Characteristics of Doctoral Scientists and Engineers in the United States: 2010

Lynn Milan,
Project Officer
Human Resources Statistics Program
(703) 292-2275

Technical Notes

Survey Overview

The Survey of Doctorate Recipients (SDR) is a panel study conducted every 2 years on a nationally representative cohort of individuals who have received a research doctorate in a science, engineering, or health (SEH) field. The National Science Foundation (NSF), through its National Center for Science and Engineering Statistics (NCSES), is the primary sponsor of the SDR. The National Institutes of Health also provides funding for the survey. The reference date for the 2010 SDR was 1 October 2010. The 2010 SDR was conducted by NORC at the University of Chicago.

The SDR is designed to complement two other surveys of scientists and engineers conducted by NCSES: the National Survey of College Graduates (NSCG, http://www.nsf.gov/statistics/srvygrads/) and the National Survey of Recent College Graduates (NSRCG, http://www.nsf.gov/statistics/srvyrecentgrads/). These three surveys share a reference date and have similar questionnaires. Results from the three surveys are combined into the Scientists and Engineers Statistical Data System (SESTAT) (see "Data Availability").

Some of the data on education and demographic information in the SDR come from the Survey of Earned Doctorates (SED), an annual census of research doctorates earned in the United States that began in 1957 (http://www.nsf.gov/statistics/srvydoctorates/). The SED provided a sampling frame for establishing the SDR in 1973 and continues to provide a sampling frame to update the SDR panel with new doctorate recipients for each new SDR survey cycle.

These notes provide an overview of the SDR protocol. Details are provided in the 2010 SDR methodology report, available upon request from the project officer.

Target Population

The 2010 SDR target population consisted of individuals with the following characteristics:

- Earned a research doctoral degree from a U.S. college or university in an SEH field by 30 June 2009
- Was under 76 years of age on 1 October 2010
- Was not terminally ill or institutionalized during the week of 1 October 2010

As in previous cycles, the 2010 SDR sampling frame was constructed from two separate listings: the returning 2008 SDR cohort, and a new cohort frame. The two cohorts are defined by the academic year of their first U.S.-granted SEH doctoral degree (see technical table B-1 for SEH fields included in the 2010 SDR sampling frame). The returning cohort frame represents individuals who received their SEH doctorate before 1 July 2007; the new cohort frame represents individuals who received their SEH doctorate between 1 July 2007 and 30 June 2009. The returning cohort frame is a *secondary frame*—it consists of the SDR sample selected for the previous survey cycle, and each frame member carries a

sampling weight from the previous cycle. The new cohort frame is a *primary frame,* including all known eligible cases from the two most recent doctoral award years.

The cases within the returning and new cohort frames were analyzed individually for SDR eligibility requirements. Individuals who did not meet the age criterion or who were known to be deceased, terminally ill, incapacitated, or permanently institutionalized in a correctional or health care facility were dropped from the sampling frames. After ineligible cases were removed from consideration, the remaining cases from the two frame sources were combined to create the 2010 SDR sampling frame. In total, there were 112,637 eligible cases in the 2010 SDR frame: 42,064 returning cohort cases, and 70,573 new cohort cases.

Sample Design

The 2010 SDR sample 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. Unlike the 2008 SDR sample design, the 2010 design incorporated a stratification factor based on the cases' last known location. The sample redesign with sample integration and revised stratification reduced undercoverage and improved operational procedures because it grouped together cases that were expected to require a similar level of effort to locate and to have similar employment and earning outcomes, depending on their expected residency.

The total number of cases selected for the 2010 SDR sample was 45,697. The sample design included 194 strata: 150 strata associated with the NSDR sample component, and 44 strata associated with the ISDR sample component. Regardless of citizenship status, all 2008 ISDR returning cases and any 2008 NSDR returning cases whose last known residence was outside the United States were classified into the 44 ISDR strata together with new cohort cases reporting plans in the SED to emigrate from the United States upon graduation. NSDR returning cases predicted to be U.S. residents and new cohort cases not reporting plans to emigrate after graduation were assigned to the 150 NSDR strata, regardless of their citizenship status. The frame was stratified by three variables—demographic group, degree field, and sex. The demographic group variable included nine categories defined by race and ethnicity, disability status, and citizenship at birth. To ensure higher selection probability for rarer population groups, classification of frame cases into these categories was done hierarchically. The goal of the 2010 sample stratification design was to create strata that represented subpopulations of greatest interest for separate estimation and reporting. The sample was then systematically selected from each stratum.

The 2010 SDR sample selection was carried out independently for each stratum and cohort substratum (i.e., the NSDR or ISDR). For returning NSDR cohort strata, the survey continued the past practice of selecting the sample with probability proportional to size, where the measure of size was the base weight associated with the previous survey cycle. Because the NSDR sample size is restricted to an allocation of no more than 40,000 sampled cases, not all of the returning cohort cases from the previous round are sampled, and this results in a maintenance cut of eligible returning NSDR cases.

For each stratum, the sampling algorithm started by identifying self-representing cases (i.e., those with a base weight = 1) and the non-self-representing cases (i.e., those with a base weight > 1). Non-self-representing cases within each stratum were sorted by citizenship, disability status, degree field, and year of doctoral degree award. The available sample (i.e., the total allocation of 40,000 sampled cases minus the number of self-representing cases) was selected from each stratum systematically with probability proportional to size. For the returning ISDR cohort strata, all cases were selected with certainty because the 2010 ISDR sample was not subject to a fixed allocation quota like the 2010 NSDR.

The new cohort samples for the NSDR and the ISDR were selected using the same algorithm that was applied to the returning NSDR cohort frame. However, because the base weight for every case in the new cohort frames is equal to 1, cases within each stratum had an equal probability of selection.

Thus, the 2010 SDR sample of 45,697 cases consisted of 40,000 cases from the NSDR sample component (36,543 cases from the returning cohort frame containing 37,267 eligible cases, and 3,457 cases from the new cohort frame containing 63,360 eligible cases), and 5,697 cases from the ISDR sample component (4,797 from the returning cohort frame containing 4,797 eligible cases, and 900 cases from the new cohort frame containing 7,213 eligible cases). The overall sampling rate was about 1 in 20 (5.2%), although sampling rates varied considerably across strata. Of the 45,697 sampled cases, a total of 31,462 cases completed the survey and were residing in the United States on the survey reference date and contributing to the U.S. SEH doctoral population estimates. An additional 4,032 cases completed the survey but were residing outside of the United States on the survey reference date and were not contributing to the U.S. SEH doctoral population estimates. All critical items (respondent's residency, employment status, and current occupation or former occupation if no longer working) must be provided for a case to be considered complete. The completed eligible cases residing in the United States consisted of 28,467 cases from the returning cohort sample and 2,995 cases from the new cohort sample.

Survey Instrument

The questionnaire comprises a large set of core data items that are retained in each survey round to enable trend comparisons and several sets of module questions that are asked intermittently on special topics of interest. The module for the 2008 SDR that gathered information on sample members' second job, if applicable, was not included in 2010, nor were questions measuring respondents' research productivity (authorships or coauthorships of papers, articles, books or monographs; number and type of patents earned). For 2010, the survey added a module of questions about college and university courses in which survey respondents were enrolled during the survey reference period and retained a question measuring technical expertise required for respondents' and respondents' spouses' primary job. See questionnaires at http://www.nsf.gov/statistics/srvydoctoratework/#qs.

As noted, critical items are required for a case to be considered complete. After indicating their residency (in or out of the United States) and employment status (working or not working) on the reference date, all respondents must provide their job title and a brief description of their duties and responsibilities for their current or most recent job, and nonworking respondents must also indicate whether or not they were looking for employment during the 4 weeks prior to the reference date.

Data Collection

Data collection for the 2010 SDR employed three main protocols. Each protocol used a different initial mode for data capture based primarily on the returning cohort's prior indication of mode preference:

- Self-administered paper questionnaire (SAQ)
- Computer-assisted telephone interview (CATI)
- Self-administered online questionnaire (Web)

After initial contact, each protocol included sequential contacts by postal mail, telephone, and e-mail that ran in parallel throughout the data collection period. In addition, sample members were encouraged to participate in the mode that was most convenient for them.

SAQ protocol (35.8% of sample members; 16,373). Initial contact was an advance notification letter from NSF. The first questionnaire was mailed 1 week after initial contact, followed by a postcard mailed 1 week later that thanked persons for participating and reminded them to complete the survey.

Approximately 6 weeks after the first questionnaire mailing, sample members who had not returned a completed questionnaire (by any mode) were sent a second questionnaire by U.S. Postal Service Priority Mail. Three weeks later, any cases still not responding received a prompting notice via e-mail to verify receipt of the paper form and to encourage cooperation. Telephone follow-up calls began 2 weeks later for all outstanding SAQ start-mode nonrespondents to request participation, preferably by the CATI mode.

CATI protocol (4.6% of sample members; 2,088)**.** Initial contact was an advance notification letter from NSF. Telephone contact and interviewing began 1 week after initial contact. Approximately 6 weeks later, sample members who had not yet responded were sent an e-mail prompt to solicit survey participation in any mode. Three weeks later, any cases still not responding received a first questionnaire mailing sent via U.S. mail, followed by a postcard mailed 1 week later that thanked persons for participating and reminded them to complete the survey. Seven weeks after the first questionnaire mailing, a second questionnaire was mailed to the remaining nonrespondents.

Web protocol (58.6% of sample members; 26,786)**.** Initial contact was a survey notification letter via U.S. mail and e-mail; this letter included a PIN and a password to access the Web survey. Two and a half weeks later, sample members who had not yet responded were sent a follow-up reminder letter via U.S. mail and also a reminder e-mail. Two weeks later, any cases still not responding received a prompting telephone call to verify receipt of the access information for the Web survey and to encourage cooperation. Four weeks later, any cases still not responding received a first paper questionnaire via U.S. mail, followed by a thank-you/reminder postcard 1 week later. Seven weeks after the first questionnaire mailing, a second questionnaire was mailed to the remaining nonrespondents.

Three additional prompting contacts were sent later in the data collection field period to any remaining nonrespondents from any of the starting mode groups in March, June, and July 2011.

Quality assurance procedures were in place at each step (address updating, printing, package assembly and mailing, questionnaire receipt, data entry, coding, CATI, and post–data collection processing). Active data collection ended in July 2011. The telephone contact and data entry processes ended on 16 July 2011 and 25 August 2011, respectively. However, Web-survey access remained available until 16 September 2011 to capture any last-minute responses. Overall, 26.4% of the responses were SAQ, 11.0% were CATI, and 62.6% were Web surveys, with 24.7% of the respondents choosing to respond in a mode other than their initial start mode.

Response Rates

Response rates were calculated on complete responses, as determined by the presence of critical items. The overall unweighted response rate was 79.8%; the weighted response rate was 79.9%. The 2010 SDR unweighted and weighted response rates are comparable to the response rates obtained in past survey cycles. Lower response rates generally occurred among groups of non-U.S. citizens (unweighted response rate = 71.0%) and among persons with missing demographic data (unweighted response rate = 47.2%). Missing demographic data typically indicated incomplete records from the SED. These cases typically are more difficult to locate. Prior experience has shown that sample members who are located usually complete the survey. Individuals who could not be located accounted for 42.7% of nonresponse cases.

Data Editing and Coding

Complete case data were captured and edited under the three separate data collection modes for the 2010 SDR. A computer-assisted data entry (CADE) system was used to process the SAQ paper forms. The CATI system, including an additional CATI instrument used to collect critical-item follow-up data, and

the Web survey had internal editing controls. Mail questionnaire data and Web-based returns were reviewed for any missing critical items (working status, job title, duties and responsibilities, and residency in the United States or elsewhere). Telephone callbacks were used to obtain this information for a complete response. All completed CATI responses included critical items. Complete responses from the three separate modes were merged into a single database for all subsequent coding, editing, and cleaning necessary to create an analytical database.

Following established guidelines for SESTAT, staff were trained in conducting a standardized review and coding of occupation and education information, "other/specify" verbatim responses, state and country geographical information, and postsecondary institution information. For standardized coding of occupation, the respondent's reported job title, duties and responsibilities, and other work-related information from the questionnaire were reviewed by specially trained coders who corrected known respondent self-reporting problems to obtain the best occupation codes. The education code for the field of study of a newly earned degree or for the first bachelor's degree earned if not reported previously was assigned solely on the basis of the verbatim response for that degree field.

Imputation of Missing Data

Item nonresponse for key employment items, such as employment status, sector of employment, and primary work activity, ranged from 0.0% to 3.0%. Nonresponse to questions deemed sensitive was higher: nonresponse to salary was 8.9%, and nonresponse to earned income was 10.9%. Personal demographic data, such as sex, marital status, citizenship, ethnicity, and race, had item nonresponse rates ranging from 0.0% to 4.2%, with sex at 0.0%, birth year at 0.3%, marital status at 4.2%, citizenship at 0.2%, ethnicity at 0.5%, and race at 1.6%. Item nonresponse was imputed using logical imputation and hot-deck imputation methods.

Logical imputation often was accomplished as part of editing. In the editing phase, the answer to a question with missing data was sometimes determined by the answer to another question. In some circumstances, editing procedures found inconsistent data that were blanked out and therefore subject to statistical imputation. During sample frame building for the SDR, some missing demographic variables, such as race and ethnicity, were imputed before sample selection by using other existing information from the sampling frame. Imputed values for race and ethnicity that were used for sampling were not included in the survey's data collection; therefore, race and ethnicity were imputed in post–data processing if this information remained missing. However, sampled cases with imputed values for race and ethnicity who responded in either the Web or CATI mode were asked these questions.

The 2010 SDR primary method for statistical imputation was hot-deck imputation. Almost all SDR variables were subjected to hot-deck imputation, with each variable having its own class and sort variables structured by a multiple regression analysis. However, imputation was not performed on critical items or on text variables. For some variables, there was no set of class and sort variables that was reliably related to or suitable for predicting the missing value. In these instances, consistency was better achieved outside of hot-deck procedures using random imputation.

Weights

With the integration of the ISDR and NSDR samples in 2010, the weights for the combined samples were developed in a single process. The weight a respondent receives approximates the number of persons in the population of recipients of U.S. doctorates that the sampled person represents. Another purpose of weights is to account for unequal selection probabilities and unit nonresponse. The first step of the weighting process calculated a base weight for all cases selected into the 2010 SDR sample. The base weight accounts for sample design, and it is defined as the reciprocal of the probability of selection under

the sample design. Then, an adjustment to the base weight was made for unknown eligibility, based on the inverse of the estimated propensity of each sample case to be assigned a known status using a logistic regression analysis; this process inflated weights to compensate for unknown cases.

In the next step, an adjustment for nonresponse was performed on completed cases to account for the sample cases that did not complete the survey. Nonresponse-adjusted weights were assigned, also using a logistic regression analysis, to respondents and to known ineligible cases (i.e., cases who were deceased, institutionalized, over 75 years of age, or living abroad during the survey reference period), but eligible nonrespondents and cases with unknown eligibility received a weight of zero.

Finally, the sum of unknown and nonresponse-adjusted weights for the known eligible cases was poststratified to align to the sum of the population counts from the 2009 Doctorate Records File (DRF), the SED's cumulative file, which includes records of all research doctorate awards granted by U.S. institutions through 30 June 2009. The total weight carried by unknown-eligibility cases was distributed to respondents assuming the same eligibility rate as observed among the respondents. Thus, the sum of weights equals the DRF frame size.

Reliability of Estimates

Sampling Error

The particular sample that was used to estimate the 2010 population of SEH doctorate recipients in the United States is one of a large number of samples that could have been selected using the same sample design and sample size. Estimates based on each of these samples would likely be apt to vary, and such random variation across all possible samples is called the *sampling error.* Sampling error is measured by the variance or standard error of the survey estimate.

The successive difference replication method (SDRM) was used to estimate sampling errors. The theoretical basis for the SDRM is described in Wolter (1984) and in Fay and Train (1995). As with any replication method, successive difference replication involves constructing a number of subsamples (replicates) from the full sample and computing the statistic of interest for each replicate. The mean square error of the replicate estimates around their corresponding full sample estimate provides an estimate of the sampling variance of the statistic of interest.

Each data table contains information on standard errors that is based on the method described above. The standard error of an estimate can be used to construct a confidence interval for the estimate. To construct a 95% confidence interval for an estimate, the corresponding standard error of the estimate is first multiplied by a z-score of 1.96 (i.e., by the reliability coefficient), then added to the estimate to establish the upper bound of the confidence interval, and then subtracted from the estimate to establish the lower bound of the confidence interval.

Nonsampling Error

Quality assurance procedures are included throughout the various stages of data collection and data processing to reduce possibilities for *nonsampling error,* which include (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 precisely measured; (3) coverage error, which arises when some members of the target population are excluded from the frame and therefore 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 occur at the point of data editing, coding, or data entry. The analyst should be aware of potential nonsampling errors, but these errors are far harder to quantify than sampling errors.

Data Limitations

Caution should be exercised when making comparisons with SDR data from previous survey cycles.

Survey Frame Changes

2006. In all cycles of the SDR except 2006, the new cohort consisted of graduates from the 2 academic years immediately preceding the survey year. In 2006, SDR collected data from graduates in the 3 previous academic years.

2003. Beginning with 2003, the new cohort frame includes all SEH doctorate recipients except those who earned an SEH doctorate in a prior year. The SDR frame is based on the first U.S. research doctorate earned in an SEH field.

2002 and prior. Recipients of two doctorates whose first degree was in a non-SEH field were not included in the SDR frame, even if their second doctorate was in an SEH field. Based on information collected annually by the SED on the number and characteristics of those earning two doctorates, this exclusion resulted in a slight undercoverage bias. Between 1983 and 2000, for example, the total number of double doctorate recipients with a non-SEH first doctorate and an SEH second doctorate was 154, representing 0.046% of the total number of SEH doctorates awarded in that period.

Questionnaire Changes

2010. The 2010 questionnaire included several changes from the 2008 version. The module questions on respondents' second jobs, patents, and publications were dropped. At the same time, the SDR reinstated from previous rounds' questionnaires a module on enrollment and course taking at a college or university and also questions on components of job satisfaction, whether employer is a new business, importance of job benefits, membership in professional associations, attendance at professional conferences, and federal agencies supporting research work. Three new questions were added: year of tenure, year of retirement, and degree of difficulty concentrating, remembering, or making decisions.

2008. The 2008 questionnaire included a module that gathered information on sample members' second job, as well as two sets of questions reinstated from the 2003 questionnaire: (1) questions measuring technical expertise required for respondents' and respondents' spouses' primary job, and (2) questions measuring respondents' research productivity (authorships or coauthorships of papers, articles, books, or monographs; number and type of patents earned). The 2006 modules on postdoctoral appointments and international collaboration were not included.

2006. The 2006 questionnaire included a module on the history of postdoctoral appointments, awarded primarily for gaining additional education and training in research, as a follow-up to a similar module included in the 1995 SDR and also a module on international collaboration among doctorate recipients.

Data Presentation Changes

2010. Due to the inclusion and exclusion of certain module questions in the 2010 questionnaire compared to the 2008 questionnaire, there are some differences in 2010 data table availability compared to 2008.

2003. Data on employed doctorate recipients were further classified to include a new category for science and engineering (S&E)-related occupations. This category includes health-related occupations, S&E managers, S&E precollege teachers, and S&E technicians and technologists.

2002 and prior. Data on employed doctorate recipients were classified into two categories: employment in an S&E occupation, and employment in a non-S&E occupation.

Definitions and Explanations

Employer location. Survey question A9 includes the location of the principal employer, and data were based primarily on responses to this question. Individuals not reporting place of employment were classified by their last mailing address.

Field of doctorate. The doctoral field is as specified by the respondent in the SED at the time of degree conferral. These codes were subsequently recoded to the field-of-study codes used in the SDR questionnaire. (See technical table B-1 for a list and classification of fields of degree reported in the SDR and in the SED sampling frame.)

Full-time and part-time employment. Full-time (working 35 hours or more per week) and part-time (working less than 35 hours per week) employment status is for the principal job only and not for all jobs held in the labor force. For example, an individual could work part time in his or her principal job but full time in the labor force. Full-time and part-time employment status is not comparable to data reported in previous years, when no distinction was made between the principal job and the other jobs held by the individual.

Involuntarily out-of-field rate. Involuntarily out-of-field rate is the percentage of employed individuals who reported, for their principal job, working in an area not related to the first doctoral degree at least partially because a job in their doctoral field was not available.

Labor-force participation rate. The labor-force participation rate (R_{LF}) is the ratio (E + U) / P, where E (employed) + U (unemployed; not-employed and actively seeking work) = the total labor force, and P = population, defined as all SEH doctorate holders less than 76 years of age who resided in the United States during the week of 1 October 2010 and who earned their doctorate from a U.S. institution.

Occupation data. The occupational classification of the respondent was based on his or her principal job (including job title) held during the reference week—or on his or her last job held, if not employed in the reference week (survey questions A19/A20 or A5/A6). Also used in the occupational classification was a respondent-selected job code (survey question A21 or A7). (See technical table B-2 for a list and classification of occupations reported in the SDR.)

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. Race and ethnicity data are from the SED and prior rounds of the SDR. The most recently reported race and ethnicity data are given precedence.

Salary. Median annual salaries are reported for the principal job, rounded to the nearest $1,000, and computed for full-time employed scientists and engineers. For individuals employed by educational institutions, no accommodation was made to convert academic-year salaries to calendar-year salaries. Users are advised that, due to changes in the salary question after 1993, salary data for 1995–2010 are not strictly comparable with 1993 salary data.

Sector of employment. Employment sector is a derived variable based on responses to survey questions A13 and A15. In the data tables, the category 4-year educational institutions includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), and university-affiliated research institutes. Other educational institutions include 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. Private, for-profit includes respondents who were self-employed in an incorporated business. Self-employed includes respondents who were self-employed or were a business owner in a nonincorporated business.

Unemployment rate. The unemployment rate (R_u) is the ratio U / (E + U), where U = unemployed (not-employed and actively seeking work), and E (employed) + U = the total labor force.

Data Availability

Additional data and reports from the SDR are available at http://www.nsf.gov/statistics/doctoratework/. Data from the SDR are also available in SESTAT at http://www.nsf.gov/statistics/sestat/. SESTAT provides an integrated database of information on employment, education, and demographic characteristics of scientists and engineers in the United States collected through the SDR, the NSCG (http://www.nsf.gov/statistics/srvygrads/), and the NSRCG (http://www.nsf.gov/statistics/srvyrecentgrads/).

References

Fay RE, Train GF. 1995. Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *ASA Proceedings of the Section on Government Statistics:*154–9.

Wolter K. 1984. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* 79(388):781–90.