-S&E degrees who enter S&E or S&E-related occupations. Such an approach would improve coverage of the target population, increasing both the accuracy and the relevance of the SESTAT data.

Adding a Field of Degree Question to the ACS

While the ACS provides opportunities for the improvement of coverage of the SESTAT target population, there is one area where the ACS is no different than the decennial Census in that it does not collect any information on the field of an individual's college degree(s). The SESTAT target population has two components (S&E or S&E-related degree OR S&E or S&E-related occupation), but it was only possible to use the latter as a sampling characteristic for the NSCG from the long form, which resulted in a sample with many non-SESTAT-eligible cases. Adding an item to the ACS related to the field of a person's college degree would greatly increase the efficiency of the ACS as a sample frame for the NSCG, as then the first part of the SESTAT population definition could be used as a sampling characteristic, and the ACS would contain data on both dimensions that determine whether an individual meets the SESTAT criteria for being classified as a scientist or engineer.

NCSES began discussing the issue of adding a field of degree (FOD) item to the ACS several years ago and has worked with other agencies, the Office of Management and Budget (OMB), the U.S. Census Bureau, and congressional staff on this process. A variety of question formats and the content of an FOD item were discussed, investigated, and tested. A new FOD item

would immediately follow the educational attainment question on the ACS and would only be asked for those whose highest level of educational attainment was a bachelor's degree or higher. Such an item could ask about FOD for one or more degrees. The test for the SESTAT definition on educational background is conducted by looking at the entire educational history (all degrees held) to determine if one or more of them are in S&E or S&E-related degrees. In 2003, approximately 40% of the NSCG reported having one or more degrees at the bachelor's level or higher. Ideally, for the purpose of NSCG sampling, ACS would collect information about the FOD for all degrees.

After initial discussions with the U.S. Census Bureau, it became evident that there would only be room on the ACS questionnaire for a question on a single degree. Therefore, it was necessary to identify a single degree that could be collected, and NSF recommended that such an item focus on the FOD for an individual's bachelor's degree, not the most obvious approach given that the ACS educational attainment question requests information on each individual's highest degree of attainment. It might have been easier to implement an FOD question related to a person's highest degree. However, this would cause significant coverage problems for the NSCG. In the 2003 NSCG population, 17% of individuals whose first bachelor's degree was in an S&E or S&E-related field reported that their highest degree was in a non-S&E field. Only a small proportion (4%) of non-S&E bachelor's-degree holders reported that their highest degree was in an S&E or S&E-related field. By asking for the field of bachelor's degree instead of field of highest degree, fewer NSCG (and SESTAT) coverage problems were likely to result. Thus, for stratification purposes for sampling, asking about a person's bachelor's degree is the best choice if FOD for only one degree is possible. The sampling efficiency for SESTAT-eligible cases would be greater (i.e., there would be fewer cases sampled that were not SESTAT eligible) asking about FOD of the bachelor's degree rather than of the highest degree. Finally, for the purpose of analysis of the degree holders in any field, not just in S&E or related fields, data about the same degree level for all sample cases would likely be useful.

NSF recommended that the content of the FOD item collect information specifically about all degree fields for the bachelor's degree and not just those in S&E or related fields. Gathering information about all degree fields would make the information much richer for analytical purposes for a wide variety of users and would improve its value for NSF purposes as well.

Such data can be used to compare patterns for those in S&E and S&E-related fields with those in non-S&E fields. Respondent accuracy in reporting degrees in S&E and S&E-related fields is likely to be better if there are specific categories for non-S&E fields, with examples, rather than simply a list of S&E or S&E-related categories and then a residual category labeled as "other" or "non-S&E."

Beginning in 2006, NCSES worked with the U.S. Census Bureau as well as two groups of academic researchers to develop and test alternative formats of an FOD question. Based on the preliminary research, two alternative formats of the question were developed and tested in the 2007 ACS Methods Test, completed in fall 2007. The result of the 2007 ACS Methods Test was to use an open ended question for FOD, which would be coded by Census. Details on the two question formats are provided later in this paper.

The evaluation of the 2007 ACS Methods Test for FOD did not reveal major problems with the FOD items. OMB approved adding an open ended question to the ACS in 2009[23]

NSCG Sampling with an ACS FOD Item

The addition of an FOD question on the ACS would affect the cost and efficiency of any of the NSCG design options outlined earlier for using the ACS as a sample frame. For example, regardless of the option, because it would be possible with the FOD item to mimic more closely the SESTAT target population with respect to educational background, the oversampling needed to find sufficient SESTAT-eligible cases in non-S&E occupational groups could be reduced substantially.

The potential for cost reductions and efficiency improvements throughout the decade will depend on how often and how extensively the ACS frame is used for drawing samples for the NSCG, on the format used to collect the FOD data, and on the accuracy of the FOD data. The sample size needed to obtain efficiency similar to the postcensal surveys in the past was not determined precisely until decisions were made and testing provided information on the quality of the FOD information collected on the ACS. However, it was reasonable to presume that the sample size of a once-a-decade sample could be cut substantially. Alternatively, the sample size could be maintained (or reduced somewhat less) to yield a larger in-scope sample, allowing

better coverage and the ability to report for rare populations or small domains, such as race/ethnicity or sex in S&E occupations.

Major improvements that could result from having an FOD question to use for sampling for the NSCG, regardless of which option(s) for sampling were chosen, included the following:

1. More efficient screening could result. The ratio of NSCG cases sampled to SESTAT-eligible respondents could decline considerably from the 2003 rate of 2.6 to 1.0; the reduction would depend on the sampling strategy chosen, the question format, and the response accuracy.

2. Improved efficiency of the sample provides the opportunity to maximize the return on a set sample size by better targeting more of the available sample for key groups (e.g., women, minorities, and persons with disabilities), which would help SESTAT to improve estimates for these populations.

3. A smaller, more efficient sample could allow for time and resources to be spent on quality improvements for the survey.

4. A smaller sample size and fewer respondents would mean a reduced overall respondent burden in terms of burden hours.

Even with an FOD item, some level of screening of cases drawn from the ACS is necessary because the combination of the FOD question and occupation does not fully identify all SESTAT-eligible cases. For example, the FOD for the bachelor's degree does not allow for the identification of non-S&E bachelor's-degree holders with non-S&E occupations who have an S&E or S&E-related degree at the master's-degree level. Additionally, there may be some accuracy in reporting of the degree field or occupation (either type 1 or type 2 errors) on the ACS that could be verified with the NSCG follow-up survey. Furthermore, there is value to NSF to collect some data periodically for comparison purposes on those who are not scientists and engineers.

The use of an ACS sampling frame provides ample opportunities for variation in the NSCG survey design. Having an FOD question would enhance all four options for utilizing the ACS as a frame for the NCSG. The impact on each of the four options is discussed below.

1. **Continue the current approach, refreshing the sample once a decade**
   With an FOD item, the sample size needed to produce a yield of SESTAT-eligible cases equivalent to that stemming from the postcensal NSCG surveys in the past several decades could be much smaller. The ACS frame could be used once per decade, with a much smaller sample, but it would not have to be limited to the decennial Census time frame.

2. **Update the entire sample more than once a decade**
   Redrawing the sample more frequently than once a decade would involve greater costs than Option 1, but the difference in costs would be much less with an FOD item than without it. Some screening would be necessary every time a new ACS sample was drawn because of measurement error (both FOD and occupation).

3. **Rotating sample approach**
   A rotating sample approach maintains some of the longitudinal aspects of the historical design. Similar to Option 2, the addition of an FOD item would reduce the size of each new sample panel from ACS that would be needed to obtain the desired number of SESTAT-eligible cases.

4. **Selective updates**
   Addition of the FOD item could facilitate this option for utilizing the ACS to the extent that the target group of interest is related to the field of a person's bachelor's degree. For example, during the rise and fall of the technology/IT firms, it would have been of great interest to compare the employment patterns of those in IT occupations with differing degree backgrounds, such as computer science, engineering, or non-S&E fields.

Based upon the recommendation by CNSTAT after presenting the four options, NCSES chose option 3.

Technical Issues for Sampling from ACS

There are several issues that impact the use of the ACS for the NSCG sample design: the form of the FOD question and how swapped and imputed data for educational attainment and FOD are assigned in the ACS. Each is discussed below.

1. Form of the FOD question

The efficiency, attractiveness, and costs of the various designs for the NSCG depend on the form of the FOD question. The more detailed the FOD information available for sampling, the better samples can be allocated to domains of interest. The U.S. Census Bureau (Rothgeb and Beck 2007), Don Dillman (Washington State University; see Dillman, Mahon-Haft, et al. 2006; Dillman, Mahon-Haft, and Wright 2006), and Jon Krosnick (Stanford University; see Cobb, Krosnick, and Bannon 2006) conducted a series of experiments that led to the development of two versions of FOD items that were tested on the 2007 ACS Methods Test—one is a categorical question, and the other is an open-ended version. Each is shown below.

Categorical version (FOD1)

This question focuses on this person's BACHELOR'S DEGREE. In which of the following major fields did this person receive his/her BACHELOR'S DEGREE(S)?
Mark (X) "Yes" or "No" box for each category.

		Yes	No
a.	Biological, Agricultural, Physical, Earth, or Other Natural Sciences	I__I	I__I
b.	Health, Nursing, or Medical Fields	I__I	I__I
c.	Engineering, Computer Sciences, or Mathematical Sciences	I__I	I__I
d.	History, Arts, or Humanities	I__I	I__I
e.	Psychology, Economics, or Other Social Sciences	I__I	I__I
f.	Business or Management	I__I	I__I
g.	Education or Education Administration	I__I	I__I
h.	Some other major field – Specify	I__I	I__I

_____

_____

Open-ended version (FOD2)

This question focuses on this person's BACHELOR'S DEGREE. Please print below the specific major(s) of any BACHELOR'S DEGREES this person has received. *(For example: chemical engineering, elementary teacher education, organizational psychology.)*

_____

_____

_____

There were issues to be resolved regardless of the version of the question that was chosen. Some of these issues affect the version that was chosen; others affect the use of the data for sampling or analysis. Table 4 lists some of the issues of concern.

TABLE 4. Issues of concern regarding the two FOD question versions

Issue	Categorical version	Open-ended version
Space on ACS	Requires approximately one-third of a column.	Requires less than one-fourth of a column.
Post-data collection coding	Only limited post-data collection coding will be needed. Non-sampling error could be added by incorrect coding of the other – *specify* item.	Extensive, ongoing coding will be required, which may delay data processing and final delivery time. Non-sampling error could be added by incorrect coding of the open-ended response(s).
Number of fields available for analysis and for use in NSCG design	Limited to 7 + 1 residual category.	It may be possible to develop a more extensive list of fields (for analysis and NSCG sampling) than with the categorical version, but this will greatly depend on the level of detail provided by the respondents.[a]
Reporting of multiple bachelor's degrees	This version of the question explicitly allows for the reporting of multiple bachelor's degrees.	The question stem indicates that more than one degree may be reported, but it is not clear that respondents will do so.
Type 1 errors	Reducing type 1 errors is a concern for this question (e.g., checking or writing having an S&E or S&E-related bachelor's degree when a non-S&E FOD is appropriate). This will have the greatest impact for NSCG sampling due to unnecessarily sampling a case that does not have a required degree. However, this can easily be resolved during the NSCG data collection, where the case can be identified as ineligible. It causes a larger problem for analysis of ACS because there is no other information on the ACS to validate the FOD.	
Type 2 errors	Reducing type 2 errors (e.g., not reporting an S&E or S&E-related FOD while actually having one) is a greater concern than type 1 errors because, in this situation, there will be population undercoverage for the NSCG. Additionally, as with the type 1 errors, it will cause problems for direct analysis of the ACS FOD data.	

[a] The initial basis for the development of the coding was the code list and coding procedure utilized for the 2003 SESTAT surveys. The field of degree (FOD) code list for the SESTAT surveys had 144 FOD codes. That list, however, had far more detail for S&E and S&E-related degree fields than for non-S&E fields. For the full evaluation of the ACS FOD data, a greater level of detail for non-S&E fields was desirable. NSF worked with the U.S. Census Bureau and other agencies to develop such a list.

A sampling design for NCSG using the ACS with FOD could be crafted to produce little or no undercoverage in an initial sample drawn from the ACS. The form of the FOD question and the accuracy of the information provided impacts the gains in efficiency. For example, how accurate will the reports on the FOD item be for those reporting for others in the household (proxy reports) compared to those reporting for themselves? If the FOD and occupation items can be used to accurately distinguish scientists and engineers from other college graduates, substantial gains in efficiency were possible. For NSCG sampling purposes, the most important concern was whether a degree is accurately reported as falling into an S&E, an S&E-related, or a non-S&E category.[24]

The accuracy of the FOD reporting will be evaluated after the first NSCG is conducted using the ACS. The information from the detailed education history collected as part of the NSCG from the individual (where there are no proxy reporters) can be compared to the information reported on FOD (and educational attainment) in the ACS. Analysis of the reinterviews in the 2007 Methods Panel testing of the two FOD items showed that responses to both versions of the FOD item were reliable and valid. However, validity on the full ACS sample is an open question.

Some number of cases apparently not meeting the criteria of being a scientist or engineer (a non-S&E bachelor's degree and a non-S&E occupation) would be drawn in the NSCG sample from the ACS frame both to provide a comparison group and to account for those in non-S&E occupations with a non-S&E bachelor's degree but an S&E or S&E-related degree at a higher level. It was advisable in drawing the first NSCG sample from the ACS to allocate part of the sample to test the efficiency of the FOD item for sampling purposes, either drawing a larger number of apparently non-S&E cases that might be done otherwise or drawing a portion of it using the long-form procedures without taking the FOD information into account.

2. Swapped and imputed data for education level and FOD

The U.S. Census Bureau regularly uses a technique called *swapping data* to create public-use data sets (a decision based on the Bureau's overall disclosure policies). Swapping is done during the survey data processing. NCSES requested that the edited ACS file, before swapping, be used for weighting and creation of the NSCG sampling frame. Using swapped data would greatly reduce the stratification efficiency, especially when disproportionate stratified sampling is used to target precision levels for selected domains. The U.S. Census Bureau allowed use of unswapped data from the ACS for sampling for the 2010 NSCG.

Another technical concern was the use of *imputed data* from the ACS. Imputed educational attainment level data (the U.S. Census Bureau calls them *allocated data*) should not be used for sampling. Imputed data create an unacceptable amount of undercoverage of those with a bachelor's degree (estimated at 3% to 7%; see Finamore,

Hall, and Fecso 2006) as well as sampling inefficiency (when those with an imputed education level of a bachelor's degree turn out not to have a bachelor's degree). Records that have imputed educational attainment level data were put aside prior to sampling, and a small sample of these ACS cases could be subsequently sampled to measure bias.

Adding an FOD question to the ACS could create an entirely new issue related to imputation. Given the relatively poor performance of the imputation methods for education level (the imputation performs much like the full-file, missing-at-random model), it is unclear how imputation should be done for missing FOD. For individuals with an S&E or S&E-related occupation, FOD imputation might perform well. For other occupations, it is not obvious that an acceptable imputation model can be developed. It may be that such cases will need to be treated as missing and reweighted. A program of research on imputation and nonresponse weighting for missing FOD is desirable.

Analytic Uses of ACS Data

The addition of the FOD item on the ACS (regardless of its final form) immediately presents opportunities to better understand the population with at least a bachelor's degree in the United States. However, analysis using the FOD item also present some challenges. For example, while analysts will have information on the highest level of educational attainment from the ACS, the FOD information will only be available for the bachelor's degree(s). This is not a common set of variables to have available together, so new analytic approaches might be needed to take this into account.

The ACS data could also be used both between NSCG cycles and to expand the types of analyses that could be done. For example, the ACS provides information for each member of the household. This information will be available for analysis and may be especially useful, along with the FOD information. Currently, the study of dual-career households is possible with ACS data, but only for occupation and education level. The FOD question on the ACS will improve such analyses for issues regarding the S&E workforce, especially since in-depth household information is not collected on the NSCG. The ACS data could also be utilized to analyze the impact of real-time events on the college-educated population. A supplemental sample to the NSCG could be considered using ACS sample households to provide deeper

information on dual-career issues. Other options for analysis or indicators of the S&E workforce trends would be possible if the ACS data were to become available on a monthly basis. Then it is might be possible to create a monthly time series for key S&E data at very aggregate levels.

The ACS could also be used in other ways to improve timeliness and relevance of the S&E workforce information. For example, the ACS could provide a frame to do more frequent (and smaller) special-topic surveys. There is also the potential to address a topical interest directly in the ACS. Adding supplemental questions (questions that appear one time or that rotate in and out of use) on subjects of current interest or to help identify special population groups could be added where a model adopted for the ACS of a core set of questions asked every year with room for the addition of other questions on a one-time or periodic basis.

Conclusion

The ACS provides the opportunity for a variety of changes to the NSCG and SESTAT in general. Some of the issues NSF must address in implementation of the rotating sample include the following:

- The issues for implementing use of ACS as a sample frame for the NSCG in the long term under option 3
- The need/advisability of phasing in use of the ACS and options for doing so during a transition period
- The frequency of using the ACS to refresh the NSCG sample
- Possibilities for other changes in the current system of the three SESTAT surveys
- Use of ACS data to supplement (or as a substitute for) data from the SESTAT surveys

These topics must be considered together rather than as a series of separate, discrete topics to consider their joint interactions and effects on survey design and operations, the ability to meet user needs, and the types of analysis that can be conducted. Similarly, considerations of cost, timeliness, and data quality must all be taken into consideration. The option that is most appealing from a data quality perspective may have unacceptable cost implications.

One important issue is the value of maintaining the longitudinal feature of the NSCG. In the past, the longitudinal feature of the NSCG (and the SDR) has not been utilized extensively. The

current design provides longitudinal data as a by-product of sample generation, but cross-sectional designs would lose that feature.

References

Cobb C, Krosnick J, Bannon B. 2006. Experimental tests of questions measuring field of degree: Surveys of Stanford and University of Virginia alumni. Memo to the Division of Science Resources Statistics from the authors.

Dillman D, Mahon-Haft T, Cook S, Wright K. 2006. Cognitive evaluations of alternative questions for determining field of study for bachelors degree(s) as a follow-up to highest degree question in the American Community Survey. Technical Report 06-037. Pullman, WA: Social and Economic Sciences Research Center, Washington State University.

Dillman D, Mahon-Haft T, Wright K. 2006. Effects of question structure on answers to field of study question proposed for the American Community Survey: A follow-up test. Technical Report 06-041. Pullman, WA: Social and Economic Sciences Research Center, Washington State University.

Fecso RS, Baskin R, Chu A, Gray C, Kalton G, Phelps R. 2007. *Design Options for SESTAT for the Current Decade: Statistical Issues*. Working Paper SRS 07-201. Arlington, VA: Division of Science Resources Statistics, National Science Foundation.

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

Finamore JM, Hall DW, Fecso RS. 2006. Dual frame estimation in the 2003 National Survey of College Graduates. *Proceedings for the Section on Survey Research Methods, American Statistical Association*, pp. 3008–15.

National Research Council (NRC). 1989. *Surveying the Nation's Scientists and Engineers: A Data System for the 1990s*. Panel to Study the NSF Scientific and Technical Personnel Data System, Committee on National Statistics. Washington, DC: National Academies Press.

National Research Council (NRC). 2003. *Improving the Design of the Scientists and Engineers Statistical Data System (SESTAT)*. Committee to Review the 2000 Decade Design for the Scientists and Engineers Statistical Data System (SESTAT). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Rothgeb J, Beck J. 2007. Cognitive interviewing research results and recommendations for the National Science Foundation's proposed field of degree question for the American Community Survey. Research Report Series Survey Methodology #2007-15. Washington DC: Center for Survey Methods Research, Statistical Research Division, U.S. Census Bureau.

Notes

1. Please contact Stephen Cohen (scohen@nsf.gov), Chief Statistician, National Center for Science and Engineering Statistics (NCSES), regarding questions on this paper. The initial draft was prepared by Ronald S. Fecso, previously Chief Statistician, and currently Chief Statistician of the U.S. Government Accountability Office (GAO). He began work on this paper at the National Science Foundation (NSF); with the approval of GAO, further work was done after he left NSF. Further revisions to the paper were made by Mary J. Frase and Nirmala Kannankutty.

2. Beginning in 2003, the coverage of the SESTAT target population was expanded to include S&E-related degrees or occupations. A major component of this population expansion was in health fields and occupations. The Survey of Doctorate Recipients (SDR) had always included those with doctorates in health fields, and these cases had been included in the integrated database. Beginning with 2003, the National Survey of Recent College Graduates (RCG) population was expanded to include recent U.S. bachelor's- and master's-degree earners in health fields, and these individuals were included in the integrated database. Individuals with health degrees or occupations were also captured in the 2003 NSCG and included in the

SESTAT database (as were individuals with other S&E-related degrees or occupations). A detailed description of S&E, S&E-related, and non-S&E fields and occupations can be found at http://sestat.nsf.gov/docs/ed03maj.html and http://sestat.nsf.gov/docs/occ03maj.html, respectively.

3. Information on each of the surveys can be found at: http://www.nsf.gov/statistics/srvygrads (NSCG); http://www.nsf.gov/statistics/srvyrecentgrads (RCG); and http://www.nsf.gov/statistics/srvydoctoratework (SDR). Information on SESTAT can be found at http://sestat.nsf.gov.

4. To resolve this issue, SESTAT has developed a statistical integration process, employing a unique linkage rule. Each survey is weighted according to the frame developed for that survey. Additionally, a series of overlap variables are calculated that allows for the identification of cases that are eligible for more than one survey. To remove these multiple selection opportunities, each case within the SESTAT target population is uniquely linked to one and only one component survey, and that individual is included in the SESTAT integrated file only when he or she is selected for that linked survey.

5. There has not been a significant amount of longitudinal analysis conducted with the SESTAT surveys. Longitudinal weights have only recently been created and have not yet been made available to non-SRS users. However, there is an interest in the user community in conducting longitudinal analysis, much of which is best done with longitudinal weights.

6. Detailed design information about the 2003 NSCG can be found in the 2003 NSCG sample design report and the 2003 NSCG methodology report, available from NCSES on request.

7. The redesign was largely based on a Committee on National Statistics (CNSTAT) report (NRC 1989).

8. Although referred to as “respondents,” in actuality the response for the individual may be provided by another household member.

9. See the 2003 NSCG questionnaire (http://www.nsf.gov/statistics/srvygrads/survey2003/grads_2003.pdf). The educational history grid can be found in Section A of the questionnaire.

10. This section draws heavily on a review of potential sampling frames for SESTAT done prior to the design of the 2003 NSCG. See Fecso, Choudry, et al. (2007).

11. This number excludes those who graduated in non-S&E fields after April 1, 2000, who were working in S&E or S&E-related occupations in 2003 as well as those with only foreign degrees who were not in the United States at the time of the decennial Census but were here working in an S&E or S&E-related occupation at the time of the 2003 NSCG.

12. For example, see the National Research Council (1989, 2003).

13. Information about the ACS can be found at http://www.census.gov/acs/www/.

14. The 2008 ACS questionnaire can be found at http://www.census.gov/acs/www/SBasics/SQuest/SQuest1.htm.

15. No address will be included in the ACS sample more than once in a 5-year period.

16. Beginning in 2006, this information was made available annually in late summer/early fall for the previous year's sample.

17. While *small-area* estimates are generally thought of as a physical location (often, as in the ACS case, government jurisdictions, such as counties), estimates for *small/rare* subpopulations (small domain estimates), such as scientists and engineers, are the equivalent statistical problem relative to sample size. NSF needs to use the ACS as the NSCG sample frame not because it wishes to produce small-area estimates but because it needs the ACS sample size for rare populations.

18. Currently, those with U.S. doctorates contained in the NSCG are not included in the SESTAT integrated database. U.S. doctorates are drawn from the SDR survey. However, sample cases in the NSCG who have doctorates from institutions outside the U.S. are included in the SESTAT integrated database.

19. For a discussion of using multiple frames for the NSCG design, see Fecso, Baskin, et al. (2007).

20. This is based on an average household size of 2.6, multiplied by the 3 million housing units surveyed in the ACS. Average household size was determined from the Census's American Fact Finder with data from the ACS for 2005.

21. It is unlikely that 12 months of ACS data would be sufficient for approximately one-third of the aggregate sampling cells that NCSES tested based on analysis of the full year of 2005 ACS data. These aggregate cells combined minority groups and used fewer occupational categories than have been used in the past. Using the 2003 sampling cells, several more years of ACS sample would be required to produce sample sizes similar to those achieved with the 2003 NSCG design. The aggregate cells that NCSES tested were important because they form the basis for many of the domains for which estimates have been produced in the past and are possible to achieve with 2 years of ACS samples.

22. It is desirable to work with the ACS data much earlier for NSCG sampling than June, but later availability of the data maybe less of an obstacle if more than 1 year of ACS sample is needed. If, for example, 2 years are needed, it may be possible to sample and field the NSCG in two waves—one based on the first of the 2 ACS years, which could be processed much in advance of the survey date, and the second, fielded slightly later, based on the second ACS year. In 2006, both the RCG and SDR were fielded in two waves for similar reasons—the late availability of the frame for part of the sample.

23. With a full year of data available from the ACS (2005), NCSES can begin to work with the U.S. Census Bureau to explore the use of the ACS without the FOD item for the NSCG (and for analysis). NCSES needed to use the ACS whether or not there was an FOD degree question.

24. In the categorical version of the FOD question tested, only one set of S&E-related fields (health) can be captured accurately. In order to identify samples in other S&E-related fields, NSF had to sample some of the non-S&E FOD categories and some non-S&E occupations. For example, in order to find individuals with degrees in science or math teacher education (an S&E-related field), it was necessary to sample some individuals with bachelor's degrees in "Education or education administration" and some secondary teachers.