Survey of Doctorate Recipients, 2015

Daniel J. Foley
Survey Statistician
National Center for Science and Engineering Statistics
dfoley@nsf.gov

Technical Notes

Survey Overview

Purpose. The Survey of Doctorate Recipients (SDR), conducted by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF), provides data on the characteristics of science, engineering, and health (SEH) doctorate degree holders. A research doctorate is a doctoral degree that (1) requires the completion of an original intellectual contribution in the form of a dissertation or an equivalent culminating project (e.g., a published manuscript) and (2) is not primarily intended as a degree for the practice of a profession. The most common research doctorate degree is the PhD. The SDR samples individuals who have earned an SEH research doctorate from a U.S. academic institution and are less than 76 years of age. The SDR provides data useful in assessing the supply and characteristics of the nation's SEH doctorates employed in educational institutions, private industry, and professional organizations, as well as federal, state, and local governments.

The SDR is designed to provide demographic, education, and career history information about individuals who earned a research doctorate in an SEH field from a U.S. academic institution and to complement another survey of scientists and engineers conducted by NCSES: The National Survey of College Graduates (NSCG, https://www.nsf.gov/statistics/srvygrads/). These two surveys share a common reference date, and use similar questionnaires and data processing guidelines.

Some of the 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 (https://www.nsf.gov/statistics/srvydoctorates/). The SED provides the sampling frame for the SDR through its annual update of the longstanding Doctorate Records File (DRF), a cumulative listing of all U.S.-earned doctorate recipients dating back to 1920.

These technical notes provide an overview of the 2015 SDR. Complete details are provided in the 2015 SDR methodology report, available upon request from the SDR Project Officer.

Data collection authority. The information collected in the SDR is solicited under the authority of the National Science Foundation Act of 1950, as amended, the America COMPETES Reauthorization Act of 2010, and the Confidential Information Protection and Statistical Efficiency Act of 2002. The Office of Management and Budget control number is 3145-0020 and expires on 31 August 2018.

Survey contractor. NORC at the University of Chicago.

Survey sponsor. The SDR is sponsored by NCSES with support from the National Institutes of Health.

Major changes to the recent cycle. The SDR sample size more than doubled for the 2015 survey cycle to 120,000 individuals, from approximately 47,000 individuals in the 2013 cycle. This sample size increase was designed to significantly improve estimation capabilities at the fine field of degree level reported in the SED. The 2015 SDR expansion required the selection of a new sample drawn from the Doctorate Records File (DRF). The DRF includes information on all research doctorates earned from U.S. institutions and

serves as the sampling frame for the SDR. The annual SED, in place since 1957, is used to update the DRF with new cohorts of U.S.-trained research doctoral graduates in all fields.

The overarching 2015 SDR sample design objectives were twofold:

- Produce reliable estimates of employment outcomes by the fine field of degree taxonomy used in the SED
- Maintain the existing estimation of various demographic characteristics and those currently used in NCSES publications such as Science and Engineering Indicators, Women, Minorities and People with Disabilities in Science and Engineering, and in the published SDR data tables.

The new sample design improves the coverage of the SDR to allow full representation of internationally residing U.S.-trained SEH doctorate recipients.

Key Survey Information

Frequency. Biennial.

Initial survey year. 1973.

Reference period. The week of 1 February 2015.

Response unit. Individuals with an SEH research doctorate from a U.S. academic institution.

Sample or census. Sample.

Population size. Approximately 1,047,900 individuals; 920,050 residing in the United States and 127,800 residing outside the United States.

Sample size. 120,000 individuals.

Survey Design

Target population. The SDR target population includes individuals that meet the following criteria:

- Earned an SEH research doctorate from a U.S. academic institution prior to 1 July 2013
- Are not institutionalized or terminally ill on 1 February 2015
- Are less than 76 years of age as of 1 February 2015

Sample frame. The SDR uses the Doctorate Records File (DRF), constructed from the annual SED, as its sampling frame. Based on the information available in the DRF, individuals who did not meet the age criterion or who were known to be deceased, or earned a degree in a non-SEH field were dropped from the frame. For those individuals who completed more than one SEH research doctorate, only the first-degree information was used for sampling eligibility.

Sample design. The SDR uses a fixed panel design with a sample of new doctoral graduates added to the panel in each biennial survey cycle up until the 2013 SDR survey cycle. For the 2015 SDR, a new sample was selected from the entire DRF via a stratified design, where the strata are defined by 215 fields of study listed in the 2013 SED (see technical table A-1). The new SDR sample includes an oversample of the following groups:

- Individuals included in the 2013 SDR
- Underrepresented minorities in the doctorate population
- Women

This targeted oversampling was implemented to continue supporting researchers who use SDR data to conduct longitudinal studies, and to improve the precision of estimates for women and minorities within the new sampling strata.

The resulting 2015 SDR sample of 120,000 cases consisted of 16,075 cases from the 2013 SDR; 10,337 cases from the DRF's new cohort of 2014 and 2015 graduates; and 93,588 cases from the 2013 DRF. The overall sampling rate was about 1 in 9 (10.9%), although sampling rates varied across strata.

Data Collection and Processing Methods

Data collection. The data collection period lasted approximately nine months. The SDR used a trimodal data collection approach: self-administered questionnaire (via mail), self-administered online survey (Web), and computer-assisted telephone interview (CATI). Individuals in the sample were started in one mode depending on their past preference and their available contact information. After an initial survey invitation, the data collection protocol included sequential contacts by postal mail, telephone, and e-mail that ran throughout the data collection period. At any time during data collection, sample members could choose to complete the survey using any of the three modes. Nonrespondents to the initial survey invitation received follow-up with alternate survey modes.

Quality assurance procedures were in place at each data collection step (address updating, printing, package assembly and mailing, questionnaire receipt, data entry, coding, CATI, and post-data collection processing). Active data collection ended in May 2016. The telephone contact and data entry processes ended on 27 May 2016 and 30 June 2016, respectively. Web-survey access remained available until 24 June 2016 to capture last-minute responses to the extent possible.

Mode. About 81% of the participants completed the Web mode, 10% of the participants completed the mail mode, and 9% of the participants completed via CATI mode.

Response Rates. Response rates were calculated on complete responses, that is, from instruments with responses to all critical items. Critical items are those containing information needed to report labor force status, including employment status, job title, and job description, as well as location of residency on the reference date. The overall unweighted response rate was 68%; the weighted response rate was 66%. Largely because of the significant change in sample design and increase in number of sample members, the 2015 SDR unweighted and weighted response rates are lower than the response rates obtained in past survey cycles.

Of the 120,000 sampled cases, a total of 78,320 cases completed the survey. Among the respondents, 67,925 cases completed the survey and were residing in the United States on the survey reference date and contributed to the U.S. SEH doctoral population estimates. An additional 10,395 cases completed the survey, but were residing outside of the United States on the survey reference date. The 10,395 non-U.S. residing cases contributed to the estimates of internationally residing U.S.-trained SEH doctoral population.

Individuals who could not be located accounted for about two-thirds of the nonrespondents.

Data Editing. Complete case data were captured and edited under the three separate data collection modes for the 2015 SDR. The Web survey captured most of the survey responses and had internal editing controls where appropriate. The CATI system also had internal editing controls. A computer-assisted data entry

(CADE) system was used to process the mail paper forms. Mail questionnaire data 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 additional information for incomplete mail responses. 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 NCSES guidelines for coding SDR and NSCG survey data, including verbatim responses, 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 (including auto-coding), 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 errors 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 based on the verbatim response for that degree field.

Imputation. Item nonresponse for key employment items, such as employment status, sector of employment, and primary work activity, ranged from zero to 3.9%. Nonresponse to questions deemed sensitive was higher: nonresponse to salary was 11.9%, and nonresponse to earned income was 19.2%. Personal demographic data, such as sex, marital status, citizenship, ethnicity, and race, had variable item nonresponse rates with sex at 0.02%, birth year at 0.78%, marital status at 13.5%, citizenship at 7.4%, ethnicity at 1.4%, and race at 8.9%. Item nonresponse was addressed using logical imputation and hot-deck imputation methods.[1]

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. All sample members with imputed values for race or ethnicity were given the opportunity to report these data if they responded in the Web or CATI modes. Respondents with missing race or ethnicity data who did not take the opportunity to report these data—and whose race or ethnicity could not be imputed by surname or birthplace—were assigned race/ethnicity values through hot-deck procedures during post-data processing.

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 to identify nearest neighbors for imputed information. However, imputation was not performed on critical items or on verbatim-based 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, such as day of birth. In these instances, consistency was better achieved outside of hot-deck procedures using random imputation.

Weighting. Because the SDR is based on a complex sampling design and subject to nonresponse bias, sampling weights are created for each respondent to support unbiased population estimates. The final analysis weights account for:

- Differential sampling rates
- Adjustments for unknown eligibility
- Adjustments for nonresponse

- Adjustments to align the sample distribution with the DRF distribution with respect to gender, race and ethnicity, degree year, and degree field

The final sample weights enable data users to derive survey-based estimates of the SDR target population. The variable name on the SDR public use data files for the SDR final sample weight is WTSURVY.

Detailed information on weighting is contained in the 2015 SDR Methodology Report, available upon request from the SDR Project Officer.

Variance estimation. 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. The 2015 SDR variance estimates are based on 104 replicate weights. Please contact the SDR Project Officer to obtain the SDR replicate weights and the replicate weight user guide.

Disclosure protection. To protect against the disclosure of confidential information provided by SDR respondents, the estimates presented in SDR data tables are rounded to the nearest 50, though calculations of percentages are based on unrounded estimates. With the new sample design, NCSES is evaluating its rounding and computation guidelines for the SDR to determine if rounding continues to be necessary.

Data table cell values based on counts of respondents that fall below a predetermined threshold are deemed to be sensitive to potential disclosure, and the letter "D" indicates this type of suppression in a table cell.

Survey Quality Measures

Sampling error. The SDR is subject to sampling error. Estimates of sampling errors associated with this survey were calculated using replicate weights and are included in each table of estimates. Data table estimates with coefficients of variation (that is, the estimate divided by the standard error) that exceed a predetermined threshold are deemed unreliable and are suppressed. The letter "S" indicates this type of suppression in a table cell.

Coverage error. The concept of coverage in the survey sampling process is the extent to which the known population that is deemed eligible for sample selection (that is, the sampling frame) "covers" the survey's target population. Any missed doctoral graduates within the DRF derived from the SED, which is a census survey of all research doctorates awarded annually in the United States, would create undercoverage in the SDR. Additional undercoverage errors may exist because of self-reporting errors in the SED that led to incorrect classification of individuals as not having earned an SEH research doctorate.

Nonresponse error. The weighted response rate for the 2015 SDR was 66%; the unweighted response rate was 68%. Results from the research and analysis of SDR nonresponse trends have been used in the development of the nonresponse weighting adjustments to minimize the potential for nonresponse bias in the SDR estimates. The SDR item nonresponse rate for key employment items, such as employment status, sector of employment, and primary work activity, ranged from zero to 3.9%. Other variables, especially those involving sensitive information, had higher nonresponse rates. For example, salary and earned income had item nonresponse rates of approximately 12% to 19%. A hot-deck imputation method was used to compensate for the item nonresponse.

Measurement error. The SDR is subject to reporting errors from differences in interpretation of questions and by modality (Web, mail, CATI). To reduce measurement errors, the SDR questionnaire items were pretested in focus groups and cognitive interviews.

Data Comparability and Changes

Data comparability. Year-to-year comparisons can be made among the 1993 to 2015 survey cycles because many of the core questions remained the same. Small but notable differences exist across some survey cycles, however, such as the collection of occupation data being based on the different versions of the occupation taxonomy. Also, due to variation in the month of the reference date in some survey cycles, seasonal differences may occur when making comparisons across cycles and decades. Thus, use caution when interpreting cross-cycle and cross-decade comparisons. Also, the definition of the SDR survey target population has experienced the following changes over time:

- Surveys conducted before 1991 included individuals who received research doctorates in fields other than SEH and individuals who received their doctorates from non-U.S. institutions.

- From 1999 to 2008, data for industrial engineers were mislabeled as "Materials/metallurgical engineers." For all years, data in this category have included only industrial engineers. For all years, data on Materials/metallurgical engineers have been included only in Other engineers.

- Surveys conducted prior to 2010 did not cover SEH doctorates residing outside of the United States.

- Since 2010, coverage of SEH doctorates residing outside of the United States only included those having graduated since 2001.

- The 2015 SDR refreshed sample improved population coverage to include all SEH doctorates awarded by U.S. institutions regardless of the academic year of award or the graduate's post-graduation residency location.

Caution is recommended when considering any analysis of trends that span pre- and post-1991 surveys, pre- and post-2010 surveys, and pre-and post- 2015 surveys because of the changes in the survey design and target population.

Overlap in sample cases across survey cycles allows for longitudinal analysis using SDR data. To link cases on the SDR public use data files across survey cycles, use the unique identification variable REFID.

Changes in survey coverage and population.

- 2015. The 2015 SDR survey cycle includes a change in both its sampling stratification and sampling frame but maintains a consistent target population. The 2015 cycle reflects a fresh sample selected from the DRF and sampling strata defined by fine field of degree. In addition, the 2015 sample represents all U.S.-trained doctorate holders with a first SEH degree regardless of their citizenship or plans to leave the United States upon graduation, which were eligibility delimiters in past cycles of the SDR. To analyze U.S.-residing cases only, use the variable FNINUS which indicates living or working in the United States on the reference date.

- 2010 and 2013. Beginning with the 2010 SDR and retained in the 2013 cycle, the sampling and weighting procedures integrated the U.S.-residing national (NSDR) and the non-U.S.-residing international (ISDR) sample components. Complete surveys from respondents located in the United States on the survey reference date were included in the SESTAT analysis dataset regardless of the initial sample component.

- 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.

- 2001 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.

Changes in Questionnaire.

- 2015. The 2015 questionnaire differed from the 2013 questionnaire by adding "National Aeronautics and Space Administration (NASA)" as response category 6 to questionnaire item A43 (Federal agencies or departments supporting your work). "National Science Foundation (NSF)" became response category 7, "Other" became response category 8, and "Don't know source agency" became response category 9. In addition, a new questionnaire item was added (E12) that included three questions to help verify information about the individual's doctorate: (1) the institution granting the doctorate, (2) the field of study of the doctorate, and (3) the month and year it was granted.

- 2013. The 2013 questionnaire differed from the 2010 questionnaire by splitting the first response category for the indicator of sample member location on the survey reference date into two categories. "United States, Puerto Rico, or another U.S. territory" became "United States or Puerto Rico" and "Another U.S. territory."

- 2010. The 2010 questionnaire differed from the 2008 questionnaire as follows. 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 coursetaking at a college or university and also questionnaire items 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 questionnaire items 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 individual's second job, as well as two sets of questions reinstated from the 2003 questionnaire: (1) questions measuring technical expertise required for the respondent's and the respondent's spouse's primary job, and (2) questions measuring respondent's research productivity (authorships or co-authorships 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.

Changes in reporting procedures or classification.

- 2015. Data tables reporting at the fine field of degree level have been added consistent with the 2015 sample design. Data tables that report on the non-U.S. residing population have been added consistent with the updated sample design that provides full coverage of the non-U.S. residing population.

- 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 with 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

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 A-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 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 non-institutionalized SEH doctorate holders less than 76 years of age during the week of 1 February 2015 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 A5/A6 or A19/A20). Also used in the occupational classification was a respondent-selected job code (survey question A7 or A21). (See technical table A-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 1995–2015 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 reported verbatim in the survey questionnaire). Users should note that prior to 2008 these other educational institutions that were written as 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 non-incorporated business.

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

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.

Notes

[1] Item nonresponse rates reflect data missing after logical imputation or editing, but before hot-deck imputation, for all variables except sex, predicted respondent location, ethnicity, and race. Demographic and location variables completed by logical imputation during frame construction were also counted as nonresponse, as well as those filled in by hot-deck imputation.