Attachment B

2015 Survey of Doctorate Recipients: Sample Design and Implementation Report

2015 SURVEY OF DOCTORATE RECIPIENTS:

Sample Design and Implementation

PREPARED FOR:

Steve Proudfoot, SDR COTR

National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230
(703) 292-5111

PREPARED BY:

Michael Yang

Karen Grigorian

NORC at the
University of Chicago
55 East Monroe Street
Chicago, IL 60603
(312) 759-4000

DECEMBER 4, 2015

NORC at the UNIVERSITY of CHICAGO

Table of Contents

Appendices

1 Overview of the 2015 SDR Sample Design

Since its inception in 1950, the National Science Foundation (NSF) has been charged to provide a central clearinghouse for the collection, interpretation and analysis of data on scientific and technical resources in the United States, and provide a source of information for policy formulation by other federal agencies. The Survey of Doctorate Recipients (SDR) has been an important means for the NSF to accomplish this objective. Conducted biennially since 1973, the SDR follows a sample of U.S.-trained doctorates in science, engineering, and health (SEH) fields throughout their careers, from shortly after degree award through age 75. The SDR is widely used by the U.S. Congress and Federal agencies, universities and professional societies, and other organizations and individuals interested in the nation's education, supply, and employment of doctorate recipients in SEH fields. Employers in universities, industry, and government sectors also use the SDR to understand and predict trends in employment opportunities and salaries for SEH doctorates.

The traditional target population of the SDR includes individuals who meet the following requirements:

- Received a doctoral degree in an SEH field from a U.S. institution;
- Age 75 years or younger on survey reference date; and
- Living in a noninstitutionalized setting on the survey reference date.

The SDR has historically featured a stratified systematic sample design, where the strata are defined by degree field, gender, race and ethnicity, citizenship, disability status, and other relevant demographic variables. The SDR sample design has undergone some significant modifications over the years in response to changes in its analytical objectives and budgetary constraints. For example, the number of strata has been reduced from over 1,000 in the early cycles to 150 as a result of the 2003 redesign. The target population of the SDR has also been redefined several times over the life course of the survey. For example, doctorates awarded in humanities were once part of the target population. Furthermore, prior to the 2003 survey cycle, the SDR restricted data collection to U.S. residents only. SEH doctorates who resided outside the U.S. on the survey reference date were excluded from the target population of the survey.

In addition to the sample redesign, the 2003 SDR included a methodological experiment which showed that data collection from international residents is operationally feasible.[1] From the 2006 cycle, the SDR sample consisted of two relatively independent components: the national SDR (NSDR) and the international SDR (ISDR).[2] While the NSDR covers doctorates residing in the U.S., the ISDR targets those residing outside of the U.S. For the 2010 SDR, the NSF decided to integrate the NSDR and ISDR to create a unified survey of U.S. trained SEH doctorates regardless of residential location.[3] The integrated sample design developed for the 2010 SDR was maintained for the 2013 SDR.[4]

The 2013 SDR features a total of 194 strata, including 150 NSDR strata and 44 ISDR strata. The NSDR strata are defined by degree field, gender, race and ethnicity, citizenship at birth, and disability status; the ISDR strata are defined by degree field, gender, race and ethnicity, and citizenship at birth. These strata were defined to align with the analytical domains used in official publications as well as those used by SDR data users.

The 2015 SDR features a substantial sample size expansion and sample redesign in response to a set of updated analytical objectives and requirements. The sample size is increased from 40,078 cases in 2013 to 120,000 cases in 2015. The main objective of this expansion is to support reliable estimates of employment outcomes by the fine field of degree (FFOD) taxonomy used in the Survey of Earned Doctorates (SED). With the marked increase in the overall sample size, the traditional SDR estimation capabilities are also expected to increase. As directed by the NSF, the overarching 2015 SDR sample design objectives are twofold:

- First, the expanded SDR is required to produce reliable estimates of employment outcomes by the fine field of degree taxonomy used in the SED;
- Second, the expanded sample is expected to maintain the existing estimation capabilities associated with analytical domains defined by various demographic characteristics and currently used in National Center for Science and Engineering Statistics (NCSES) publications such as Science and

[1] Grigorian, Karen and Tom Hoffer (2005). Non-U.S. Citizen Undercoverage Feasibility Study Report. Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.
[2] Cox, Brenda G., Karen Grigorian and Michael Yang (2006). The 2006 International Survey of Doctorate Recipients (ISDR): Sample Design. Report submitted to the National Science Foundation by Battelle under subcontract to the National Opinion Research Center at the University of Chicago, IL.
[3] Cox, Brenda. G., Karen Grigorian, Fang Wang, and Rebecca Wang (2012b). 2010 Survey of Doctorate Recipients: Sample Design and Implementation. Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.
[4] Cox, B. G., K. Grigorian, Y.M. Yang, M. Sinclair, 2013. 2013 Survey of Doctorate Recipients: Sample Design and Implementation. Prepared for the National Science Foundation, January 2013. Chicago, IL: NORC.

Engineering Indicators , Women, Minorities and People with Disabilities report, and detailed data tables.

For more detailed discussions of the 2015 SDR sample design objectives, please refer to the attached document "Requirements of Sample Expansion and Sample Redesign of the Survey of Doctorate Recipients" and its addendum, both can be found in Appendix A.

The expansion of the SDR, along with its new estimation objectives, required a significant redesign of the SDR sample. As specified by the NSF, the broad objectives of the SDR redesign include meeting the newly defined estimation objectives, resolving any longstanding sampling issues to improve efficiency, creating a more unified sample design which eliminates the NSDR and ISDR distinction, and constructing a flexible and sustainable design for the growing demands of SDR data.

The NSF and NORC conducted extensive research in order to meet these new requirements. Many design options were considered, simulated, and reviewed during the course of the research. Appendix B contains comprehensive discussions of these options. Upon evaluating all these options, the NSF decided to select a fresh new sample from a new sampling frame constructed from the original Doctorate Record File (DRF). By so doing, the existing SDR panel will cease to exist after the 2013 cycle and a new panel will start from the 2015 cycle. The main advantage of a fresh new sample is its unparalleled simplicity as it eliminates all the sample frame, sample design, and database maintenance complexity accumulated over the past 40 years under the old design. The drawback of a fresh new sample design, however, is the interruption of a prominent longitudinal data series.

This report documents the 2015 SDR sample design and selection procedures. Section 2 summarizes the major sample design changes from the previous cycle. The remaining sections discuss in detail the main parameters of the 2015 SDR design. Section 3 describes the frame construction process which is considerably different from the prior cycles because a completely new frame is required under the fresh new sample approach. Section 4 presents the 2015 SDR sample design, including sample size, sample stratification, and sample allocation procedures under the new sample design to meet the new analytical objectives. Section 5 describes the sample selection procedures, including methodical oversampling of the traditional SDR analysis domains under the new design. Section 6 briefly discusses how the SDR design will be maintained in 2017 and beyond. Finally, Section 7 provides some concluding remarks regarding data processing procedures under the 2015 design.

2 Sample Design Changes from the 2013 Cycle

The changes between the 2013 and 2015 SDR sample designs are substantial. Design changes in a longitudinal study such as the SDR must be documented so that data users can properly analyze the data and interpret their findings, especially when they employ SDR data from multiple cycles with different sample designs. Before presenting the 2015 SDR sample design and implementation in detail, this section highlights the most significant sample design changes from the 2013 SDR and their implications for analysts, as follows.

- Through the 2013 cycle, the SDR had been a longitudinal survey with a significant panel component. Well over 90 percent of the previous cycle sample is retained in the current cycle sample. With a fresh new sample selected from a newly constructed sample frame, the 2015 SDR no longer retains the previous panel, i.e., no explicit longitudinal panel is automatically carried forward from the 2013 sample. Through oversampling, about one third of the 2013 SDR panel cases is included in the 2015 SDR sample. In general, however, the 2015 sample does not support longitudinal analyses. It practically represents the starting point of a new panel sample.
- The 2015 SDR sample frame is constructed afresh from the most recent version of the DRF. In the past, the SDR sample frame at each cycle consists of two components: the panel (old cohort frame) from the previous cycle and the new cohort doctorates awarded after the previous cycle (new cohort frame). Note that the old cohort frame is a secondary frame because it is a sample itself. Conceptually, the 2015 SDR frame contains three components: (1) the 2013 SDR sample that remains eligible for the 2015 SDR (*n*=45,936); (2) the new cohort cases from the 2012 and 2013 SEDs, and (3) the 2015 “expansion cohort” (*n* = 979,526) constructed from the 2013 DRF. The expansion cohort is made up of the following:
  - Those that were selected into the SDR sample but later dropped from the panel due to ineligibility discovered during subsequent SDR data collection, including the deceased, no degree earned, and maintenance cut, i.e., deselection from the sample during the 1995-2013 cycles;
  - Those that were eligible for sample selection but were never selected during the past cycles;
  - Those that had been ineligible for selection based on previous target population definitions.
- The 2015 sample size is increased to 120,000 cases from a sample size of 47,078 for the 2013 SDR cycle.

- Instead of defining the sampling strata by degree field and demographics, as had been the case in the past, the 2015 SDR strata are defined by fine field of degree alone, reflecting the emphasis on the new analytical objectives at the fine field level.
- The 2013 SDR sample allocation is mostly proportional, with additional allocation to small domains to guarantee a minimum sample size for these domains. The 2015 SDR sample allocation involves a two-step process to achieve a compromise between the two sets of analytical goals: the first step allocation to the fine fields is intended to meet the analytical goals at the fine field level; the second step allocation by the broad field of degree is designed to maintain and improve the existing analytical capabilities by the traditional analysis domains. The result is a much more disproportional allocation across the explicit sampling strata. Based on the variation of the base weight, the overall 2015 SDR design effect is 1.59, while the 2013 overall design effect is 1.09.
- Under the 2015 design, the traditional analytical capabilities are maintained through oversampling women and underrepresented minorities (URM). The 2013 panel cases were also oversampled to support limited longitudinal analysis. For the key traditional SDR domains, a series of tables in Appendix C compares the coefficient of variation for a typical sample estimate between the 2013 and 2015 SDRs. With rare exceptions, the 2015 SDR is projected to achieve better precision than the 2013 SDR.

3 Sample Frame Construction

The 2015 SDR employed a completely new original sample frame constructed from the DRF. This section discusses the frame construction procedures in detail. The goals of frame construction are twofold: one is to include all doctorates in the target population so they all have a non-zero probability of being selected into the sample; the other is to define auxiliary frame variables to support sample design and survey operations. Subsection 3.1 discusses the identification of frame cases; subsection 3.2 discusses the construction of key frame variables. The layout of the frame file is presented in Appendix D.

3.1 Identifying Frame Cases

Prior to the current expansion, the SDR sample of each cycle consists of two components: an old cohort sample and a new cohort sample. While the new cohort sample is selected from the new cohort portion of the frame, the old cohort sample is selected from the old cohort frame that is composed of the previous cycle's sample. That is, the old cohort frame is a so-called secondary frame rather than an original frame constructed from the DRF. The old cohort frame represents the old cohort population through the base weight, and the old cohort sample represents the longitudinal panel that gets updated at each cycle through maintenance cut. Given the sample expansion, however, the 2015 SDR needs to redefine its sampling frame from the original DRF because a fresh new sample requires a fresh new sampling frame. The DRF is a database that contains educational information for all doctorate recipients from U.S. universities since 1920. The DRF is updated annually based on the SED which collects information annually from all doctorates awarded by U.S. institutions about their educational history, funding sources, and post-doctoral plans.

The target population for the 2015 SDR remains the same as the previous cycle except for the addition of the new cohort doctorates awarded in academic years 2012 and 2013. Specifically, it includes individuals who meet the following requirements regardless of residency location:

- Received a doctoral degree in an SEH field from a U.S. institution;
- Seventy five years of age or younger on 1 February 2015; and
- Living in a noninstitutionalized setting on 1 February 2015.

The final 2015 SDR sampling frame includes 1,102,985 cases, consisting of six groups of doctorates, as shown in Table 3.1, based on their historical relationship with the existing SDR design. These six groups may be combined into three broad categories: the panel, the new cohort, and the expansion cohort, as described below.

TABLE 3.1 The Six Groups of 2015 SDR Frame

Cohort	Frame Group	Description	SED Academic Years (AY)	Number of Cases
Panel	1	2013 SDR sample cases that remain eligible for 2015 SDR	1960-2011	45,936
Expansion Cohort	2	Permanently ineligible cases determined in past cycles of the SDR accrued since 1973 forward (i.e., deceased, no degree earned, non-U.S. citizens located abroad 2 cycles in a row)	1964-2011	2,292
	3	Maintenance cut cases removed from the sample during 1995-2013 sample selection (proportionally deselected regardless of response outcome)	1960-2009	64,532
	4	Eligible for primary selection from SED 1960-2011, but not selected	1959-2011	859,891
	5	Not eligible for primary selection from SED 1975-2000 that are now considered eligible (i.e., new graduates with plans to leave the U.S. after degree award)	1975-2000	52,811
New Cohort	6	New cohort cases from SED 2012 and 2013	2012-2013	77,523
Total				1,102,985

The panel portion of the frame is identified from the 2013 SDR sample of 47,078 doctorates. Of these cases, 45,936 meet the target population definition and are included in the 2015 sample frame. There are 893 cases determined to be out of scope for the 2015 frame based on information available in the DRF; 887 cases determined to be out of scope due to age ineligibility and 6 cases classified as double doctorates. There are an additional 249 cases known to be out of scope for the SDR based on information available from the 2013 SDR. These 249 cases are transferred to the eligible expansion frame case set and give a chance at selection. Unlike past cycles, the panel cases on the 2015 frame no longer carry a base weight; they represent no other cases other than themselves on the frame.

The expansion cohort is constructed from the 2013 DRF. These doctorates are needed on the 2015 frame because they are no longer represented by the panel cases through the base weight, as it was the case in prior SDR cycles. The expansion cohort frame consists of four groups of cases:

- **Permanently ineligible cases (Group 2):** These are cases that had been selected into the SDR sample in a previous cycle but were later dropped from the sample due to ineligibility discovered during subsequent SDR cycles. These include the deceased, those with no eligible degree, and non-U.S. citizens located abroad for two consecutive SDR cycles. NSF decided to include these known ineligible cases on the frame to simplify database maintenance as these cases, if dropped from the frame, will need to be brought back during post-survey data processing.

- **Maintenance cut cases (Group 3):** These are the cases that have been dropped from the SDR old cohort sample during 1995-2013 through random subsampling to maintain a stable sample size. Without such maintenance cut, the SDR sample size would have increased over time due to the addition of a new cohort sample at each cycle.
- **Non-selected cases (Group 4):** These are doctorates that had been eligible for sample selection but were never selected into the SDR sample in the previous cycles.
- **Previously ineligible cases (Group 5):** These are doctorates that were not eligible for the SDR based on previous target population definitions, i.e., new graduates with plans to leave the U.S. after degree award. These cases are eligible for the 2015 SDR.

The 2015 new cohort frame includes 38,140 cases from the 2012 SED and 39,383 from the 2013 SED. To ensure that all frame cases in these groups are defined consistently, only data available in the 2013 DRF are used as inputs, with the only exception being that data collected in previous SDR cycles are used to determine age eligibility.[5] The protocols for building the 2015 new cohort frame variables are applied to all eligible cases in the 2015 SDR sampling frame and are described in Section 3.2.

For each frame component, Table 3.2 shows the frequency of eligible and ineligible cases for all records in the 2013 DRF. This table accounts for eligible cases as well as cases determined to be ineligible for inclusion on the 2015 SDR frame. The final 2015 SDR frame contains 1,102,985 cases.

[5] Six cases—5 in the expansion cohort and 1 in the panel—are coded as “age eligible” and included in the 2015 frame based on SDR data although the DRF indicates that they are ineligible.

TABLE 3.2 2015 SDR frame eligibility status for all cases in the 2013 DRF

2015 SDR Frame Status		Sample Frame Component			Overall
		2013 Panel*	Expansion Cohort	New Cohort	
Eligible					
00	Frame Eligible	45,936	979,526	77,523	1,102,985
Ineligible					
01	Age ineligible	887	222,251	8	223,416
03	Deceased, according to the DRF	0	780	14	794
11	Non-SEH doctoral degree field	0	642,287	26,143	668,430
13b	Double Doc; first SEH doctorate earned before SED 2012/2013	6	261	49	316
Overall		**46,829**	**1,845,105**	**103,737**	**1,995,971**

* The 2013 SDR sample included 249 cases determined to be ineligible for the 2015 SDR based on information obtained during the 2013 survey. Most are known deceased cases. These cases are included in the 2015 SDR eligible frame shown here in the Expansion cohort case count and are given a chance of selection. If selected, these will be immediately finalized with an ineligible outcome for the 2015 cycle.

3.2 Construction of frame variables

Frame variables are used to support the sample design, including stratification variables, sorting variables, and sample selection variables. All frame variables are constructed from DRF data with age as the only exception. The primary variables used to stratify, sort, or assess eligibility for the 2015 SDR frame are as follows:

- PHDFIELD – doctorate degree field reported in the SED
- SDRFLD15 – based on PHDFIELD, aggregated recoding of the doctorate degree field
- PHDFY – year of doctorate degree award reported in the DRF
- CENTURY – based on PHDFY, the century of doctorate degree award
- RACETH15 and RACE15– these are racial group assignment derived from ethnicity and race data reported in the DRF; the component variables from the DRF are renamed ASIAN15, BLACK15, HISPANIC15, NATIVE15, PACIFIC15, and WHITE15 in the frame
- URM15 – based on RACETH15, underrepresented minority status
- BIRCIT15 – citizenship at birth based on data reported in the DRF

- SEX15 – gender reported in the DRF
- LOCSTAT15– predicted residency location based on information provided in the SED at the time of degree award
- AGE15 – age of each frame member relative to the 2015 SDR reference date based on age data reported in the DRF

When frame variables have missing data in the DRF, they are systematically imputed using a set of imputation rules. Therefore, constructing the frame variables amounts to imputing missing data on these variables. Missing data in the following frame variables are imputed: RACETH15, URM15, SEX15, LOCSTAT15, BIRCIT15, and AGE15. PHDFY and PHDFIELD are key design variables that do not contain any missing data on the DRF.

The details associated with each of these critical frame variables is described below including the imputation rules and the amount of missing data for each of the frame variables where applicable.

PHDFIELD and SDRFLD15. PHDFIELD is used to define the 2015 SDR sampling strata, and SDRFLD15 is used to support sample allocation as well as sample selection. PHDFIELD is never missing in the DRF, so no imputation is required for this variable. Since SDRFLD15 is derived from PHDFIELD, no imputation is required for SDRFLD15 either. The NSF required that all PHDFIELDs that represented fields of degree obtained in the 21st century (i.e. from academic year 2001 and later) be included in the frame and be used to form sampling strata. There are 36 eligible SEH fine fields of degree that are recorded in the DRF but were discontinued before academic year 2001. Under the 2015 design, each fine field of degree is its own sampling stratum, but these “20th century-only” fine fields are grouped together to form a single composite stratum. These discontinued fields contain a total of 26,825 cases. Table 3.3 details these discontinued fields of degree, displaying their codes, descriptions and period in which sample-eligible frame members earned degrees in those fields.

TABLE 3.3 20th Century discontinued fields of study[a] in 2015 SDR Frame

PHDFIELD Code	Fine field of degree	Years in DRF	Number of Cases
007	Animal Husbandry	1962-1982	565
032	Plant Protection/Pest Management	1988-1991	13
040	Food Sciences	1969-1989	1,720
042	Food Distribution	1994-1994	1
045	Soil Sciences	1968-1988	1,284
054	Fish and Wildlife Science	1964-1982	679
060	Wildlife	1983-1988	142
065	Forestry Science	1964-1988	1,309
140	Hydrobiology	1964-1979	132
156	Microbiology/Bacteriology	1961-1982	4,651
171	Genetics	1961-1982	1,918
186	Animal/Plant Physiology	1960-1960	1
205	Dentistry	1968-1968	1
219	Public Health/Epidemiology	1966-1982	973
224	Hospital Administration	1967-1977	37
225	Medical/Surgery	1964-1976	25
235	Optometry/Ophthalmology	1966-1966	1
322	Electrical Engineering	1961-1985	7,157
323	Electronics Engineering	1961-1983	1,081
354	Naval Architecture/Marine Engineering	1983-1991	64
506	Astronomy/Astrophysics	1962-1969	147
521	Agricultural/Food	1965-1979	221
545	Geophysics, Solid Earth	1962-1976	428
547	Fuel Technology/Petroleum Engineering	1967-1979	70
549	Mineralogy/Petrology/Geological Chemistry	1963-1969	95
554	Applied geology	1969-1991	279
555	Applied Geology/Geological Engineering	1965-1968	18
562	Electron Physics	1984-1991	23
563	Electromagnetism	1961-1979	135
567	Mechanics	1961-1976	50
573	Thermal Physics	1961-1981	161
575	Theoretical Physics	1961-1962	2
619	Human Engineering	1966-1966	1
679	Political Science/Public Administration	1960-1976	3,441
		Total of 20th century discontinued fields	26,825
		Total of 21st century fields	1,076,160
		Overall	1,102,985

[a] Two additional PHDFIELDS, Textile Engineering (375) and Experimental/Comparative & Physiological Psychology (616) were also identified as discontinued fields. However, no frame case with these PHDFIELDS was age eligible for selection into the 2015 SDR.

Table 3.4 presents the full distribution of SDRFLD15. Please see the field of study coding taxonomies crosswalk in Appendix E for the collapse of PHDFIELD into SDRFLD15.

TABLE 3.4 Frame Distribution of SDRFLD15

SDR Field	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Chemistry	91,847	3,726	83,212	4,909
Physics/Astronomy	65,215	2,587	58,284	4,344
Earth/Ocean/Atmospheric Sciences	29,405	1,312	26,376	1,717
Mathematics	55,147	2,234	48,971	3,942
Computer/Information Sciences	31,530	1,434	26,358	3,738
Agricultural Sciences	43,898	1,782	39,781	2,335
Medical Sciences	49,856	2,357	42,648	4,851
NIH Biological Sciences	112,094	4,651	99,048	8,395
Other Biological Sciences	111,123	4,681	97,940	8,502
Psychology	148,409	6,030	134,603	7,776
Economics	45,983	1,916	41,396	2,671
Anthropology/Archaeology/Sociology	42,754	2,043	38,034	2,677
Other Social Sciences	64,454	2,710	57,251	4,493
Electrical/Electronics/Communications Engineering	59,871	2,545	52,410	4,916
Other Engineering	151,399	5,928	133,214	12,257
Overall	1,102,985	45,936	979,526	77,523

PHDFY and CENTURY. PHDFY represents the academic year (called 'fiscal year' in the SED) of doctoral receipt. This variable is used to define the new cohort. It is also used to construct the CENTURY indicator as one of the sorting variables to support systematic sample selection. For cases earning a degree in the 20th century (PHDFY<2001), CENTURY is set to "20"; and those earning their degree in the 21st century (PHDFY≥2001) are set to "21." Since 1958, when the SED began to field its annual survey, PHDFY is never missing. Therefore, PHDFY and CENTURY contain no imputed data. Tables 3.5 and 3.6 illustrate the distribution of these variables in the final frame.

TABLE 3.5 PHDFY Distribution by Cohort

PHDFY	2013 Panel	Expansion Cohort	New Cohort	Total Cases
1959-1969	812	24,702	0	25,514
1970-1979	5,322	153,116	0	158,438
1980-1989	7,174	188,901	0	196,075
1990-1999	10,704	259,931	0	270,635
2000-2009	17,690	285,296	0	302,986
2010-2011	4,234	67,580	0	71,814
2012-2013	0	0	77,523	77,523
Overall	45,936	979,526	77,523	1,102,985

TABLE 3.6 CENTURY Distribution by Cohort

PHDFY	2013 Panel	Expansion Cohort	New Cohort	Total Cases
20th century	25,218	653,166	0	678,384
21st century	20,178	326,360	77,523	424,601
Overall	45,936	979,526	77,523	1,102,985

RACETH15, RACE15 and URM15. RACETH15 represents race and ethnicity, and URM15 represents underrepresented minorities. They are constructed from the separate race/ethnicity variables ASIAN15, BLACK15, HISPANIC15, NATIVE15, PACIFIC15, and WHITE15 after they are imputed. RACE15 represents racial group independent of ethnicity and collapses individuals selecting more than one race as multiracial.

There is a considerable amount of imputation in the 2015 SDR racial variables. Data on race and ethnicity are entirely missing before 1973 since the SED only started collecting this data with the 1973 cycle. In addition, the racial category of "Native Hawaiian/Pacific Islander" did not exist in the SED until the 2001 survey. The NSF-approved rules for assigning race and ethnicity are as follows:

1. Use reported data from the SED;
2. When ethnicity is missing, use the U.S. Census Bureau Hispanic surname list and impute any matches as Hispanic ethnicity (if race is also missing and the surname is Hispanic, impute the race to white);[6]
3. When race is missing, and ethnicity is either missing or non-Hispanic, use the GENESYS Asian surname list[7], and logically impute any matches as NH Asian;
4. When ethnicity is still missing, but race is reported, use place of birth to logically impute ethnicity;
5. When race and ethnicity are both still missing, use place of birth to logically impute race and ethnicity;

[6] The 2015 new cohort cases and 2015 panel cases that joined the panel in the 2013 survey round were updated using the Hispanic surname list based on the 2000 U.S. Census available as of 2011 located at http://www.census.gov/genealogy/www/data/2000surnames/index.html. The 2015 panel cases that joined the panel prior to the 2013 survey round were updated using the Hispanic surname list based on the 1990 U.S. Census.

[7] Market Systems Group provides the GENESYS Sampling Systems suite of sampling tools, which includes this algorithm that matches surnames to an Asian surname list for a nominal fee (http://www.m-s-g.com/Web/genesys/index.aspx).

6. When race and ethnicity are both still missing and place of birth is missing, impute to NH white.

The crosswalk of birth places to race and ethnicity imputation assignments is located in Appendix F, Table F2. The sources for race and ethnicity data in the 2015 SDR frame are detailed in Tables 3.7 and 3.9.

TABLE 3.7 Race Data Sources: 2015 SDR Frame

Race Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported	947,363	39,825	836,286	71,252
Surname imputation (Asian)	15,588	549	13,601	1,438
Birthplace imputation	95,403	3,854	90,790	759
Hot deck imputation	0	0	0	0
Default imputation (white)	44,631	1,708	38,849	4,074
Overall	1,102,985	45,936	979,526	77,523

After all missing data on race are imputed, the variable RACE15 is created to tabulate race classifications independent of ethnicity. In cases where one race is identified, the value of RACE15 is assigned to that race. Otherwise, in cases where self-report indicates more than one race, RACE15 is assigned to ‘6’ for “more than one race”. The frequencies of RACE15 are shown in Table 3.8 below.

TABLE 3.8 Frame Distribution of RACE15

Ethnicity Data Source	2013 Panel	Expansion Cohort	New Cohort	Total Cases
Asian	11,095	228,920	24,727	264,742
Black	2,605	29,599	3,398	35,602
Native	199	2,646	287	3,132
Pacific	84	696	112	892
White	31,414	712,301	47,263	790,978
More than one race	539	5,364	1,736	7,639
Overall	45,936	979,526	77,523	,102,985

TABLE 3.9 Ethnicity Data Sources: 2015 SDR Frame

Ethnicity Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported	962,550	40,837	850,247	71,466
Surname imputation (Hispanic)	3,552	199	3,068	285
Birthplace imputation	87,032	3,221	83,279	532
Hot deck imputation	0	0	0	0
Default imputation (non-Hispanic)	49,851	1,679	42,932	5,240
Overall	1,102,985	45,936	979,526	77,523

RACETH15 is defined in the following hierarchical manner:

- If a case is Hispanic or Latino, assign the case to the Hispanic value regardless of race;
- If a case is not Hispanic (NH) and is black, assign the case to the NH black regardless of other race selections;
- If a case is not Hispanic or black, and is Asian, assign the case to the NH Asian regardless of other race selections;
- If a case is not Hispanic, black, or Asian, and is American Indian or Alaskan Native, assign the case to the NH American Indian regardless of other race selections;
- If a case is not Hispanic, black, Asian, or American Indian, and is Native Hawaiian or other Pacific Islander, assign the case to the NH Pacific Islander regardless of other race selections; and
- Otherwise, assign the case to NH white.

The distribution of the resulting race/ethnicity group assignments is shown in Table 3.10.

TABLE 3.10 Race/Ethnicity Assignment: 2015 SDR Frame

Race/ethnicity Group	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Hispanic	46,114	3,406	37,976	4,732
NH-American Indian	4,157	246	3,494	417
NH-Asian	266,747	11,267	230,139	25,341
NH-Black	36,046	2,628	29,854	3,564
NH-Pacific Islander	984	102	748	134
NH-White	748,937	28,287	677,315	43,335
Overall	1,102,985	45,936	979,526	77,523

URM15 is defined in the following manner:

- If a case is Hispanic or Latino, assign the case to URM regardless of race;
- If a case is not Hispanic (NH) and is American Indian, black, or Pacific Islander, or reports more than one race, assign the case to URM; and
- If a case is not Hispanic or not American Indian, black, or Pacific Islander, and is Asian or White, assign the case to non-URM.

The distribution of the resulting URM15 variable is shown in Table 3.11.

TABLE 3.11 Frame Distribution of Underrepresented Minority (URM15)

URM	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Yes	90,355	6,616	74,171	9,568
No	1,012,630	39,320	905,355	67,955
Overall	1,102,985	45,936	979,526	77,523

BIRCIT15. The BIRCIT15 variable indicates the sample member's citizenship at birth, defined as either U.S. or non-U.S. For all cases in the frame, this information is obtained from the SED. Cases that do not have valid information on birth citizenship are imputed to be non-U.S. The sources for birth citizenship data in the 2015 SDR frame files are detailed in Table 3.12. The distribution of the resulting birth citizenship variable is shown in Table 3.13.

TABLE 3.12 Citizenship at Birth Sources: 2015 SDR Frame

Citizenship at Birth Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported in SED	1,055,996	44,329	938,676	72,991
Citizenship imputed from DRF with BIRTHPL and PDLOC	1,117	32	1,048	37
Default imputation (non-U.S. born)	45,872	1,575	39,802	4,495
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.13 Frame Distribution of Citizenship at Birth

Citizenship at Birth Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
U.S. born	656,847	27,839	589,281	39,727
Not U.S. born	446,138	18,097	390,245	37,796
Overall	1,102,985	45,936	979,526	77,523

SEX15. SEX15 is defined from data in the SED. Cases with missing data on sex are imputed to be female, giving these cases a higher probability of selection. The data sources for the sex variable in the 2015 frame are shown in Table 3.14. The distribution of the resulting sex variable is shown in Table 3.15.

TABLE 3.14 Data Sources for SEX15

Sex Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported	1,101,208	45,880	977,859	77,469
Default imputation (female)	1,777	56	1,667	54
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.15 Frame Distribution of SEX15

Sex Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Male	749,244	29,321	675,528	44,395
Female	353,741	16,615	303,998	33,128
Overall	1,102,985	45,936	979,526	77,523

LOCSTAT15. The LOCSTAT15 variable indicates the last known residence location of the sample member prior to the 2015 SDR; it is either in the U.S. or out of the U.S. The 2010 SDR was the first cycle to use this variable.[8] In the past two cycles, this variable was used to distinguish between NSDR and ISDR cases. Under the 2015 SDR design, it is used as one of the sorting variables. For all cases in the

[8] For more details about the LOCSTAT variable development for the 2010 SDR and continued for the 2013 and 2015 SDR, see the memoranda "2010 SDR Sample Frame Development Memo #3 – Sample Member Location Variable" sent to Daniel Foley and Steve Cohen, NSF, on April 23, 2010 from Karen Grigorian, NORC, and Brenda Cox, SRA, and "2015 SDR Frame Decisions – Frame File Layout" sent to Steve Proudfoot, NSF, on March 28, 2014 from Karen Grigorian and Lance Selfa, NORC.

2015 frame, LOCSTAT15 is derived from responses to the SED question about planned post-graduation location. Any cases with no residency data from the SED are imputed to be in the U.S.

The sources for the location data in the 2015 SDR frame files are detailed in Table 3.16. The distribution of the resulting location variable is shown in Table 3.17.

TABLE 3.16 Location Data Sources: 2015 SDR Frame

Location Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
SED	1,064,194	44,297	948,386	71,511
Default imputation (in the U.S.)	38,791	1,639	31,140	6,012
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.17 Frame Distribution of LOCSTAT15

Location Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
In the U.S.	987,174	40,832	877,149	69,193
Out of the U.S.	115,811	5,104	102,377	8,330
Overall	1,102,985	45,936	979,526	77,523

BIRCIT15. The BIRCIT15 variable indicates the sample member's citizenship at birth, defined as either U.S. or non-U.S. For all cases in the frame, this information is obtained from the SED. Cases that do not have valid information on birth citizenship are imputed to be non-U.S. The sources for birth citizenship data in the 2015 SDR frame files are detailed in Table 3.18. The distribution of the resulting birth citizenship variable is shown in Table 3.19.

TABLE 3.18 Citizenship at Birth Sources: 2015 SDR Frame

Citizenship at Birth Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported in SED	1,055,996	44,329	938,676	72,991
Citizenship imputed from DRF with BIRTHPL and PDLOC	1,117	32	1,048	37
Default imputation (non-U.S. born)	45,872	1,575	39,802	4,495
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.19 Frame Distribution of Citizenship at Birth

Citizenship at Birth Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
U.S. born	656,847	27,839	589,281	39,727
Not U.S. born	446,138	18,097	390,245	37,796
Overall	1,102,985	45,936	979,526	77,523

AGE15. The AGEYR15variable indicates the sample member's year of birth and is used to create AGE15 and IAGE15. The primary sources of AGEYR15 are birth year data reported on the SED, supplemented with birth year information collected from the SDR. Any missing data on AGEYR15 are imputed from sample members' bachelor's degree year, if known, or from their doctorate degree year, which is known for all sample members. The birth year imputation rules assume that sample members are 18 when they earned their bachelor's degree, 21 when they earned their doctoral degree. These age assumptions may not be realistic; they are intended to minimize frame undercoverage which could arise if we eliminate those doctorates who are missing birth year but have earned a doctoral degree at a young age. The sources for age in the 2015 SDR frame files are detailed in Table 3.20. The distribution of the resulting age variable is shown in Table 3.21.

TABLE 3.20 Age Source: 2015 SDR Frame

Age Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported in SED	1,051,578	43,833	935,485	72,260
BA Year Imputation	15,875	652	14,022	1,201
PhD Year Imputation	35,532	1,451	30,019	4,062
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.21 Frame Distribution of Age

Age Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Under 35	111,950	3,445	57,776	50,729
35-39	129,168	6,579	106,668	15,921
40-44	123,227	6,872	110,924	5,431
45-49	123,076	6,063	114,770	2,243
50-54	127,462	5,357	120,737	1,368
55-59	121,617	4,887	115,755	975
60-64	123,245	4,545	118,140	560
65-75	243,240	8,188	234,756	296
Overall	1,102,985	45,936	979,526	77,523

HCAPIN15. The disability status variable, HCAPIN15, is not used in the sampling process, but has been included on the frame and in this reporting section as disability status is important to future reporting and analysis. The HCAPIN15 variable indicates the sample member's disability status – either disabled or not disabled. For all cases in the 2015 sample, the disability information is obtained from the SED, which has gathered data on disability since 1985. The historical data on disability in the DRF are recorded in the variable DISABILITY1. Starting with the 2012 cycle, the SED is using the identical disability question and code frame (summarized in DISABILITY2) as does the SDR. Therefore, defining disability status for frame cases requires using both DISABILITY1 and DISABILITY2. Cases that never reported disability status, including those who completed the SED before the disability questions were introduced to the SED, are imputed to be non-disabled. This imputation means that the proportion of disabled doctorates in the population should be much higher than known on the frame because the cases with unknown disability status have been imputed to be not disabled. The sources for disability status in the 2015 SDR frame files are presented in Table 3.22. The distribution of the resulting disability status variable is shown in Table 3.23.

TABLE 3.22 Disability Status Source: 2015 SDR Frame

Disability Status Data Source	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Self-reported in SED	679,405	30,631	579,270	69,504
Default imputation (not disabled)	423,580	15,305	400,256	8,019
Overall	1,102,985	45,936	979,526	77,523

TABLE 3.23 Frame Distribution of Disability Status

Disability Status Assignment	Total Cases	2013 Panel	Expansion Cohort	New Cohort
Disabled	13,866	559	9,220	4,087
Not disabled	1,089,119	45,377	970,306	73,436
Overall	1,102,985	45,936	979,526	77,523

SUMMARY OF FRAME VARIABLES DATA SOURCES. Table 3.24 summarizes the data sources for the key frame variables subject to imputation. These results are shown by variable and by the three main sample frame components.

TABLE 3.24 Data Sources for Sample Frame Variables Subject to Imputation and/or Derivation: 2015 SDR Frame

		2015 SDR Sample Frame Cases		
Sample Frame Component	Sample Frame Variable	Reported Values in the Final Frame	Imputed from a Non-default Rule	Assigned Default Imputation
2013 Panel	Race (RACE15)	39,825	4,403	1,708
	Ethnicity (HISPANIC15)	40,837	3,420	1,679
	Sex (SEX15)	45,880	n/a	56
	Location (LOCSTAT15)	44,297	n/a	1,639
	Citizenship at birth (BIRCIT15)	44,329	32	1,575
	Disability status (HCAPIN15)	30,631	n/a	15,305
	Birth year (AGEYR15)	43,828	2,108	n/a
Expansion Cohort	Race (RACE15)	836,286	104,391	38,849
	Ethnicity (HISPANIC15)	850,247	86,347	42,932
	Sex (SEX15)	977,859	n/a	1,667
	Location (LOCSTAT15)	948,386	n/a	31,140
	Citizenship at birth (BIRCIT15)	938,676	1,048	39,802
	Disability status (HCAPIN15)	579,270	n/a	400,256
	Birth year (AGEYR15)	935,484	44,042	n/a
New Cohort	Race (RACE15)	71,252	2,197	4,074
	Ethnicity (HISPANIC15)	71,466	817	5,240
	Sex (SEX15)	77,469	n/a	54
	Location (LOCSTAT15)	71,511	n/a	6,012
	Citizenship at birth (BIRCIT15)	72,991	37	4,495
	Disability status (HCAPIN15)	69,504	n/a	8,019
	Birth year (AGEYR15)	72,260	5,263	n/a
Overall	Race (RACE15)	947,363	110,991	44,631
	Ethnicity (HISPANIC15)	962,550	90,584	49,851
	Sex (SEX15)	1,101,208	n/a	1,777
	Location (LOCSTAT15)	1,064,194	n/a	38,791
	Citizenship at birth (BIRCIT15)	1,055,996	1,117	45,872
	Disability status (HCAPIN15)	679,405	n/a	423,580
	Birth year (AGEYR15)	1,051,572	51,413	n/a

4 Sample Design

4.1 Precision Requirements

The SDR sample design has undergone several major changes since its inception in 1973, reflecting changing estimation objectives and budgetary situations. For the past few cycles, the SDR was designed to produce estimates by various analytical domains defined by aggregated field of degree, gender, race and ethnicity, citizenship at birth, and disability status. The existing SDR sample stratification and allocation system reflects these estimation goals. A new significant change to the 2015 SDR design is a major sample size expansion to support employment outcome estimates by fine field of degree (FFOD). The sample size of the 2013 SDR is 47,078 cases, while the current expansion calls for a sample size increase to 120,000 cases for the 2015 SDR.

With the marked increase in the overall sample size, the estimation capability of the 2015 SDR is expected to increase substantially. To guide the SDR sample redesign, the NSF specified general requirements regarding the analytical objectives of the 2015 sample. The following comes from the document "Requirements of Sample Expansion and Sample Redesign of the Survey of Doctorate Recipients" and its addendum (full text in Appendix A):

- **Specified precision at FFOD level**: Producing employment outcome estimates at the SED fine field of degree (FFOD) level for the entire SDR eligible population regardless of their residential location and time of receiving doctorates. The precision is required to be within 5% margin of errors at the 95% confidence level for important outcomes.[9]
- **Maintain precision of key 2013 domains**: The overall expanded sample should maintain the existing 2013 estimation capability at the aggregated degree levels and for domains defined by various demographic characteristics currently used in NCSES publications. This set of requirements will be examined by comparing the estimates' precision levels derived under the proposed designs and the actual 2013 SDR sample results.

The expansion of the SDR, along with its new estimation objectives, calls for a significant redesign of the SDR sample. The objectives of the SDR redesign include meeting the newly defined estimation objectives, resolving any longstanding sampling issues to improve efficiency, creating a unified sample design for NSDR (National SDR) and ISDR (International SDR), and constructing a flexible and

[9] This requirement was later relaxed to a 5% margin of error at the *90%* confidence level.

sustainable design for the growing demands of SDR data. The rest of this section describes design considerations and the final design chosen for the 2015 SDR sample redesign.

4.2 Design Approaches Considered

This subsection briefly discusses the two broad design approached considered for the 2015 SDR design. More detailed descriptions of these approaches and design options are presented in Appendix B.

Broadly speaking, the 2015 SDR frame may be thought of as consisting of two overlapping frames. Frame A, which contains the first four frame groups in Table 3.1, covers the portion of the SDR population that is represented by the existing 2013 SDR sample, or the panel sample. Frame B encompasses Frame A as well as the population that is not represented by the panel sample, namely, frame groups 5 and 6 in Table 3.1, including new cohort doctorates awarded in 2012 and 2013 as well as those that were excluded from the SDR frame prior to 2000. Therefore, Frame A is completely nested within frame B which includes all six frame groups in Table 3.1

Under the guidance of NCSES, NORC considered two major design approaches for the 2015 SDR redesign: dual frame design, and single frame design with a fresh new sample, as discussed below.

Dual Frame Design

The main motivation of the dual frame design is to preserve the existing SDR panel, both to reduce data collection cost and improve data utility. Under this approach, the 2015 SDR sample would include two independent and partially overlapping samples: the panel sample (i.e., the existing 2013 SDR sample) from Frame A and an independent expansion sample from Frame B. For estimation, these two samples would be combined through a dual frame method to derive the overall 2015 estimates.

Let's designate the existing panel sample from Frame A as sample a and the new expansion sample to be selected from Frame B sample b. The two samples would be first be weighted separately according to their respective sample design, generating two sets of sampling weights $w_j^{(a)}$ and $w_j^{(b)}$. Then a single set of weights would be created for the combined sample through a combining factor. For a sample member j selected into either sample, its sampling weight will be calculated as

$$w_j = \begin{cases} w_j^{(b)},\ if\ \ j \in sample\ b\ and\ not\ \ frame\ A \\ \lambda w_j^{(a)}, if\ j \in sample\ a\ and\ \ frame\ A \\ (1-\lambda)w_j^{(b)}, if\ \ j \in sample\ b\ and\ \ frame\ A \\ \lambda w_j^{(a)} + (1-\lambda)w_j^{(b)},\ if\ \ j \in both\ sample\ a\ and\ sample\ b\ and\ \ frame\ A \end{cases}$$

The first category includes cases that are in Frame B but not in Frame A, representing the non-overlapping portion of the frame. For this portion of the population, the estimate will be based only on a subset of sample b. The other three categories include samples from the overlapping portion of the population. For this portion of the population, there are two estimates, one based on sample a and the other based on a subset of sample b.

The combining factor λ is defined as

$$\lambda = \frac{n_{eff}^{(a)}}{n_{eff}^{(a)} + n_{eff}^{(b)}}$$

In this expression, n_{eff}^{a} is the effective sample size associated with sample a selected from Frame A; and n_{eff}^{b} is the effective sample size associated with sample b selected from Frame A. The effective sample size is the expected number of complete surveys divided by the design effect due to unequal weighting.

With a single set of weights defined, the usual Horvitz-Thompson estimator can be used to derive point estimates after proper adjustments for eligibility, nonresponse, and frame coverage.

Under this dual frame estimation approach, the effective sample size from sample a $n_{eff}^{(a)}$, is known for each fine field. Therefore, sample size determination is to estimate $n_{eff}^{(b)}$ such that $1.96 * \sqrt{V(\hat{p})} \le .05$ when estimating a population proportion P for a fine field. The quantity $n_{eff}^{(b)}$ can then be converted to a nominal sample size based on the design effect and expected completion rate.

Single Frame with a Fresh New Sample

Under this design approach, a fresh new sample would be selected from Frame B, and the existence of the panel sample has no bearing on the 2015 SDR design. The sample will be stratified by FFOD only. Sample allocation to the strata is determined to balance the competing estimation goals discussed earlier. For a sample member j, its sampling weight will be

$$w_j = \frac{1}{p_j}$$

where p_j is the inclusion probability under the sample design. The Horvitz-Thompson estimator can be used to derive point estimates after proper adjustments for eligibility, nonresponse, and frame coverage.

4.3 Sample Allocation

We now discuss the Fresh New Sample approach that the NSF decided to adopt. Under this approach, the SDR sample is stratified by FFOD to 216 sampling strata, including the discontinued 20th century fields strata. As discussed below, the sample of 120,000 cases is allocated to the strata in two steps. The two-step allocation is implemented to achieve a balance between satisfying the fine field level estimation requirement and maintaining the existing estimation capabilities of the SDR with regard to the key analytical domains under the prior design. In particular, the second step is intended to improve the representation of the population by the 15-category aggregate degree fields.[10] This measure is implemented because, after step one, aggregate fields with a large population but consisting of a small number of fine fields (e.g., Computer/Information Sciences) are underrepresented, while aggregate fields with a small population but consisting of a large number of fine fields (e.g., Agricultural Sciences) are overrepresented. The objective of the two step allocation is to make the representation of aggregate degree fields more proportional to the population.

NSF decided to allocate 1,000 cases to the stratum that represents the discontinued fields. The remaining 119,000 cases are allocated to the other 215 strata as described in Steps 1 and 2 below.

[10] The 15 categories are: Chemistry, Physics/Astronomy, Earth/Ocean/Atmospheric Sciences, Mathematics, Computer/Information Science, Agricultural Sciences, Medical Sciences, NIH Biological Sciences, Other Biological Sciences, Psychology, Economics, Anthropology/Archeology/Sociology, Other Social Sciences, Electrical/Electronics/Communications, and Other engineering.

Step 1

The first step features an equal allocation to each stratum. For a population proportion centered at 50 percent, the first step allocation is designed to achieve a 5 percent margin of error (MOE) with 90% confidence. The following summarizes how the first step allocation is derived.

- Assume that the population proportion is $\hat{P} = 0.50$ to derive the most conservative sample size estimate;
- The number of complete surveys needed per stratum is $n_{completed} = \left(\hat{P} * \left(1 - \hat{P}\right)\right) * \left(\frac{Z_{90}}{MOE}\right)^2$, where Z_{90} is the critical value of the standard normal distribution for a 2-tailed test at a 90% confidence level (approximately 1.645), and MOE = 5%
- Assume that the completion rate is 70%, the number of cases to sample per stratum is $n_{sampled} = \frac{n_{completed}}{0.70}$
- In strata where the allocated sample size exceeds the number of cases on the frame, the stratum sample size is set to equal to the frame size, $n1_{FFOD} = \min(N_{FFOD}, n_{sampled})$, where N_{FFOD} = Total frame size for the FFOD
- The total sample allocated in step one is $Total\_Step1 = \sum_{FFOD} n1_{FFOD}$

The nominal sample size allocated to each stratum is 387, which will produce 271 complete surveys with an expected completion rate of 70%, enough to satisfy the stratum level precision requirements. Note that the finite population correction factor (FPC) is not incorporated in the sample size estimation described above. When sampling from a finite population, the variance of the mean is reduced by a factor $(N-n)/N$, called FPC, where N is the population size and n is the sample size. For sample size estimation, the application of the FPC helps to reduce the sample size necessary to meet the specified precision requirement. To be conservative, NCSES and NORC decided not to apply the FPC when estimating the sample size per stratum. The effect of ignoring the FPC is to overestimate the standard error of the mean; but it offers additional insurance in case the completion rate is lower than expected.

For fine fields with less than 387 cases in the frame, all frame cases are included in the sample. A total of 77,965 cases, or 65 percent of the total sample, are allocated in the first step. The first step allocation represents the minimum allocation to each stratum and assures that the final sample will achieve the required level of precision at the fine field level.

Step 2

The second step allocates the remaining 35 percent of the sample proportionately to the 15 SDR broad field categories, as represented by the design variable SDRFLD15. This second step allocation is designed to allocate the balance of the sample in such a way as to minimize the variation in sampling weights for the full sample, given the first step allocation. The second step allocation is carried out as follows:

- First, calculate the expected proportional allocation to the 15 broad field categories based on the overall frame distribution across the 15 broad fields. The fine fields within the discontinued 20th century fields stratum also participate in this calculation. The expected proportional allocation to each broad field of degree (BFOD) is: $Expected\_Allocation_{BFOD} = Total\_Sample * \frac{N_{BFOD}}{N}$, where N_{BFOD} is the total number of frame cases per broad field.
- Second, subtract the total first step allocation for each broad field from the expected allocation to get the second step allocation per broad field category. For those broad field categories (Agricultural Sciences and Earth/Ocean/Atmospheric Sciences) that have already exceeded the expected allocation after the first step allocation, allocate 0 cases in the second step. If the step one allocation per broad field is $Step1_{BFOD} = \sum_{FFOD \in BFOD} n1_{FFOD}$, then the step two allocation to each broad field is
- $Step2\_Allocation\_a_{BFOD} = \max(0, Expected\_Allocation_{BFOD} - Step1_{BFOD})$. The step two allocation to the broad field is then adjusted to reflect the fact that two broad fields would not receive additional allocation in step two. The adjusted step two allocation to the broad field is: $Step2\_Allocation\_b_{BFOD} = Step2\_Allocation\_a_{BFOD} * \frac{Total\_Step2}{Total\_Step2\_Allocation\_a}$, where $Total\_Step2 = 119{,}000 - Total\_Step1$, and $Total\_Step2\_Allocation\_a = \sum Step2\_Allocation\_a_{BFOD}$
- Third, within each broad field category, proportionately allocate the second step allocation to each fine field stratum based on the frame count per fine field stratum[11]. This allocation is calculated as $n2_{FFOD} = Step2\_Allocation\_b_{BFOD} * \left(\frac{N_{FFOD}}{\sum_{FFOD \in BFOD} N_{FFOD}}\right)$
- The final allocation to each fine field stratum is the sum of step one and step two allocations or the frame size if the sum exceeds the frame size, i.e., $Final\_Allocation_{FFOD} = \min(N_{FFOD}, n1_{FFOD} + n2_{FFOD})$

[11] The fine fields that make up the discontinued fields stratum do not receive any allocation.

When a fine field does not have enough cases to support the final allocation, the total allocation is equal to the frame total. In that situation, the shortage is allocated to the discontinued field stratum. For this reason, the final allocation to the discontinued stratum is slightly over 1,000. In the final allocation, a total of 118,916 cases are allocated to the 215 fine field strata, with the remaining 1,084 cases allocated to the 216th stratum representing the 20th century discontinued fields. Appendix Table G.1 shows the step one, step two, and total allocation by 2015 sampling stratum. For comparison, Appendix Table G.2 shows the total sample allocation by 2013 sampling stratum.

5. Sample Selection

Within each of the 216 strata, a random sample is selected systematically with probability proportionate to size (PPS). PPS sampling is adopted as a vehicle to oversample underrepresented racial and ethnic minorities (URM), women, and the 2013 panel cases. Oversampling of URM and women allows the sample to sustain the estimation capabilities under the prior SDR design. The addition of a panel oversample maintains the simplicity of a fresh new sample, but allows for limited longitudinal analysis using earlier waves of the SDR. The oversampling is achieved by assigning a measure of size to each frame member and then selecting the sample systematically with PPS within each stratum. Each frame case is assigned a measure of size (MOS), as follows:

- Male URM: 2.0
- Female URM: 2.5
- Female non-URM: 1.5
- Panel cases, regardless of gender or URM status: 5.0
- All other cases: 1.0

Under PPS sampling, the selection probability for a case i in stratum h is $p_{hi} = n_h * MOS_i / \sum MOS_i$, where n_h is the stratum sample size, MOS_i is the measure of size for case i, and the summation is over all frame cases within a stratum. For cases with large MOS, the selection probability may be equal to or greater than 1. Such cases are identified first because they would be selected with certainty into the sample. For each stratum, the certainty cases are identified as follows:

1. Sort the frame cases in descending order by MOS;
2. Sum MOS across all frame cases to get the total MOS *Total_MOS*;
3. Denote the total allocated sample size as *Total_Allocated* ;
4. Carry out the following procedures, starting at the top of the sorted frame
    a. If MOS ≥ (Total_MOS/Total_Allocated) then this case is a certainty selection:
        i. Set sampling weight = 1;
        ii. Move the case to a separate file that contains all certainty cases;

   iii. Recompute Total_MOS = Total_MOS – MOS;

   iv. Recompute Total_Allocated = Total_Allocated – 1;

   v. Return to 4.a to evaluate the next case on the sorted frame.

b. If MOS < (Total_MOS/Total_Allocated) then this case and all cases following this case are not certainty cases

   i. All non-certainty cases constitute the frame for systematic PPS sampling;

   ii. The final sample consists of the certainty cases and those selected from the rest of the frame through systematic PPS sampling.

Before systematic sampling, the frame is sorted by the following variables to impose an implicit stratification within each stratum. The sorting variables are:

- CENTURY
- RACETH15
- BIRCIT15
- SEX15
- LOCSTAT15
- SDRFLD15
- PHDFY

The purpose of implicit stratification is to improve the representation of the sample with respect to the sorting variables. Note that sorting by SDRFLD15 (the 15 broad fields) is only effective in the discontinued stratum because all the other strata represent a single field. The purpose of sorting by SDRFLD15 within the discontinued fields is to ensure a proportional representation of the broad fields within the discontinued stratum.

With the certainty cases set aside, the SAS procedure PROC SURVEYSELECT is used to carry out the systematic sampling within each stratum. Systematic sampling selects cases at a fixed interval throughout the stratum after a random start. PROC SURVEYSELECT uses a fractional sampling interval to provide exactly the specified sample size. The interval within a stratum is $\frac{\sum MOS_i}{n}$. The expected number of hits (selections) for a case is $\frac{n * MOS_i}{\sum MOS_i}$. The sampling weight is the inverse of the expected number of hits. The final sample includes all the certainty cases and those selected through PROC SURVEYSELECT.

Subsequent to selection, the selected sample along with the quality assurance procedures were sent to the NSF for review and approval.[12]

[12] See the memo sent to Emilda Rivers and Steve Proudfoot at the NSF from Michael Yang, Lance Selfa, and Karen Grigorian at NORC entitled “2015 SDR – Sample Selection Quality Control Procedures and Results” issued on 16 February 2015 as well as the sample review tables in the attachment file named ” 2015 SDR Expansion Sample Allocation and Review Tables.zip.”

6. SDR 2017 and Beyond

The NSF adopted the fresh new sample approach to meet new estimation goals, to simplify the sample design and estimation procedures, and to resolve longstanding frame and sampling issues accumulated over time. With a fresh new sample, the 2015 SDR represents a significant turning point in the hitherto longitudinal sample design dating back to 1973. Although the SDR sample has undergone several major redesigns, for the first time since the 1975 SDR, the new sample does not include a substantial panel component. Instead, the 2015 SDR is expected to be the starting point of a new longitudinal data series. The NSF has not yet provided any guidelines for the 2017 SDR sample design, but it is most likely that the 2015 SDR sample will form the sampling frame for the old cohort sample for 2017 while a new cohort sample will be selected from the new cohort frame consisting of SEH doctorates awarded in 2014 and 2015 academic years. Unless the analytical objectives change, we expect the 2017 SDR to follow the same stratification scheme and sample collection procedures.

Assuming that the 2017 SDR sample size will be kept at the current level, a maintenance cut to the old cohort sample will be necessary while adding a new cohort sample. To preserve the oversampling of URM, women, and the 2013 panel cases, we expect the 2017 SDR old cohort sample to be a straightforward equal probability random sample within each stratum. Like the 2015 SDR, the 2017 new cohort sample will be stratified by FFOD, and the sample allocation will be guided by the analytical objectives specified by the NSF. If the analytical objectives stay the same, for example, the 2015 allocation procedures may be adapted to allocate the 2017 new cohort sample.

7. Concluding Remarks

The current SDR sample redesign may have significant implications for post-survey data processing procedures such as weighting adjustment, missing data imputation, and variance estimation. We conclude this report by discussing these likely implications. Additional research will be needed to modify these procedures if they turn out to be necessary.

Starting from the 2010 cycle, the SDR has moved from the traditional weighting class method to the model-based propensity score method for noncontact and nonresponse weighting adjustments. A propensity score is predicted from a logistic regression model for both eligibility determination and interview cooperation; these scores are then used, either directly or indirectly, to adjust the original sampling weight to compensate for noncontact and nonresponse prior to a final poststratification adjustment. Given the changes to the sample design, the 2015 noncontact and nonresponse models are likely to be different from the models of the prior rounds. For example, additional predictor variables may need to be included to capture the noncontact and nonresponse pattern associated with the expansion cohort cases that appear in the SDR sample for the first time. Furthermore, with the newly defined estimation goals, the poststratification procedures may also need to be revamped to match the poststrata with analysis domains. For example, it may be necessary to define the poststrata by fine field of degree, among other key factors.

The SDR conducts extensive missing data imputation, using basically the same set of imputation procedures in the past few cycles. With the sample redesign, these imputation procedures may need to be adapted to the 2015 SDR. For example, many variables are imputed through multivariate regression models; these models may need to be updated to reflect the new sampling frame and relevant features of the sample design. The sorting variables under hot deck imputation may also need to be updated by new variables and new models.

The SDR has used a replication method for variance estimation to account for its complex design features that cannot be adequately captured with the Taylor Series linearization method. The current successive difference replication method (SDRM) is designed for systematic samples where implicit stratification puts similar cases close to each other on the sampling frame. Under the existing strategy initially developed by the Census Bureau, the replicates are formed as if a systematic sample is selected from a single large stratum. While this method may effectively account for the reduction in variance resulting from implicit stratification within the explicit strata, it may not account for the impact of explicit stratification on the sampling variance. NORC has proposed to the NSF to explore alternative variance estimation approaches to improve potentially both statistical and cost efficiency. With the 2015 sample redesign, it may be a good time to revisit the current procedures. For example, we would like to compare the SDRM with a simpler and more efficient procedure based on a Taylor Series or Jackknife method. In case the SDRM does not lead to noticeable reduction in the variances, the Taylor series or Jackknife methods would make more efficient alternatives.

In addition to SDRM replicate weights, the SDR also provides estimated Generalized Variance Functions (GVFs) for a set of key NSDR and ISDR domains. The GVFs are valuable because they provide a mechanism for data users to compute the variance of estimates not directly provided by the SDR. With the sample redesign, it may be necessary to redefine the GVF definitions so that they match with key analysis domains.

Appendices removed