SECTION B: Collection of Information Employing Statistical Methods

Survey Data Collection Procedures Background

The SED questionnaire is distributed to new doctorate recipients by the graduate deans of the approximately 432 doctorate-granting institutions, and approximately 570 independent programs within those institutions, in the United States. The SED questionnaires (either web or paper) are filled out at the time the individuals complete all requirements for their doctoral degrees. If paper questionnaires are completed, they are returned to NSF's survey contractor by the graduate dean's office. Because doctorates complete the requirements for graduation throughout the year, the questionnaire distribution and completion process is continuous.

The institution (usually the graduate dean's office) is the main SED interface with the doctorate recipient and experience shows that the interface is highly effective. The distribution of the questionnaire by the university itself, the clear nature of the questionnaire, and the cooperation of the graduate deans all combine to keep survey response rates above 90 percent.

When the completed paper survey questionnaires are received by the survey contractor, they are edited for completeness and consistency and then entered directly into the survey contractor's computer-assisted data entry (CADE) program. Surveys received via the web survey mode do not need to be data entered and are edited mainly through a series of pre-programmed skip patterns and range checks, which allow obvious errors to be corrected immediately.

The survey contractor works with ICs to obtain contact information for students who have not submitted their SED questionnaires. An Address Roster is sent to ICs asking for the addresses of the non-respondents. The survey contractor also utilizes web-based locating sites to identify contacting information for non-respondents. A series of letters or emails is sent to any graduate who did not complete the survey through their graduate school, requesting their participation and containing a PIN/password for web access (see Attachment 9 for a sample letter).

Finally, any graduate who does not complete the SED through their graduate school and does not return a survey through the non-respondent mailing effort is given the opportunity to complete a slightly shortened version of the survey over the telephone. If, by survey close-out, an individual has not responded, public information from the commencement programs or other publicly accessible sources is used to construct a skeletal record on that individual. The institution may also be asked to help provide data to complete skeletal records for these non-respondents. The skeletal record contains the name, PhD institution, PhD field, degree type, calendar year that the doctorate was earned, month that the doctorate was earned, and (usually) the sex of the doctorate earner. If a survey questionnaire is later received from a previous non-respondent, the skeletal record is replaced by the information provided by the respondent.

B.1. Universe and Sampling Procedures

The SED is a census of all students receiving a research doctorate between July 1 and June 30 of the following year. Because it is a census, no sampling is involved. All institutions identified in

IPEDS as granting doctoral degrees are asked to participate *if*: (1) they confer "research doctorates" and (2) they are accredited by one of the regional accreditation organizations recognized by the Department of Education. If so, the schools are asked to distribute the link to the online questionnaire, or to distribute paper questionnaires, to their research doctoral recipients at the time of graduation. The SED maintains the universe of research doctorate-granting institutions each year by comparing the list of institutions from IPEDS against the schools participating in the SED.   If a new institution is found to be offering a research doctorate, the institution is contacted and added to the SED universe.

A high rate of response is essential for the SED to fulfill its role as a key part of the universe frame for longitudinal sample surveys, such as the Survey of Doctorate Recipients, and as the only reliable source of information on very small groups (racial/ethnic minorities, women, and persons with disabilities) in specialized fields of study at the PhD level.

The feasibility of conducting the SED on a sample basis, and the utility of the resulting data, have been considered and found to be unacceptable. One reason many institutions participate in the survey is to receive complete information about all of their doctorate recipients in order to make comparisons with peer institutions. In addition, it is highly unlikely that the 570 graduate offices that voluntarily distribute the SED questionnaire could effectively carry out a sampling scheme. Schools often refer their students to an online graduation checklist, where the SED is but one step in the graduation process. In addition, conducting the SED on a sample basis would produce poor estimates of small groups (in particular, racial/ethnic minorities) earning degrees in particular fields of study, and such data are important to a wide range of SED data users.

A second sampling option – a mailing to doctorate recipients *after* graduation – would likely result in a much lower response rate because of difficulties in obtaining accurate addresses of doctorate recipients, particularly the foreign citizens who represent an ever growing proportion of the doctorates recipient universe each year. Such a technique would impose on the universities the additional burden of providing current addresses of new graduates, a somewhat ineffective process because the addresses of new doctorates are outdated almost immediately after graduation.

A third alternative, sending the questionnaire to doctorate recipients at a selected subset of institutions, would result in only a marginal decrease in respondent burden because the largest universities, all of which would need to be included in such a scheme, grant a disproportionate number of doctoral degrees. For example, the 50 largest institutions annually grant slightly over 50 percent of all doctoral degrees. Application of these sampling techniques would reduce both the utility of the data and the overall accuracy of the collected data. Matrix or item sampling – a widely used technique in achievement testing – would not be feasible because the characteristic information is needed for each doctorate recipient for use in selecting the sample for the follow-up SDR. It would reduce the utility of the information to request, for example, sex, race, or field of degree information for some doctorate recipients and not for others. These characteristics are not evenly distributed across the doctorate population, and the extensive uses made of the data base rely on the completeness and accuracy of the information on doctorate recipients.

Therefore, sampling doctorates would decrease the utility of the data while increasing burden on the graduate schools which administer the survey and decrease the incentives for the institutions to participate.

B.2. Survey Methodology

Because there is no sampling involved in the SED, there has traditionally been no weighting necessary. Basic information about non-responding individuals is obtained, where possible, from public records at their graduating institutions, graduation lists, etc. Both unit and item nonresponse are handled by including categories of "unknown" for all variables in tabulated results. The statistical and methodological experts associated with this survey are Stephen Schacht, Senior Research Scientist at NORC (773-256-6016) and Michael Yang, Senior Statistician at NORC (301-634-9492). At NSF, Lynn Milan, Project Officer for this survey (703-292-2275) and Jeri Mulrow, acting NCSES Chief Statistician (703-292-4784), provide statistical oversight.

B.3. Methods to Maximize Response

The SED has enjoyed a high response rate during its existence, with an average of 92% completions over the past 30 years. It owes this high rate, in part, to the use of the data by the graduate deans, who go to extraordinary lengths to encourage participation on the part of their graduates. Each graduate dean receives a profile of their graduates, compared with other institutions in their Carnegie class, soon after the data are released each year. It is also due to extensive university outreach efforts on the part of the survey contractor, NORC at the University of Chicago, and National Science Foundation staff, and to the importance the universities themselves place on the data.

Throughout the data collection period, schools are constantly monitored for completion rates. Data on doctorates awarded on each commencement date are compared to data from the previous round in order to flag fluctuations in expected returns. Schools with late returns or reduced completion rates are individually contacted. Site visits, primarily to institutions with low response rates, by NSF staff and survey contractor staff are also critical to maintaining a high response rate to this survey. NORC's electronic monitoring systems are particularly important to these efforts, as each institution's graduation dates or SED submission dates can vary from monthly to annually.

In addition to the broad efforts to maintain high completion rates, targeted efforts to prompt for missing surveys and critical items are also key. The survey contractor works with ICs and also utilizes web-based locating sites to contact students by mail and email for missing surveys. A series of letters is sent to any graduate who did not complete the survey through their graduate school, requesting their participation and including a PIN/password for web access as well as a paper questionnaire. Additionally, any non-respondent who does not complete the SED through their graduate school and does not return a survey through the non-respondent follow-up effort is given the opportunity to complete a slightly shortened version of the survey over the phone.

Finally, a Missing Information Roster (MIR) is sent to ICs who can sometimes provide critical item information (sex, race/ethnicity, citizenship etc.) in addition to addresses. Data received via the different modes are merged and checked to avoid duplicate requests going out to the various sources. The results of these varied efforts significantly increase the number of completions as well as reduce the number of missing critical items, thereby improving the quality of the SED data.

The response rates of institutions as well as the response rates to questionnaire items are evaluated annually. For example, the evaluation of the response rate for 2013 indicated that over half of the non-response was due to 20 institutions. Institutions with poor response rates were targeted for special letters or site visits by NSF or survey contractor staff and, to a large extent, these efforts have been successful in raising the response rates at institutions.

B.4. Testing of Procedures

The SED has undergone extensive review and testing of the questionnaire and the methods employed in conducting the survey in recent years. The changes made to the SED 2016 survey version are a result of many activities which have helped inform changes to instruments and procedures over time. The following major activities have been conducted since the previous OMB clearance submission (see Attachment 10.1 for a list of the methodological studies conducted over the past 15 years). The NSF project officer will be pleased to provide any of the documents referred to in this section or those referred to throughout the supporting statement.

Data Collection Related Tests

The accuracy of the data from the SED has been one of its strongest assets. An ongoing evaluation of the accuracy of coding, editing, and data entry processes is conducted. It consistently indicates that the error rate is very low (less than one percent). During data collection, the frequency distribution of variables is monitored on a continuous basis, so that emerging problems, such as high item non-response rates, can be identified early in the data collection phase and appropriate corrective measures implemented, if necessary. Additional quality control checks on the merger of paper and electronic questionnaires as well as the merger of missing information into the master database are also ongoing. The survey questionnaires are constantly compared with the universities' graduation lists and commencement programs to make sure that only those persons with earned research doctorates are included.

Additional research that has been conducted in the last two years related to data collection operations and strategies are summarized below. (See Attachment 10.1 for additional details.)

- **Institution Eligibility Criteria:** This study was undertaken in 2013 to examine the eligibility criteria for institution inclusion in the SED against a broader national and international context as well as the adjudication process for determining the eligibility of institutions and programs not currently in the SED but appearing to meet the criteria for inclusion. The study's final report is under consideration by NSF. No changes have yet been implemented to the SED eligibility review process.

- **Confidentiality Issues:** This study included cognitive interviews and focus groups conducted in 2013 and 2014 with doctorate recipients, graduate deans, institution contacts, and institutional researchers concerning the confidentiality procedures employed by the SED. No changes have been implemented based on the findings.
- **Web Survey Breakoff Conversion:** Two studies were conducted in 2013 and 2014 discussing strategies to increase survey completion of sample members who had begun but not completed the SED web survey, as well as the impact on data quality. Findings demonstrated that prompting did not have adverse effects on survey data quality, as measured by item nonresponse. However, the results of the two studies indicated different outcomes regarding the success of standard email prompts over mail prompts in converting breakoffs. These results informed the 2015 nonrespondent contacting experiment, which is examining if sending email prompts before mail prompts results in a higher survey completion rate (see further details in the next section, "2015 Experiments").
- **Mode Effects on Item Response Rates:** A 2013 study reviewed SED item response rates by mode, controlling for time of completion, and found that web item response rates tend to be higher than hardcopy item response rates. The study concluded that prompts may play an important role in increasing item response rates on the web. Thus, the web prompts have remained in the web questionnaire. The redesign of the web questionnaire being done in 2015 will further improve the prompts both for clarity and user experience. The effectiveness of these redesigned prompts will be included in the cognitive interview activities.

2015 Experiments

Strategies to prompt survey completion for non-respondents are continually examined with a view to maximizing response rates and reducing data collection costs. During the 2015 data collection cycle, two experiments are being employed to test improving response rates of non-respondents. First, the inclusion of a progress bar in the web survey will be tested. Non-respondents who are contacted through follow-up efforts will be selected for the control or treatment group upon logging into the SED survey. Treatment group members will see a progress bar that displays their advancement through the survey both by visual increase of the bar and by percentage. Control group members will not see a progress bar, which follows the current web survey design. The experiment is designed to test if the inclusion of a progress bar in the web survey reduces the number of breakoffs and, ultimately, leads to more completed web surveys.

The second experiment will test contacting strategies for nonrespondents for whom both a mailing address and email address are available. The current follow-up protocol for SED nonrespondent is to send all nonrespondents with a mailing address up to five mail prompts (four letters and one postcard) before any other treatment, regardless of the presence of an email address in the sample database. Nonrespondents are then sent up to two email prompts if an email address is available. The experiment will test the effectiveness of making email prompts primary over mail prompts for SED nonrespondents for whom the doctorate-granting institutions have provided both a mailing and an email address. Under the experimental design, non-

respondents selected for the treatment group will first be sent the pair of prompting emails and, if necessary, will then start the series of five mail prompts.

Nonrespondents eligible for this experiment are unique in that both email and mailing addresses are provided by the degree-granting institutions and this information is used in the first prompt contacts. Due to the transitional nature of this population at the time of their graduation, the use of institution-supplied mailing information as quickly as possible is critical in order to reach nonrespondents before they have moved. Email addresses may be more effective in reaching nonrespondents who have already relocated.

The experiment results will examine if there would be a cost benefit to prioritizing email follow-up before mail for SED nonrespondents. This experiment also has the potential benefit of reducing respondent burden. Sending a paper invitation for a web questionnaire requires that the letter include a URL address for the survey, which recipients must manually enter into a computer (as opposed to clicking on a link in an email), before typing in their access code (also provided in the letter) in order to start the web questionnaire.

When it is possible to obtain nonrespondent email addresses early in the prompting cycle for a web-savvy target population with a general preference for the web mode (such as recent doctorate recipients), starting respondents with the web mode (i.e., email prompts) can potentially increase response rates while decreasing respondent burden and data collection costs. Considering that the web is the dominant questionnaire completion mode in the SED (90% of FY 2014 respondents completed the web-based questionnaire), that the SED potentially possesses an adequate email address list for its nonrespondents, and that the survey targets web-savvy and highly educated individuals, the SED could save resources and increase the data collection pace if nonrespondents could be prompted via email instead of conventional mail prompting methods.

Survey Quality Tests and Research

Several tasks were completed since the last OMB package, including several that informed the recommendations for the next cycle. These tasks ranged from continuous assessments of everyday processes to overarching reviews of the institutions and degrees included in the survey to confirm the completeness and accuracy of the SED universe.

The following tasks are conducted regularly throughout each survey round:
- Review of systems, programming, and quality control data preparation processes with a goal of shortening data collection and an earlier delivery of the final data file. Based on the system review, additional reports were developed to assist in tracking institutions late in returning their materials. Aggressive interventions with these institutions aided in shortening the data collection period for the 2013 and 2014 survey rounds.
- Merging data on a flow basis to identify and correct data inconsistencies and to reduce the amount of time between the close of data collection and the release of the data. Based on this review for the 2014 round, the majority of the file preparation case-level data consistency and doctorate eligibility reviews were conducted before data collection closed, giving file preparation staff more time to focus on other quality assurance tasks that can only occur after data collection closes.

These tasks are completed annually, prior to the beginning of data collection or the start of data preparation:

- Comparison of the IPEDS database of doctorate-granting institutions to the SED universe to identify institutions newly offering doctorate programs that are not currently in the SED. Based on this review six new institutions were deemed eligible for participation in the SED for the 2015 round and eight for the 2016 round.
- Review of the IPEDS database and the Interim Results Form to determine if any institutions currently participating in the SED are offering eligible degrees that are not currently being included. Based on this review, six programs at institutions already in the SED were deemed eligible for participation in the SED for the 2015 round and five programs for the 2016 round.
- Discussion of possible improvements in the coding and editing processes to ensure faster data entry resulting in more timely follow-up with non-respondents. In 2015, enhanced auto-coding rules were implemented in order to capture more cases and reduce the records that require manual coding. In addition, a new coding interface was designed for more efficient manual coding.
- Consultation with data processing managers on issues of paper and electronic data handling and mergers. In the 2014 round, based on this consultation, a change was made in how prior round hardcopy questionnaires' data were incorporated into the merged data set for the current round. Rather than having to link the previous round's hardcopy data file as an additional source to the current round's merged dataset, the original hardcopy surveys were data entered into the current round's instrument to ensure data editing consistency and reduce the complexity of the data merge process. Special rules were created to instruct these editing and data entry processes to ensure data quality.
- In-depth analysis of confidentiality issues, particularly of data products that will be publicly available. For the 2013 and 2014 rounds, staff worked closely with NCSES project officers to modify the structure of the race and ethnicity construct to meet new requirements to more closely match U.S. Census categories while minimizing the impact of additional data suppression across different reports.
- Coordination of items common to the SDR and SESTAT instruments (see section A.4). Included in the 2016 questionnaire revisions is the ability for respondents to indicate currency type when reporting salary amount in the SED web questionnaire. This revision was made in order to match the SDR item format and collect more accurate salary information from respondents.

The following tasks are completed annually at the end of each data collection period. The results are compiled and reviewed before each new OMB clearance cycle to inform possible changes:

- Extensive reviews of unit and item-by-item frequencies and item analysis for floor and ceiling effects (i.e., whether quantitative response options go low enough and high enough for the range of SED responses). The 2013 and 2014 data frequencies were reviewed to determine if there was a need for the expansion of salary categories in questionnaire item B8. (The current categories go from "$30,000 or less" to "$110,001 or above" in $5,000 increments at the lower end of the range and $10,000 increments at the higher end.) It was determined that a new category would be too small to provide any

utility and would likely be aggregated for reporting purposes; thus, no change was made at this time.

- Review of all respondent comments for concerns over confidentiality or item improvements. For the 2016 instrument, there was consideration to remove the five leading "X" placeholders when asking for Social Security number in item C15 to allay respondent concerns about being asked to provide SSN. Based on expert methodologist review, it was determined that this revision should be tested in the cognitive interview activities being conducted in CY 2015. In questionnaire item C15, the term "last four" will be italicized in order to reinforce that the request is only for partial SSN.
- Also, the addition of "Associate's Degree" as a response category when asking about highest level of parental education (item C4) was initiated by feedback from respondents who felt the existing categories of "some college" and "Bachelor's degree" do not properly represent those who have earned an Associate's degree.
- Review of "other, please specify" information in consideration of expanding or changing answer options. No revisions to the 2016 instrument resulted from this review.
- Coordination of data post-processing rules for items common to the SDR and SESTAT instruments, including the race, ethnicity and disability (i.e., "specific functional limitation") items (see section A.4). After the revisions of the functional limitations questions in 2012, SED staff coordinated with SDR to understand how they edited hardcopy questionnaires and handled the responses in post-processing steps to ensure that data was interpreted according to the same rules, as appropriate.

In addition, the following tasks were conducted during the last OMB clearance cycle, and will be conducted periodically in the future:

- Detailed review of emerging and declining fields of study and alignment with the CIP (Classification of Instructional Programs). The result of the review completed in preparation for the 2016 SED is the addition of eight new fields and the removal of one field from the SED field of study taxonomy.
- Review of the non-PhD doctorate degrees included in the SED to confirm that they are research degrees and thus eligible for the survey. Based on this review in 2014, it was determined that a newly offered Doctorate of Design is eligible for the SED.
- Literature reviews on targeted topics, such as disclosure avoidance and other confidentiality issues, as well as an initial review of the accreditation requirements for academic institutions. In 2013, a review of the institution eligibility criteria for participation in the SED was conducted, including literature reviews and interviews with select institutions and accreditation agencies. Recommendations from the accreditation review included adding two additional agencies to the list of qualifying accrediting agencies when considering approving new institutions for the SED. This recommendation is still under consideration at NSF.

Finally, the following specialized studies were conducted during the last OMB cycle, the findings of which will be used to inform future SED processes.

- **Timeline Data Quality Improvements for the Survey of Earned Doctorates:** An analysis of the current approach the SED employs to collect, edit and report timeline data, resulting in recommendations for improved data quality through potential questionnaire, editing, and data presentation changes. Findings were used to inform a number of process

revisions, including: expansion of the auto-coding process for timeline variables; modification of rules used to flag nontraditional timeline sequences; addition of timeline variables to the DRF for use in further research; and revision of select imputation rules. In addition, revisions to timeline questions in the instrument are included in the future cognitive interview activities.

- **Enhancements in Auto-Coding in the Survey of Earned Doctorates**: A study to assess the feasibility of employing an automated coding application to additional SED variables that are currently manually coded, resulting in the definition of additional coding rules needed to apply these changes while improving data quality and reducing labor costs. Findings from this study are being implemented in the SED 2015 coding activities.
- **Department Coding Feasibility Study:** An examination of the feasibility and cost of coding respondents' verbatim responses to their department (item A3), which has up until now not been cleaned, coded or stored in the DRF. Findings indicated that coding this item would be feasible for respondents whose field of study was in science and engineering. The findings were considered by NSF, but no changes have yet been implemented.
- **Disclosure Analyses of Tabular Data:** An analysis of cell suppression processes using log-linear modeling as a method of checking the underlying trends over different subgroups, which could have broad utility for SED data products and beyond. The results indicated this approach does not increase disclosure risk and also preserves unsuppressed cells and marginal counts. The findings were considered by NSF, but no changes have yet been implemented.

Proposed Tests and Research

Over the course of the proposed OMB cycle (May 2015 – December 2017), NSF anticipates conducting several methodological research tasks and analyses of data user needs, some involving cognitive interviews. The tasks associated with these research studies and user analyses will be conducted under the Generic Clearance of Survey Improvement Projects (OMB #3145-0174), as needed.

The first effort will involve the redesign of the SED web survey to create a visual design that is more appealing and reduces potential confusion, measurement error, and break-offs. These changes are intended to be applied to the 2016 SED web survey. The new design will be mobile compatible and 508 compliant and will also include an overall redesign incorporating best practices. This will include:

1) movement of button locations for optimal usage;
2) reorganization of question matrices to improve visual grouping;
3) compatibility with various mobile platforms (e.g., smart phone and tablet);
4) improved spacing of question stems and response categories;
5) implementation of an advanced search function to reduce respondent burden and improve data quality; and
6) overall modernization of web survey design which is more aligned with the experiences of the doctoral graduate population that is completing the survey.

The second is a larger research effort to investigate a redesign of SED survey questions, including question order, language, and response options. Drawing on previous work and the survey literature, the objectives of this work are to:

- redesign the entire SED instrument to improve its measurement properties,
- test and fine-tune the revised instrument through a series of cognitive interviews, and
- evaluate the potential benefits and possible drawbacks of the various elements of the redesigned instrument.

This work not only has the potential to enhance the overall quality of the SED data, but also to reduce the burden placed on respondents as they complete the survey and the costs of the survey by cutting down the time spent resolving data discrepancies. The item revisions will be tested through a series of four rounds of cognitive interviews, with up to 25 respondents per round, which will allow evaluating the effectiveness of a set of proposed instrument changes, revising the instrument as needed, and testing the revisions. The first two rounds of interviewing will be devoted to testing content changes to the questionnaire, such as improvements to question order, wording, response categories, and instructions to the respondent. The third and fourth rounds will include the new web survey redesign discussed above in order to test the revised SED content in the new web survey environment. Cognitive interview respondents will include graduate students who are about to complete their PhD or who have recently completed it. The findings are intended to inform the 2017 SED, which will include the submission of an OMB addendum, as needed, to gain approval of questionnaire revisions.

The draft SED 2016 questionnaire was first reviewed in December 2014, and the final questionnaire changes were reviewed and approved by the sponsors in January 2015. (See Attachment 5 for the list of persons who were consulted or who reviewed the questionnaire.) See Attachment 2 for a list detailing the changes made to the SED 2016 questionnaire from the 2015 version and the rationales for those changes.

B.5. Individuals Consulted

NORC at the University of Chicago is the organization contracted to collect and analyze the SED data for the 2016-2017 survey rounds. Staff from NORC who have consulted on the aspects of the design are listed in Attachment 5.

Additional individuals both inside and outside of NSF who have consulted on the statistical and methodological aspects of the design are also listed in Attachment 5.