Scientists and Engineers Statistical Data System (SESTAT), 2010

Flora Lan
Project Officer
Human Resources Statistics Program
(703) 292-4758

Technical Notes

The Scientists and Engineers Statistical Data System (SESTAT) comprises three demographic surveys of scientists and engineers sponsored by the National Science Foundation (NSF): the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR). The three component surveys are conducted every 2–3 years and use similar questionnaires, survey reference dates, data collection periods, and data-processing procedures to facilitate integration for SESTAT. The three surveys are designed to provide maximum coverage of the target population—namely, scientists and engineers—with special emphasis given to relatively rare populations (e.g., doctorate recipients, recent graduates, and minorities). Overall, SESTAT provides a comprehensive picture of the number and characteristics of individuals in the United States with a bachelor's or higher-level degree and their employment, with a focus on those having science and engineering (S&E) degrees or working in S&E occupations. In the 2000s, this definition was expanded to include S&E-related degrees and occupations.

Mathematica Policy Research, Incorporated, under NSF contract number NSFDACS1066103, prepared the data tables under the direction of Daniel Foley and Flora Lan, Human Resources Statistics Program, National Center for Science and Engineering Statistics, National Science Foundation. Mathematica staff members who worked on this report included David Edson, Xiao Fu, Amang Sukasih, and Donsig Jang.

Target Population and Coverage

The 2010 SESTAT target population includes individuals who had the following characteristics as of the component surveys' reference week of 1 October 2010:

- Were living in the United States (the 50 states, the District of Columbia, Puerto Rico, or another U.S. territory)
- Were living in a non- institutional setting and were not terminally ill
- Were 75 years of age or younger
- Received a bachelor's or higher-level degree in an S&E field or S&E-related field or received a bachelor's or higher-level degree in a non-S&E field but worked in an S&E or S&E-related occupation

Individuals with the following characteristics are not covered:

- Those who received a bachelor's or higher-level degree in an S&E or S&E-related field between 1 July 2009 and the reference date of the surveys (1 October 2010)
- Those who received their first bachelor's degree in a non-S&E field between 1 January 2010 and 1 October 2010 and worked in an S&E or S&E-related occupation on the survey reference date
- Those who earned all of their postsecondary degrees from foreign institutions and entered the United States after 1 January 2010

Because the 2010 SESTAT was created from three component surveys, cases identified in one component survey might also be eligible for another survey. Consequently, SESTAT uses a unique

linkage rule when integrating the component sample surveys in which each survey sample member is weighted according to the frame developed for that survey. Next, a series of overlap variables is calculated and assessed to identify 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.

Under the unique linkage rule, the vast majority of respondents with doctorates in science, engineering, or health (SEH) fields from U.S. academic institutions are SDR sample cases; those with research doctorates in other fields or those with research doctorates awarded by foreign institutions are sample cases from either the NSCG or NSRCG. In these SESTAT component surveys, doctoral-degree information is obtained by different processes. The SDR identifies only doctorate holders in SEH fields from U.S. doctoral research training programs in institutions that are eligible for participation in the Survey of Earned Doctorates (SED). In the NSCG and NSRCG, the number of individuals with a nonprofessional doctoral degree, their field of study, and their academic institution is based on respondents' self-reporting that their highest degree is a doctorate (e.g., PhD, DSc, EdD) and a verbatim response for their field of study and the name and location of the awarding academic institution. The different processes yield a more constrained definition and estimate of research doctorates for the SDR and a broader and more general definition and estimate of research doctorates for the NSCG and NSRCG.

Component Surveys

The 2010 National Survey of College Graduates

The NSCG has been conducted by the U.S. Census Bureau on behalf of NSF since 1993 and is the largest of the three component surveys, representing approximately 90% of the SESTAT target population. The NSCG is used to study the occupations and career paths of U.S. residents with a bachelor's or higher-level degree (particularly in an S&E field). The 2010 NSCG incorporated a dual-frame sample design in that 65% of its sample was selected from the 2009 American Community Survey (ACS) respondents who indicated they had a bachelor's or higher-level degree in any field of study. The remaining portion of the 2010 NSCG sample was selected from respondents to the 2008 NSCG. Prior to 2010, the NSCG was designed as a decade-long panel study of college graduates based on a sample of respondents from each decennial census long-form sample. For the portion of the 2010 NSCG sample from the 2009 ACS, the NSCG questionnaire collected the respondent's full postsecondary educational history, birth year, sex, race, and ethnicity. The questionnaire for the portion of the 2010 NSCG sample from the 2008 NSCG asked these questions in their 2003 NSCG baseline survey cycle. The total sample size for the 2010 NSCG was 100,488, and the weighted response rate was 77.8%.

The 2010 National Survey of Recent College Graduates

The NSRCG has been conducted since 1974 and provides data on recent graduates. The NSRCG is a cross-sectional survey that is used to study the continuing graduate education of recent bachelor’s and master’s graduates, their early employment experiences, and the attributes of their employment, particularly in finding work in their field of study.

The 2010 NSRCG consisted of individuals who recently received bachelor's or master's degrees in SEH fields from a U.S. college or university within the preceding 2 academic years (defined as July 2007–June 2008 and July 2008–June 2009). The 2010 NSRCG used a two-stage sample design. In the first stage, a stratified, nationally representative sample of 302 colleges and universities from a universe of approximately 2,200 U.S. academic institutions was asked to provide lists of their graduates for sampling. In the second stage, graduates with bachelor's or master's degrees in SEH fields were identified and included in the 2010 NSRCG sampling frame. Of the 302 sampled institutions in the first stage, 290

provided lists of their graduates for sampling, representing a weighted response rate of 95.7%. Data collection in the second stage resulted in a weighted response rate of 72.6%.

The 2010 Survey of Doctorate Recipients

The SDR has been conducted since 1973. The SDR is a representative panel study of S&E doctorate recipients from U.S. academic institutions. The SDR is used to study the career paths of this highly trained cohort of scientists and engineers. Recipients of professional degrees—such as those awarded in medicine, law, or education—are not included in the SDR. The 2010 SDR consisted of doctorate recipients with a degree in an SEH field from U.S. academic institutions between 1 January 1948 and 30 June 2009. Baseline data on education and demographic characteristics among SDR sampled members come from the SED, an annual census of research doctorates earned in the United States (http://www.nsf.gov/statistics/srvydoctorates/). The annual SED provides a sampling frame for updating the SDR panel over time with a supplemental sample of new U.S. SEH doctorate recipients added into each survey cycle.

The 2010 SDR included two sample components: the National Survey of Doctorate Recipients (NSDR), which includes U.S.-degreed doctorate recipients predicted to be living in the United States after graduation, and the International Survey of Doctorate Recipients (ISDR), which includes U.S.-degreed doctorate recipients predicted to be living outside the United States. The 2010 NSDR sample consisted of 40,000 cases selected systematically across strata, including 36,543 from the returning cohort and 3,457 from the new cohort. The overall weighted response rate was 79.9%. The 2010 ISDR sample consisted of 5,697 cases, including 4,797 from the returning panel and 900 from the new cohort. The overall weighted response rate was 75.4%. Only the NSDR respondents are included in SESTAT.

Editing Guidelines and Procedures

Because the three SESTAT component surveys typically are conducted by different survey data collection contractors, NSF uses standardized guidelines for quality assurance in data editing and data processing. Multiple coding procedures are involved in processing a unit response in terms of a respondent's occupation, education, "Other (specify)" verbatim responses, geographic coding, and educational institution coding. In addition, several questionnaire items are deemed critical data elements—such as residence information, employment status, and type of occupation if employed—and must be completed by the respondent to be considered an acceptable unit response. Quality assurance guidelines also address editing rules for "refused," "don't know," or blank or missing responses and for ensuring proper skip patterns.

Imputation

If necessary, telephone follow-ups are used to obtain answers to critical data elements (required for an acceptable unit response) and other noncritical but important items, such as degree information and employer location. Except for items with verbatim responses, missing data for noncritical items are imputed. Imputation does not begin until after all logical editing is complete. Sequential hot-deck imputation is used for missing data. Before imputation, serpentine sorting is used to ensure that adjacent data records are as similar as possible. After imputation of the data, postimputation edit checks are used to ensure that imputed values remain consistent with nonmissing data and adhere to the editing guidelines and procedures described above.

Sample Weights

Sample selection probabilities for SESTAT component surveys vary substantially and reflect the differential sampling rates used to create sufficient sample sizes to produce reliable estimates of domains of interest in each survey's target population. For SESTAT data, sampling weights are developed for

respondents in each component survey and for the combined and integrated SESTAT. For each component survey, sampling weights are adjusted for the differential selection probabilities and also for nonresponse and undercoverage. The fully adjusted sampling weights become the analysis weights and are added to each respondent record in SESTAT (variable name: Z_WEIGHTING_FACTOR_SURVEY). These weights should be used only when making estimates from each component survey in SESTAT. For the three combined component surveys, sampling weights are adjusted further for cross-survey multiplicity when estimates are made based on SESTAT. The integrated SESTAT weight (variable name: Z_WEIGHTING_FACTOR) should be used when making estimates for the overall target population.

Reliability of Estimates

Sampling Errors

Estimates are subject to sampling errors because SESTAT comprises three sample surveys. To measure the precision of the SESTAT estimates, standard errors (SEs) were calculated and provided for each estimate reported in the data tables. The SEs can be used to construct confidence intervals for the estimates. To construct a 95% confidence interval about an estimate, multiply the SE of an estimate by a z-score of 1.96. Add the result to the estimate to establish the upper bound of the confidence interval, and subtract it from the estimate to establish the lower bound of the confidence interval.

Nonsampling Errors

Quality assurance procedures are included throughout the various stages of data collection and data processing to reduce *nonsampling errors*, including (1) *nonresponse error*, which arises when the characteristics of respondents differ systematically from nonrespondents; (2) *measurement error*, which arises when the variables of interest cannot be measured precisely; (3) *coverage error*, which arises when some members of the target population are excluded from the frame and thus do not have a chance to be selected for the sample; (4) *respondent error*, which occurs when respondents provide incorrect data; and (5) *processing error*, which can arise at the point of data editing, coding, or data entry. The analyst should be aware of potential nonsampling errors, but these errors are more difficult to detect and quantify than sampling errors.

Definitions and Explanations

Disability. The SESTAT component surveys ask the degree of difficulty—none, slight, moderate, severe, unable to do—an individual has in seeing (with glasses or contact lenses); hearing (with a hearing aid); walking without assistance; lifting 10 pounds; or concentrating, remembering, or making decisions. Those respondents who answered "moderate," "severe," or "unable to do" for any activity were classified as having a disability.

Education data. These data were derived from responses to several questions on type of degree and field of study earned by the respondent. The education categories of respondents in the SESTAT data tables were based on the respondents' field of study for their highest degree held in the reference week. The SESTAT surveys collect the respondent's full degree inventory. Additional information on the degree inventory can be found at http://sestat.nsf.gov/docs/inventory.html. The list of SESTAT education categories and the aggregation into minor and major groups reported in the tables are shown in technical table T-1.

Employment sector. Sector of employment is a derived variable based on responses to multiple survey questions. In the data tables, the category 4-year educational institution includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), and university-affiliated research institutes. Other educational institution includes 2-year colleges, community colleges, technical institutes, precollege institutions, and other educational institutions (which respondents wrote

verbatim in the survey questionnaire). Users should note that, prior to 2008, these other educational institutions that were written verbatim by respondents were grouped with 4-year educational institutions rather than with 2-year colleges. The category business or industry includes self-employed individuals, nonprofit organizations, and other unspecified types of employers.

Occupation data. These data were derived from responses to several questions on the kind of work performed by the respondent in his or her principal job. The occupational classification (i.e., job code) of the respondent was based on his or her principal job (including job title) held during the reference week or on the last job held, if the respondent was previously employed but not employed in the reference week. The list of SESTAT occupation codes and the aggregation into minor and major groups reported in the tables are shown in technical table T-2.

Race and ethnicity. Ethnicity is defined as Hispanic or Latino or not Hispanic or Latino. Values for those selecting a single race include American Indian or Alaska Native, Asian, black or African American, Native Hawaiian or Other Pacific Islander, and white. Those persons who report more than one race and who are not of Hispanic or Latino ethnicity also have a separate value.

Salary. Median annual salaries are reported for the principal job and are rounded to the nearest $1,000. All respondents are asked to report annual salaries, even if their annual salary is provided for less than 12 months.