Scientists and Engineers Statistical Data System (SESTAT), 2013

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

Technical Notes

Overview

Purpose. The Scientists and Engineers Statistical Data System (SESTAT) provides a comprehensive picture of the number and characteristics of individuals in the United States with a bachelor's or higher-level degree, with a focus on those having science and engineering (S&E) degrees or working in S&E occupations. In 1993, when SESTAT was developed, the definition of scientists and engineers included only those with S&E degrees and S&E occupations. In 2003 and subsequent years, the definition of scientists and engineers was expanded to include those with S&E-related degrees and S&E-related occupations. (See technical tables A-1 and A-2 for more information on what is included in these broad categories.)

Prior to 2013, SESTAT was comprised of three demographic surveys from the National Science Foundation's National Center for Science and Engineering Statistics (NSF's NCSES): the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR). Starting in 2013, SESTAT included data from only SDR and NSCG and not NSRCG. NSRCG was discontinued after the 2010 survey cycle given the availability of the American Community Survey (ACS) as a sampling frame for NSCG. ACS was used to add a large number of younger graduates to the NSCG sample, which then offset the need to conduct the more costly NSRCG.

Data collection authority. SESTAT is an integrated data system from SDR and NSCG. The information from SDR and NSCG is solicited under the authority of the National Science Foundation Act of 1950, as amended.

SESTAT contractor. Mathematica Policy Research, Incorporated.

SESTAT sponsor. The National Science Foundation.

Key Information

Frequency. Every 2 years.

Initial year of SESTAT. 1993

Reference period. 1 February 2013

Response unit. Individual

Sample or census. Integrated sample

Population size. 28,950,000

Sample size. 115,152

SESTAT Design

Target population. The 2013 SESTAT target population consists of individuals who earned a bachelor's degree or higher before 1 January 2012 and have the following characteristics as of the survey reference date (1 February 2013):

- Lives in the United States (50 states, the District of Columbia, Puerto Rico, or other U.S. territories)
- Is not institutionalized or terminally ill
- Is less than 76 years of age
- Has earned a bachelor's degree or higher in an S&E field or an S&E-related field before 1 February 2013 or was working in an S&E or S&E-related occupation during the week of 1 February 2013

SESTAT integration. Given that the 2013 SESTAT is comprised of two component surveys, cases identified in one component survey might also be eligible for the other survey. SESTAT uses a unique linkage rule to integrate the two component sample surveys.

Under the unique linkage rule, respondents with doctorates in science, engineering, or health (SEH) fields from U.S. academic institutions who are identified through the NCSES Survey of Earned Doctorates (SED) are SDR sample cases; those with research doctorates in other fields or those with research doctorates awarded by foreign institutions are NSCG sample cases. Consequently, the SESTAT estimate of U.S.-trained research doctorate holders based on the SDR sample is more aligned to the "true" population of U.S. trained research doctorate holders than the NSCG estimate of research doctorates that is not linked to SED eligibility criteria.

Data Collection and Processing Methods

Data collection. The 2013 SESTAT is an integration of data from two NCSES surveys: NSCG and SDR.

The 2013 National Survey of College Graduates. NSCG has been conducted by the U.S. Census Bureau on behalf of NSF since 1993 and is the larger of the two component surveys, representing approximately 97% of the SESTAT target population. 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). In 2013, approximately 60% of the NSCG sample was selected from the 2011 ACS respondents with a bachelor's or higher-level degree in any field of study and represented a sample size of 83,000 college graduates. The remaining 40% of the 2013 NSCG sample were respondents from the 2010 NSCG. Respondents from the 2010

NSCG were selected from the 2009 ACS (sample size of 48,000) and the 2010 NSRCG (sample size of 13,000). For first-time sample members (e.g., the 2013 NSCG sample respondents selected from the 2011 ACS), the NSCG questionnaire collected the respondent's full postsecondary educational history, birth year, sex, race, and ethnicity. The 2013 NSCG sample from the 2010 NSCG and 2010 NSRCG received these questions in their 2010 baseline survey cycle.

The 2013 Survey of Doctorate Recipients. SDR has been conducted since 1973 as a longitudinal study of SEH research doctorate recipients from U.S. academic institutions. Recipients of professional degrees in medicine, law, or education are not included in the survey. The 2013 SDR consisted of doctorate recipients between 1 July 1960 and 30 June 2011 who were age 75 or younger. The annual SED provides a sampling frame for SDR panel updates over time with a supplemental sample of new U.S. SEH doctorate recipients added into each survey cycle. Baseline data on education and demographic characteristics among SDR sample members come from SED, an annual census of research doctorates earned in the United States (http://www.nsf.gov/statistics/srvydoctorates/). SESTAT includes all of the 2013 SDR respondents who lived in the United States as of the survey reference date.

Mode. Both NSCG and SDR sample members can complete the survey using any of the three modes: self-administrated paper questionnaire (SAQ); computer-assisted telephone interview (CATI); and self-administered online questionnaire (Web).

Response rates. The total sample size for the 2013 NSCG was 144,000, and the weighted response rate was 74%. The 2013 SDR sample consisted of 40,000 cases selected systematically across strata, including 36,666 from the returning cohort and 3,334 from the new cohort. The weighted response rate for SDR was 76.4%.

Data editing. 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. 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. Consequently, these fields are the first to undergo review and data editing. In addition, quality assurance guidelines also address editing rules for "refused," "don't know," or blank or missing responses and for ensuring proper skip patterns.

Imputation. 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, post imputation edit checks are used to ensure that imputed values remain consistent with nonmissing data and adhere to the editing guidelines.

Weighting. SESTAT sampling weights are developed for respondents in each component survey and for respondents in the integrated SESTAT. For each component survey, sampling

weights are adjusted for the differential selection probabilities and also for nonresponse and undercoverage. The fully adjusted survey-specific sampling weights (variable name: Z_WEIGHTING_FACTOR_SURVEY) should be used only when making estimates from each of the two component surveys in SESTAT. For SESTAT, sampling weights are adjusted further for removing the possibility of a respondent being represented in both surveys and counted twice. The integrated SESTAT weight (variable name: Z_WEIGHTING_FACTOR) should be used when making estimates for the overall target population.

Variance estimation. SESTAT variance estimation is based on replicate weights constructed for SDR and NSCG and uses the variability across replicate estimates to approximate the variance of the estimate. Please consult the methodology report for details.

Survey Quality Measures

Sampling error. Estimates are subject to sampling errors because SESTAT comprises two sample surveys. Standard errors were calculated and provided for each estimate reported in the data tables and can be used to construct confidence intervals for the estimates. To construct a 95% confidence interval for an estimate, multiply the standard error 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.

Coverage error. SESTAT data are subject to coverage error, which occurs when a group of individuals from the target population are left out of the frame.

Nonresponse error. SESTAT data are subject to nonresponse error, which occurs when characteristics of the respondents differ systematically from nonrespondents.

Measurement error. SESTAT data are subject to measurement error which arises when variables of interest cannot be precisely measured.

Data Comparability (Changes)

Changes in SESTAT coverage or population

- 2013. A major change in SESTAT is that the population of recent college graduates is no longer surveyed via the NSRCG. Rather, this population is surveyed as part of the new cohort of NSCG.

- 2013. The component surveys' reference date for SESTAT changed to 1 February 2013. (Survey reference dates for the 2003, 2006, 2008, and 2010 survey rounds of SESTAT are 1 October 2003, 1 April 2006, 1 October 2008, and 1 October 2010, respectively.)

Changes in questionnaire

Minor changes were made in the 2013 SDR and NSCG questionnaire, and these changes are detailed below.

- 2013 SDR questionnaires. More information is available at http://www.nsf.gov/statistics/srvydoctoratework/surveys/srvydoctoratework_nat2013.pdf.

- 2013 NSCG questionnaires for new respondents. More information is available at http://www.nsf.gov/statistics/srvygrads/surveys/srvygrads-newrespond2013.pdf.

- NSCG questionnaire for returning respondents. More information is available at http://www.nsf.gov/statistics/srvygrads/surveys/srvygrads-returnrespond2013.pdf.

Changes in reporting procedures or classification. Not applicable.

Definitions

- *Disability.* The data are derived from responses to the degree of difficulty questions—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.* The data are derived from responses to several questions on type of degree and field of study earned by the respondent. Although SESTAT component surveys collect the respondent's full degree history, the education categories of respondents listed in the SESTAT data tables were based on the respondents' field of study for their highest degree held in the reference week. The list of SESTAT education categories and the aggregation into minor and major fields of study reported in the tables are shown in technical table A-1.

- *Employment sector*. The data are derived from responses to multiple survey questions regarding type of principal employer and type of educational institution if applicable. 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. The category “Other educational institution” includes 2-year colleges, community colleges, technical institutes, precollege institutions, and other educational institutions provided as verbatim responses. Users should note that prior to 2008, the 4-year educational institution category included the verbatim responses for other educational institutions. The category business or industry includes self-employed individuals, nonprofit organizations, and other unspecified types of employers.

- *Occupation data.* The data are derived from responses to multiple survey 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 A-2.

- *Race and ethnicity.* The data are derived from responses to ethnicity and race survey questions. Ethnicity is defined as being Hispanic or Latino or not Hispanic or Latino. Categories for race include American Indian or Alaska Native, Asian, black or African American, Native Hawaiian or Other Pacific Islander, and white. Persons who report more than one race and who are not of Hispanic or Latino ethnicity also are categorized separately.

- *Salary.* The data are derived from responses to the basic annual salary question for their principal job, even if their annual salary is provided for less than 12 months. In the data tables, median annual salaries are reported by principal job categories and are rounded to the nearest $1,000.