SF-83-1 SUPPORTING STATEMENT (Part B)

for the

2021

Survey of Doctorate Recipients

TABLE OF CONTENTS

B. COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

1. RESPONDENT UNIVERSE AND SAMPLING METHODS

The 2021 SDR sample size is set at 125,938 cases including 115,937 returning sampled cases from the last survey cycle, and 10,001 new cohort sampled cases who are recent doctorate recipients from academic years 2018 and 2019.

Approximately 40,000 of the 2021 SDR sample who participated in the 2015 SDR make up the SDR longitudinal sample representing the 2015 SDR target population moving forward into the 2021 survey cycle of data collection. This panel will be weighted and maintained up through the 2025 cycle of the biennial SDR to provide longitudinal data for the 10-year time period 2015-2025. (See "Consultation Outside the Agency" within Section A.8 for further background information on its development).

1.1 Frame

The source of the primary sampling for the SDR is the Doctorate Records File (DRF). The DRF is a cumulative file listing research doctorates awarded from U.S. institutions since 1920. It is updated annually with new research doctorate recipients through NCSES's Survey of Earned Doctorates (SED). The 2019 SDR sample selected from the 2017 DRF represented a surviving population of nearly 1.2 million Science, Engineering, and Health (SEH) doctorate holders less than 76 years of age. The 2021 SDR is expected to represent about 1.23 million SEH doctorate holders from the 2019 DRF, including over 85,000 from the two most recent academic years, 2018 and 2019. In total, the 2019 DRF contains 2,324,954 records.

The target population for the 2021 SDR includes individuals who must:

- Have earned a research doctoral degree in a SEH field from a U.S. institution, awarded no later than academic year 2019, and
- Be less than 76 years of age on 1 February 2021 based on their month and year of birth, and
- Be living in a noninstitutionalized setting on 1 February 2021, and not terminally ill.

The final 2021 SDR sampling frame can be classified into four target populations as shown in Table 1.

1. Frame Group 1 contains individuals eligible for the 2015 SDR target population. These cases were awarded doctorate degrees in academic years 2013 and earlier.
2. Frame Group 2 contains individuals that became newly eligible for inclusion in the 2017 SDR target population. These cases were awarded doctorate degrees in the 2014 and 2015 academic years.
3. Frame Group 3 contains individuals that became newly eligible for inclusion in the 2019 SDR target population. These cases were awarded doctorate degrees in the 2016 and 2017 academic years.
4. Frame Group 4 contains individuals that became newly eligible for inclusion in the 2021 SDR survey cycle. These cases were awarded doctorate degrees in the 2018 and 2019 academic years.

Table 1: The 2021 SDR Frame Groups by Sample Component

Frame Group	Description	SED Academic Years (AY)	Population Size	Sample Component	Sample Size
1	2015 SDR target population that remain eligible for 2021	1960-2013	974,717	2015 SDR returning sample	81,745
				2015 Supplemental sample	13,514
2	2017 SDR newly sampled cases that remain eligible for 2021	2014-2015	82,372	2017 new cohort	10,717
3	2019 SDR newly sampled cases that remain eligible for 2021	2016-2017	83,211	2019 new cohort	9,961
4	New cohort cases from SED AY 2018 and 2019	2018-2019	85,148	2021 new cohort	10,001
Total			1,225,448		125,938

Based on the prior three SDR cycles, NCSES expects approximately a 70% response rate in each sample component in 2021.

1.2 2021 Sample Design

In the 2015 survey cycle, the SDR sample size increased from 45,000 to 120,000 individuals. The goal of the large sample size increase was to improve the precision of estimates for key analytic domains of interest, especially the fine field of degree (FFOD) categories reported in the SED. Over 200 FFODs served as the explicit sampling strata in the sample design for both the 2015 and 2017 cycles. For the 2019 survey cycle, adjustments to the SDR sample design were made based on feedback from SDR stakeholders in combination with evaluations of the reliability and utility of the 2015 and 2017 estimates at the 200+ FFOD stratification levels. As with the 2019 SDR, the 2021 SDR now stratifies the sample frame by 77 detailed fields, and sex and minority status, rather than only roughly 220 FFODs used in the 2015 and 2017 SDR cycles. The 2021 survey cycle will use the same general approach that was used in 2019. The 2019 sampling design had the following modifications:

- In 2019, the 2015 and 2017 sampling strata of over 200 FFODs were replaced by a set of 308 sampling strata defined by crossing 77 detailed fields of degree (DFOD) with gender (2 categories: male and female) and underrepresented minority (URM) status (2 categories: URM and non-URM). These stratification changes were designed to produce sustainable and reliable estimates for population subgroups that can be supported by the sample size and are aligned with the NCSES taxonomy of disciplines (TOD).
- The second adjustment in 2019 was to sample allocation within strata to achieve estimation precision requirements for several types of domains. Instead of setting the precision requirements at a single FFOD level as was done in the 2015 and 2017 cycles, the estimation precision requirements for 2019 were set for three levels of aggregation over the 308 sampling strata. The 2021 sample will use the same estimation precision requirements as in 2019 and are shown in Table 2.

Table 2: The 2021 SDR Overall Precision Requirements

Domain	Margin of Error	Minimum Number of Completes
DFOD	5%	270
DFOD x SEX	6%	190
DFOD x URM	7%	135

In Table 2, the margin of error in the first column is two times the standard error associated with estimating a population proportion of 50% at the 90% confidence level. The second column shows the required minimum number of completed surveys to achieve the precision requirement per domain. Based on these constraints, minimum sample sizes for each sampling strata could be determined. Finally, the allocation was performed by finding an allocation that is as close as possible to the proportional allocation, subject to the minimum sample size constraints.

- The third adjustment was to the returning cohort's retention rule. For the 2019 SDR survey cycle, sample cases were dropped from subsequent cycles if they did not respond in the preceding two cycles of data collection after entering the SDR. This was a change from the 2015 and 2017 cycles where all previously sampled cases were carried forward regardless of prior response status. This decision was made in response to cost and inefficiencies in carrying these nonresponding cases forward in subsequent survey cycles. Because fewer than 1,500 cases met the 2-cycle nonresponse criteria at the conclusion of the 2019 cycle, this adjustment to the retention rule will not apply in 2021. Therefore, all sampled cases from the 2019 cycle who remain age eligible will be carried forward into the 2021 sample.

As in the 2019 cycle, the 2021 sampling design includes oversampling of underrepresented minorities and women, allowing the SDR sample to sustain and strengthen the estimation capabilities of the 2013 and prior cycles' SDR design based on smaller overall sample sizes.

As noted above, the 2021 SDR sample consists of 115,937 individuals who were included in the 2019 SDR sample and remain eligible for this cycle (i.e., less than 76 years of age, not permanently institutionalized, etc.). In addition, the 2021 sample includes 10,001 individuals sampled from among those who obtained their SEH doctorate degree since the 2019 SDR sample selection. This sample, referred to as the 2021 "new cohort" sample, follows the same stratification design as the 2019 sample and applies similar sample allocation for the overall sample to the new cohort sample of 10,001 individuals. Furthermore, when drawing the new cohort sample, demographic variables such as race/ethnicity categories, citizenship at birth, predicted resident location, disability status, age group, and doctorate award year, are used as implicit sorting variables within each stratum to improve their representation in the sample. The new cohort sample is then drawn systematically with equal probability within each stratum (See appendix F). After combining the continuing and new cohort, the 2021 SDR sample will consist of 125,938 individuals.

Furthermore, the continuing cohort also contains a longitudinal sample of 40,000 individuals that were selected in 2019 from among sample respondents to the 2015 SDR who were less than 66 years old on February 1, 2015, the survey reference date. This longitudinal sample will continue to be followed in 2021 and in the 2023 and 2025 survey cycles. Because these cases are part of the cross-sectional sample, data collection, editing, and other processing steps will not be treated differently. However, the longitudinal sample requires statistical procedures that differ from those used for the cross-sectional sample, as described in section 2 below.

2. STATISTICAL PROCEDURES

The SDR statistical data processing procedures have several components including sampling weight adjustments to compensate for the stratified sampling design features and differential response rates, imputation procedures to address item nonresponse, and estimation procedures for calculating sampling error.

2.1 Weighting

A final weight will be computed for each completed interview in the cross-sectional sample including its longitudinal sample cases. These weights are intended to be used to conduct cross-sectional statistical analysis of the data from all 2021 SDR respondents so that the results represent the eligible population of doctorate recipients (i.e., individuals who earned a research doctoral degree in a science, engineering, or health field from a U.S. institution awarded no later than academic year 2019 and are less than 76 years of age). The weighting procedures consist of a series of statistical adjustments to the original sampling weights and will follow methods similar to those applied in the development of the 2019 SDR weights. These methods are briefly described below.

For a sample member j, its original sampling weight will be computed as

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

where p_j is the inclusion probability under the sample design.

The sampling weight will be adjusted in sequence for unknown eligibility, unit nonresponse, and frame coverage based on similar methodologies developed for the 2019 SDR. First, for cases whose eligibility status is not determined by the end of the survey, their assigned base weights are transferred to cases whose eligibility is known. Next, among known eligible cases, the weights of nonrespondents are transferred to the respondents so that the respondents represent all eligible cases in the sample. Finally, a raking adjustment aligns the sample to the frame population so that the sample estimates agree with the frame counts with respect to factors not explicitly controlled for in the sample design.

As in the 2019 SDR, logistic regression models will be used to derive unknown eligibility and nonresponse weighting adjustment factors for different segments of the sample. Resulting propensity scores will be used to define weighting classes, and extreme weights will be trimmed to reduce the variation of the weights prior to raking. With a final weight, the Horvitz-Thompson estimator will be used to derive point estimates for SDR variables.

In addition to the weights for the cross-sectional sample, weights will also be created for the longitudinal sample. Methods similar to those used for the 2019 longitudinal weighting will be followed. Because this sample is drawn from the 2015 SDR respondents, the target population for these weights is the estimated 860,264 individuals who had received a research doctoral degree in a science, engineering, or health field from a U.S. institution by June 2013 and are less than 66 years of age on 1 February 2015. As with the cross-sectional weights, the longitudinal weighting procedures consist of a series of statistical adjustments to their sample weights as well. To account for the two-phase nature of the sample, the sampling weights are the final 2015 cross-sectional weights divided by the selection probabilities associated with inclusion into the longitudinal sample. These sampling weights are then adjusted for unknown eligibility and nonresponse, similar to the procedures for the cross-sectional sample. Finally, they are adjusted by raking to the 2015 frame totals noted above and to the cross-sectional population estimates of the 2017, 2019 and subsequent 2021 SDR.

2.2 Item Nonresponse Adjustment

Historically, the SDR has conducted comprehensive imputation to fill in item-level missing data in the cross-sectional sample. Two general methods of imputation, logical imputation and hot deck imputation, have been used. The logical imputation method is employed during the data editing process when the

answer to a missing item can be deduced from past data, or from other responses from the same respondent. For those items still missing after logical imputation, a hot deck imputation method is employed. In hot-deck imputation, data provided by a donor respondent in the current cycle is used to impute missing data for a respondent who is similar to the donor respondent based on a propensity model. The 2021 SDR will use similar imputation techniques, although the actual imputation models may differ since we will have additional data from the 2019 cycle to identify donors, instead of only considering same-cycle (2021) data.

For the longitudinal sample, item-level imputations from the cross-sectional sample will be retained for respondents to each cycle year. If a full cycle of data is missing for longitudinal sample members, a combination of logical and hot-deck imputation will be used to fill in complete information for the missing cycle year. In the longitudinal hot-deck imputation method, donors are selected based on similarity in key variables in the observed cycle years (i.e., responses in any of the 2015, 2017 and 2019 cycles).

2.3 Variance Estimation

The SDR has used the Successive Difference Replication Method (SDRM) for variance estimation since 2015. The SDRM method was designed to be used with systematic samples when the sort order of the sample is informative. This is the case for the 2021 SDR, which employs systematic sampling after sorting cases within each stratum by selected demographic variables. As in prior cycles, a total of 104 replicates will be used for both the cross-sectional and the longitudinal samples for the 2021 SDR. Within each replicate, the final weight is developed using the same weighting adjustment procedures applied to the full sample. In the case of the longitudinal sample, the two-phase nature of the sampling weights will be incorporated into the variance estimation by applying the raking step for each replicate to control totals that are derived from the cross-sectional replicates instead of the fixed control totals used for the cross-sectional sample. The SDRM replicate weights can be used to estimate the variance of point estimates by using survey variance estimation software packages such as SAS or R.

3. METHODS TO MAXIMIZE RESPONSE

3.1 Maximizing Response Rates

The weighted response rate for the 2019 SDR was 69% (unweighted, 68%). To attain a targeted response rate of 70% for 2021, extensive locating efforts, nonresponse follow-up survey procedures, and targeted data collection protocols will be used during data collection. In addition, both an early-stage and late-stage monetary incentive will be offered as outlined in section A.9 above and section B.4 below.

3.2 Locating

Continuing sample members who are categorized as locating problems in 2021 and new sample members with incomplete contacting data will first need to be located before making a request for survey participation. The 2021 SDR will follow a locating protocol similar to the approach implemented in prior cycles. The contacting information obtained from the 2019 SDR and prior cycles will be used to locate and contact the continuing sample members; the information from the SED will be the starting information used to locate and contact the new sample members in 2021.

2019 SDR Locating Protocol Overview. As in prior SDR cycles, there will be two phases of locating for the 2021 SDR: prefield locating and main locating. Prefield locating activities include batch processing of

sampled cases through LexisNexis® (formerly, Accurint®)[1] and online searches, address review, and individual case locating (also called manual locating). Prefield locating occurs approximately three months before the start of data collection and is used to ensure the initial invitational outreach by mail and email requesting survey participation is sent to as many sample members as possible. Prefield individual case locating includes online searches, limited telephone, mail and email contacts to sample members, and telephone calls and emails to contact persons who may know how to reach the sample members. No more than one mail or one email contact with sample members will be attempted as part of prefield locating. Main locating includes manual locating and additional LexisNexis® processing as needed. Main locating activities will begin at the start of data collection and will include contact (by mail, telephone, or email) with sample members and other contact persons. Both the prefield and main locating activities will be supported by an interactive (i.e., real time) online case management system (CMS). The case management system will include background information for each case, all the locating leads, all searches conducted, and all outreach attempts made which lead to the newly found contacting information (including mailing addresses, telephone numbers, and email addresses). CMS information also will be integrated with survey paradata and monitoring metrics that support an adaptive design approach (See section B.4.4. below for more information on the adaptive design plans for 2021).

Prefield Locating Activities. The prefield locating activities consist of four major components as follows:

1. For both the returning sample component and the new cohort sample component, the U.S. Postal Service's (USPS) automated National Change of Address (NCOA) database will be used to update addresses. The NCOA incorporates all change of name/address orders submitted to the USPS nationwide for residential addresses; this database is updated biweekly. The NCOA database maintains up to 36 months of historical records of previous address changes. However, the NCOA updates will be less effective for the new sample because the starting contacting information from SED could be up to three years out of date.

2. After implementing the NCOA updates for the returning panel and new cohort component, the sample will be assessed further to determine which cases require prefield locating. This assessment is different for the returning panel cases than for the new cohort sample component. Prefield locating will be conducted on panel cases which could not be found in the prior round of data collection or ended the round with unknown eligibility (meaning we could not confirm if the sample member received our contacts). A LexisNexis® batch search also will be run on the returning cohort using the available prior survey cycle information as necessary.

3. For the new cohort, a LexisNexis® batch search will be run using the available information provided in the 2018 and 2019 SED. The returned results will be assessed to determine which cases are ready for contacting and which require further prefield locating. There are four potential data return outcomes from the LexisNexis® batch search for both the returning panel and the new cohort:

    a. Returned with a date of death. For those cases that return a date of death, the mortality status will be confirmed with an independent online source and finalized as deceased. When the deceased status cannot be confirmed, the cases will be queued for manual prefield locating

[1] Accurint® is a widely accepted locate-and-research tool available to government, law enforcement, and commercial customers. Address searches can be run in batch or individually, and the query does not leave a trace in the credit record of the sample person being located. In addition to updated address and telephone number information, Accurint® returns deceased status updates.

and the possible deceased outcome will be noted in the case record so further searching on the possible date of death may be conducted.

b. Returned with existing address confirmed. For cases where LexisNexis® confirms the prior survey data or the SED address as current (i.e., less than two years old), the case will be considered ready for data collection and will not receive further prefield locating.

c. Returned with no new information. For cases where LexisNexis® provides no new information or the date associated with new contacting information is more than two years out of date, the cases will be queued for manual prefield locating.

d. Returned with new information. When LexisNexis® provides new and current contacting information, the new information will be used and the case will be considered ready for data collection with no further prefield locating.

4. The manual locating effort throughout prefield locating involves a specially trained locating team that will conduct online searches and make limited calls to sample members and outreach to contact persons for those individuals not found via the automated searches. Only publicly available data will be accessed during the online searches. The locating staff will use search strategies that effectively combine and triangulate the sample member's earned degree and academic institution information, demographic information, prior address information, any return information from LexisNexis®, and information about any nominated contact persons. Locators will search employer directories, education institutions sites, alumni and professional association lists, white pages listings, real estate databases, online publication databases (including those with dissertations), online voting records, and other administrative sources. Locating staff will be carefully trained to verify they have found the correct sample member by using personal identifying information such as name and date of birth, academic history, and past address information from the SED and the SDR (where it exists).

   Additionally, the 2021 SDR will use LexisNexis® to conduct individual matched searches, also known as AIM searches. AIM allows locators to search on partial combinations of identifying information to obtain an individual's full address history and discover critical name changes. This method has been shown to be a cost-effective strategy when locating respondents with out-of-date contact information in prior SDR cycles as well as other studies. The AIM searching method will be implemented by the most expert locating staff and will be conducted on the subset of cases not found with regular online searches.

Main Locating Activities. Cases worked in main locating will include those not found during the prefield locating period as well as cases determined to have outdated or incorrect contacting information from failed 2021 data collection outreach activities. Prior to beginning the main locating work, locating staff who worked during the prefield period will receive refresher training that focuses on maintaining sample members' confidentiality particularly when making phone calls, or supplementing online searches with direct outreach to the sample members and other individuals, and gaining the cooperation of those sample members and other individuals successfully reached. The locating staff will continue to use and expand upon the online searching methods from the prefield period and, ideally, gain survey cooperation from the found individuals. In addition to outreach to sample members, main locating activities during data collection will include calls and emails to dissertation advisors, employers, alumni associations, and other individuals who may know how to reach the sample member.

3.3 Data Collection Strategies

As with prior cycles, the 2021 SDR will continue using a multi-mode data collection protocol including self-administered web forms, mailed paper self-administered questionnaires (SAQ), and computer-assisted telephone interviews (CATI) to facilitate survey participation, data capture, and sample member

convenience. The 2021 SDR data collection protocols and contacting methods build upon the methodology used in prior cycles, reflecting NCSES' objective of increasing alignment with their other similar surveys, such as the National Survey of College Graduates (NSCG). Therefore, the 2021 SDR will mimic the 2021 NSCG in its contacting protocol (i.e., types and relative timing) in the fielding of the survey over its 6-month period of data collection. Like NSCG, the 2021 SDR data collection field period will include five phases: invitational, reminder, additional mode, late, and CIO (critical item only phase). The general contact strategy for sample members believed to live within the US is to send a postal mailing every four weeks, starting at week 1. The format and content of mailings will differ across each mailing. An email will go out within a few days of each mailing, and a reminder email will go out about a week later. An additional email will be sent in week 15, about halfway through the data collection period, to sample members who started but have not yet completed the online questionnaire. Calls prompting response will occur in week 4 targeting sample members with missing or unconfirmed mailing and email addresses. Similar prompts will occur in week 7 for those with missing email addresses even if they have confirmed mailing address if the sample member has not yet responded. Sample members believed to reside outside of the US will receive a reduced number of mail contacts (3 rather than 7) recognizing the delays in delivery overseas. However, the email and phone contacts will be the same regardless of location.

In the invitation stage, all sample members will initially be encouraged to respond to the 2021 SDR via the online questionnaire as was done in 2019. Sample members who participated in 2019 will initially receive an email inviting them to participate, followed by a postal letter a week later encouraging them to participate in the 2021 SDR by completing the online questionnaire. New 2021 sample members, and continuing sample members who did not respond in 2019 will receive their survey invitation by postal letter first, followed by an email invitation. The invitation postal letter will include the 2021 SDR URL and their Personal Identification Number (PIN). The invitation email will include a live link that will take the sample member directly to the starting page of the SDR 2021 web instrument. In 2019, 93% of respondents participated via the online questionnaire, an increase of more than 10 percentage points from the prior cycle. NCSES expects this trend to continue in 2021.

In the additional mode phase, hard-copy paper questionnaires will be offered at week 9 to those who completed a paper questionnaire in 2019 (about 5% of all respondents completed 2019 surveys). In addition to these targeted mailings, the contractor will send paper questionnaires upon request throughout the data collection cycle. CATI response will be offered at week 12 for sample members who completed by CATI in 2019 (about 2% of all respondents in 2019). Additionally, CATI interviewers will attempt to conduct interviews with a subset of nonresponding sample members in key analytic groups identified by the adaptive design model with below target threshold response rates (See section 3.4 for description of the adaptive design model).

The last stage of the contacting protocol, the Critical Item Only (CIO) phase, offers an abbreviated version of the 2021 SDR as a method to motivate participation among these most reluctant sample members who may not have time available to complete the full survey. The CIO instrument has been used in the prior SDR cycles as a method to motivate response as a last call for participation. Sample members in key analytic groups with response rates below target thresholds will be eligible for a CIO interview in this stage.

As with the 2019 SDR and prior cycles, sample members who were a hostile refusal in a prior cycle will receive limited contact requesting their participation in the 2021 SDR. We will only contact the prior round hostile refusals with the initial survey invitation letter and email. In 2019, the contractor fielded initial invitations only to 84 sample members who had closed out in the 2017 cycle as hostile refusals and six completed the survey. Table 3 shows the maximum number of contacts to be made by mode, cohort and location (domestic or international) for all other sample members.

Table 3: Maximum number of contacts by cohort, location, and mode of contact

Cohort	Domestic		International	
	Mail	Email	Mail	Email
New	7	15	3	15
Continuing	7	15	3	15

3.4 Incentive Plan for 2021

As with the 2019 SDR protocol, the 2021 protocol includes an early-stage and a late-stage incentive for U.S.-residing nonrespondents to reduce the potential for nonresponse bias. Sample members determined to be out of the U.S. will be excluded from the SDR incentive offer, as will those who work for the National Science Foundation. The 2021 SDR will use prepaid debit cards as the incentive, a change from prior cycles which used checks. Internationally residing sample members who are not eligible for a monetary incentive but who have low response propensity or who are in an underperforming analytic domain will be offered the abbreviated version of the questionnaire referred to as the Critical Items Only (or CIO) survey as an additional motivator to participate.

In 2021, NCSES will include an incentive experiment with the goal of reducing the number of incentives offered and the total dollar amount of incentives used in this and subsequent SDR surveys. In 2019, 36% of the full sample was offered an incentive. By comparison, in 2017, only 23% of the sample was offered an incentive. The increase in 2019 resulted from a significant reduction in the level of verified contact information (including verified mailing addresses) and the associated lower response rate at the start of the late-stage incentive offer. The proposed incentive experiment for the 2021 SDR is expected to reduce the proportion of cases offered an incentive to approximately 25%. The experimental design will assess the effect on response rates of reducing or eliminating monetary incentives in the early stage of data collection, while keeping the primary purpose of the incentive focused on increasing the representativeness of the sample. For details about the planned use of incentives in the 2021 SDR, refer to Supporting Statement Part A, section 9.

4. TESTING OF PROCEDURES

The SDR and NSCG are complementary workforce surveys. Therefore, the two surveys must be closely coordinated to provide comparable data. Many of the questionnaire items in the two surveys are the same, including the reference date of 1 February 2021.

The complementary survey questionnaire items are divided into two types of questions: core and module. Core questions are defined as those considered to be generic to both the SDR and NSCG. These items are essential for sampling, respondent verification, basic labor force information, and NCSES analyses of the science and engineering workforce. SDR and NSCG surveys ask core questions of all respondents each time they are surveyed to establish baseline data and to update the respondents' labor force status, changes in employment, and other characteristics. Module items are special topics that are asked less frequently on a rotational basis. Module items provide the data needed to satisfy specific policy or research needs.

As in the 2019 SDR, sample members living inside and outside of the U.S. will receive the same questionnaire content. However, the 2021 questionnaire will reflect two types of modifications relative to the prior cycle.

- First, in alignment with the National Survey of College Graduates, the 2021 SDR will include modifications to collect data that speaks to the effect of the coronavirus pandemic on the sample members' employment situation. NCSES has modified the response categories for five of the current SDR employment questions and has added follow-up questions to the salary and earned income items, as well as added a new measure regarding telework. Each of these changes will allow NCSES to collect information important to understanding some of the employment and related economic implications of the coronavirus on this unique population. See Appendix D.2 for the set of questions that have been added or changed from the 2019 form found in Appendix D.1.

- Second, based on promising results from an experiment conducted in the fall of 2020 to test dependent interviewing methodologies to strengthen the SDR longitudinal component, the 2021 instrument will have two versions of the electronic web and CATI questionnaires: a prefilled form, and a standard form without prefills.

  This examination of dependent interview methodologies and inclusion of a prefilled form into the 2021 SDR data collection effort responds to the following terms of clearance from the 2019 SDR Information Collection Request:

  Approved consistent with the understanding that the design continues to support biennial cross-sectional estimation (including collection for the 2019 reporting period) as well as establishing the baseline for a new longitudinal component. In the period between the 2019 data collection and the planned 2021 data collection, a follow up instrument will be designed for the longitudinal component.

  As in 2019 and prior cycles, the paper SAQ will continue to use the standard form without prefills. The two electronic versions will cover the same topics and constructs as the paper form but will field a dependent interview approach as follows:
  - Continuing sample members who indicated that they were working in their last cycle of participation in the SDR will see prefilled responses from their last cycle for several employment-related questions. Respondents are asked to review their response from their last cycle and then asked to update their information as necessary in reference to February 1, 2021.
  - New cohort sample members will receive the traditional SDR version of the instrument without prefills, as will continuing respondents who reported that they were not working on the reference date in their prior cycle of participation.

  The questionnaire items that will use a dependent interviewing approach as part of the 2021 SDR are as follows:
  - Questionnaire item A9. Name, department, and address of principal employer.
  - Questionnaire item A10. Employer's main business.
  - Questionnaire item A11. Employer size
  - Questionnaire item A13. Type of principal employer (sector of employment)
  - Questionnaire item A14. Indicator for educational institution employer*
  - Questionnaire item A15. Type of educational institution employer

The analytic report from the 2020 SDR Dependent Interviewing Survey Pilot Study is included in Appendix G.

4.1 Survey Contact Materials

Survey contact materials will be tailored to fit sample member's current information from paradata on location and past participation to gain their cooperation in 2021. Contact materials that request sample member participation via the web survey will include access to the survey online. As has been done since 2003 SDR, the 2021 SDR letterhead stationery will include project and NSF/NCSES website information, and the data collection contractor's project toll-free telephone line, USPS, and email addresses. Stationery will contain a watermark that shows the survey's established logo as part of an effort to brand the communication to sample members for ease of recognition. The back of the stationery will display the basic elements of informed consent. See Appendix E for draft copies of the contacting materials.

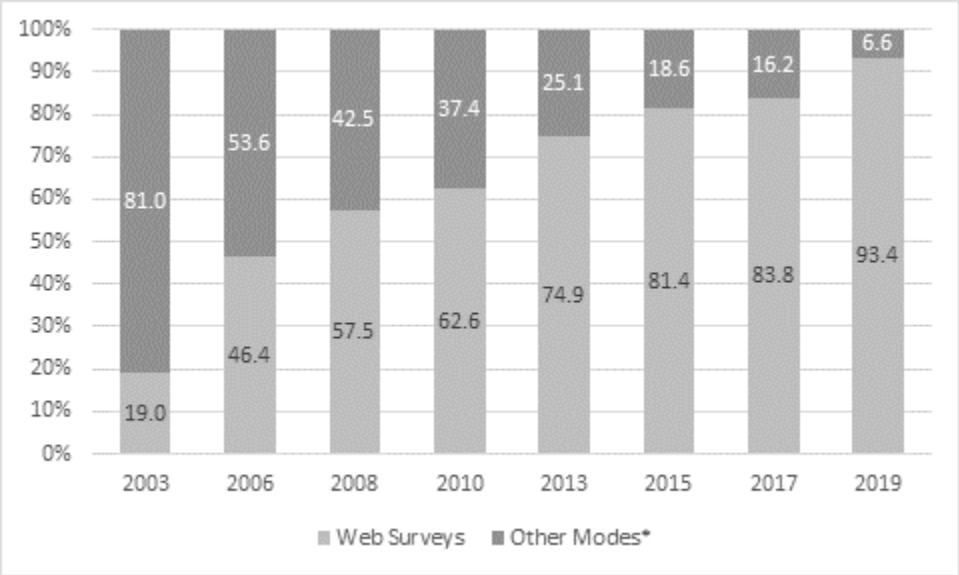
4.2 Questionnaire Layout

There are no changes from the 2019 SDR questionnaire layout for the 2021 survey. Through cognitive research, testing, and other policy relevant interests such as the pandemic, NCSES continues to review and revise the content of its survey instruments. After the 2019 data collection, NCSES made minor modifications to question lead-ins and response categories common to both the NSCG and SDR to increase consistency between the two surveys. NCSES will review the information after the 2021 round and will propose and test changes and content improvements for the 2023 survey cycle.

4.3 Web-Based Survey Instrument

In the 2003 SDR, the online mode was introduced. Figure 2 shows the rate of SDR web survey participation from the 2003 through 2019 survey cycles.

Figure 2: Web Mode Participation Rate: 2003-2019 SDR



*Other response modes are self-administered mail-in form or telephone interview.

As in 2019, the 2021 online survey will be a mobile aware survey that renders in a user-friendly format on mobile devices (e.g., smartphones and tablets) so that the respondent experience with the online survey will be similar regardless of the screen size or web browser used to access the survey. Over 90% of the SDR respondents are expected to participate via web based on the online participation in the last survey cycle (93% in 2019). Of web respondents, 11% participated via a mobile device in 2019.

4.4 Adaptive Design Goals and Monitoring Metrics

The 2021 data collection will include an adaptive design strategy to help achieve a balanced sample to minimize nonresponse bias and achieve targeted numbers of completes for key analytic domains. This is the 4th cycle of SDR to apply an adaptive design approach. The 2021 emphasis will continue building on the procedures implemented during the 2019 cycle, including the development of improved monitoring metrics that assess the adaptive design's effects of treatments and interventions used in prioritizing nonrespondents for data collection.

As shown in Figure 1 in Part A9, the 2021 SDR data collection will be implemented in five phases over the course of 24 weeks. The first two phases will apply the same overall protocol to all sampled cases, pushing all sample members to a web response before using adaptive design techniques for nonresponse follow-ups. Messages within contacts will include a data visualization tailored to sample member nonresponse type such as domestic or international. See Appendix E for contacting materials.

Use of adaptive design strategies will begin in the "additional modes" or third phase of the data collection protocol at week 8 as shown in Part A9 Figure 1. Adaptive design prioritization models will run again at the start of each subsequent data collection phase to update priority scores.

The locating effort will also use adaptive design modeling results to prioritize cases for main field locating effort. Locating propensity models will be run at the start of prefield locating, at the end of prefield locating prior to the start of data collection, and then again at the end of week 5 after processing postmaster returns that indicate a problem with the mailing address from either of the first two mailings. Throughout all phases, flow processing metrics such as R-indicators and preliminary weighted key domain estimates will be monitored and compared to those from outcomes in the prior survey cycle. These monitoring metrics will assess the impact of the locating, contacting strategies, and use of incentives throughout data collection efforts.

Two prioritization metrics will be used for the adaptive design approach. The first is a "contact unit quality score" that is intended to guide which "contact unit" (email address, phone number, mailing address) to use in contacting sample cases as well as deciding which cases need priority in tracing/locating efforts. This quality score is determined by an algorithm that uses information on past success rates of the source or vendor(s) that provided the contact unit, whether the contact unit came from the respondent in a prior cycle or a designated contact of the respondent, and whether the contact unit is associated with one or more information items used to verify the case's identity (e.g., degree granting institution, year of graduation, etc.). This score will be constructed based on qualitative information rather than on statistical models with a focus on the quality of the current locating vendor data and sample member paradata.

The second metric combines two factors: the importance of the case in achieving sample size and precision goals in a set of target analytic domains, and a model-predicted propensity to respond for the case. The importance factor was used previously in the 2019 cycle and is based on a count of key analytic domains in which the case belongs and sample size falls short of the domain precision goals. The response propensity model is based on data from the 2019 cycle and is fitted separately for continuing cohort cases with prior responses, continuing cohort cases without prior responses, and new cohort cases. Model selection methods are used to determine the relevant variables for each of subset of cases. These two factors are then combined into an overall metric that is used to prioritize cases for tracing/locating and for data collection efforts.

4.5 Developing a Non-Production Bridge Panel

The nature of research doctoral training, the labor market, and workforce-related activities is changing. In response to this changing environment, and further motivated by a recommendation from the National Academies of Science, Engineering, and Medicine's Committee on National Statistics,[2] NCSES would like to explore modifications to the SDR survey content. Standard cognitive testing of question wording modifications provides a qualitative measure of quality and usability but does not assess or quantify the potential impact on survey estimates. Given the importance of maintaining the SDR's trend data, NCSES plans to include a small, representative, non-production sample (referred to as a bridge panel) to quantify the potential impact of question wording modifications on key survey estimates.

The bridge panel would allow NCSES to compare current SDR survey estimates (using responses from the SDR production sample) with estimates resulting from the modified questions (using responses from the bridge panel). Thus, the bridge panel would serve as a bridge to our current questions and could aid in the transition of our survey to possible question wording modifications. In future cycles, the bridge panel would provide NCSES the opportunity to assess and quantify the impact to survey estimates of potential methodological changes.

Sample Design and Selection

The 2021 SDR non-production bridge panel will include 5,000 sample cases selected from the 2019 Doctorate Records File (DRF). As noted earlier, the 2021 SDR target population includes all U.S. residents under age 76 with a science, engineering, or health research doctorate degree from a U.S. academic institution earned prior to 1 July 2019. Like the 2021 SDR production sample, the bridge panel will be representative of the SDR target population.

As part of the 2021 SDR sample selection effort, the 5,000-case non-production bridge panel will be sampled separately from the 125,938 -cases in the production sample. The sample selection for the bridge panel will use stratification variables similar to those used for the new sample cases, as discussed in Section B 1.2, but with further aggregated categories for the 77 detailed fields of study. The stratification variables will include sex, minority status, and 26 aggregated minor fields of study reported in the Doctorate records File (DRF). The multiway cross-classification of these stratification variables produces 104 non-empty sampling cells. The 5,000 bridge panel sample cases will be allocated across the 104 sampling cells in a manner that aligns with how the 10,000 new sample cases were allocated across similar sampling cells. After determining the sample allocation per sampling cell, cases will be selected

[2] At NCSES's request, CNSTAT convened an expert panel to review, assess, and provide guidance on NCSES's efforts to measure the S&E workforce population in the United States. Recommendation 5.2 of the panel's consensus study report noted that NCSES "*should continue to monitor, and formally evaluate as needed, the content of its survey questionnaires to ensure that the concepts and terminology are up to date and familiar to respondents. Changes should be implemented with careful consideration of their impact on trend data.*"

using systematic probability proportional to size sampling. The use of aggregated versions of the new sample stratification variables will enable comparison of key estimates between the bridge panel and the production sample.

Weighting Procedures, Replicate Weights, and Standard Errors
Estimates from the 2021 SDR bridge panel will be based on standard weighting procedures. As was the case with sample selection, the weighting adjustments will be done separately for the bridge panel and production sample cases. The goal of the separate weighting processes is to produce final weights for the bridge panel. To produce the final weights, the bridge panel cases will follow the weighting methodology outlined in Section B 2.1. In addition, sets of replicate weights, using the successive difference method, will be constructed to allow for separate variance estimation for the bridge panel.

Questionnaire and Survey Content
The 2021 SDR bridge panel questionnaire will include content similar to the 2021 questionnaire included in Appendix E with two modifications:

1) For the questionnaire items that were modified for 2021 to include coronavirus pandemic response options (i.e., employment status, part-time employment, job benefits, earnings, and conference attendance), the question wording from 2019 without the coronavirus pandemic response options will be used.
2) The questionnaire item measuring gender will be modified to offer response options beyond the binary responses of male and female.

Respondent Locating Techniques and Data Collection Methodology
As described in Section B 3.2, NCSES will use a combination of locating and contact methods based on past SDR surveys to maximize the survey response rate among the bridge panel cases. In terms of data collection methodology, the bridge panel will use a single-mode, web-based data collection protocol. The bridge panel cases will follow the 'web first pathway' outlined in Figure 1 in Part A9 with two modifications given the single web mode: (1) Bridge panel cases will not receive a paper questionnaire as part of the contacts at weeks 8 and 20; (2) No outgoing telephone calls will be made to the bridge panel cases during weeks 12-22. The bridge panel will receive survey mailing materials similar to those planned for the production sample cases (see Appendix E).

Similar to the incentive approach planned for the production sample, we plan to offer a $30 prepaid debit card to a subset of highly influential bridge panel cases at week 1 of the data collection effort. "Highly influential" refers to cases that represent small but important subpopulations in key analytic domains and a low response/locating propensity. We expect to offer $30 debit card incentives to approximately 1,000 of the 5,000 SDR 2021 bridge panel cases. These debit cards will have a six-month usage period at which time the cards will expire and the unused funds will be returned to NCSES.

Evaluation of Bridge Panel Question Modifications
As noted earlier, the bridge panel is designed to provide NCSES with an opportunity to compare current SDR survey estimates (using responses from the SDR production sample) with estimates resulting from the modified questions (using responses from the bridge panel sample). This comparison is designed to aid in the SDR transition to possible question wording modifications.

To determine if the modified questionnaire items should be included in subsequent SDR survey cycles, NCSES will conduct an evaluation of the bridge panel question modifications at the completion of the 2021 SDR data collection effort. The evaluation will include three components:

1) Comparison of survey estimates between the production sample and bridge panel sample for the modified questionnaire items. This comparison will be at both the total U.S.-trained SEH doctorate population level and the level of the bridge panel stratification variables.
2) Comparison of overall nonresponse rates and item-level nonresponse rates between the production sample and bridge panel sample.
3) Comparison of web instrument paradata (e.g., breakoff rates, changed answer rates, etc.) for both the production sample and bridge panel sample to assess the user experience associated with the modified questionnaire items.

This evaluation approach is designed to provide insight on both the estimation impact of these question modifications as well as any nonsampling error issues that would be introduced through the inclusion of these questions.

5. CONTACTS FOR STATISTICAL ASPECTS OF DATA COLLECTION

The NCSES contacts for statistical aspects of the SDR data collection are John Finamore, NCSES Acting Chief Statistician (703-292-2258), Daniel Foley, SDR Project Officer (703-292-7811) and Wan-Ying Chang, NCSES Mathematical Statistician and the lead SDR sampling statistician (703-292-2310).