

## MEMORANDUM

May 18, 2012

To: Shelly Martinez, Desk Officer  
Office of Management and Budget

From: John R. Gawalt, Acting Director  
National Center for Science and Engineering Statistics

Via: Suzanne Plimpton, Clearance Officer  
National Science Foundation

Subject: Notification of information collection under generic clearance

The purpose of this memorandum is to inform you that the National Center for Science and Engineering Statistics (NCSES) of the National Science Foundation (NSF) plans to conduct a methodological study for the Early Career Doctorates (ECD) Project under the generic clearance for survey improvement projects (OMB number 3145-0174).

### Background

NCSES has determined based on results from the multi-year Postdoc Data Project (PDP) that it is necessary and feasible to gather in-depth information about postdoctoral researchers (postdocs) working in all employment sectors in the United States. Early activities for the PDP focused on determining best methods for filling data gaps related to coverage of and data collected on and from postdocs. To address the sampling and data collection challenges that surfaced related to defining a postdoc in a measureable way, NCSES expanded the PDP target population beyond postdocs to include early career doctorates (ECDs) - defined as individuals receiving their doctorate in the last ten years. Expanding the target population addresses two challenges NCSES has experienced: development of a sampling frame given the variation in postdoc definitions across and within organizations, and development of a robust dataset which allows analysts to compare characteristics of postdocs and non-postdocs.

In consultation with several stakeholders, NCSES developed a questionnaire and pretested it in the fall of 2009. Results from the pretest indicate that most of the questions work as intended, but modifications are needed to address the differences in employment practices across sectors and to reduce the number of early breakoffs. A continued challenge to collecting data from ECDs is how to reach them.

With an expanded target population and revised questionnaire, NCSES proposes this research to test a data collection strategy that uses institutional contacts as the conduit for questionnaire dissemination to reach ECDs in the following employment settings: U.S. academic institutions, Federally Funded Research and Development Centers (FFRDCs), and National Institutes of Health (NIH) Intramural Research Programs. These organizations were involved in prior research to build a sampling frame and indicated relative ease in developing a list of individuals in the ECD target population. This approach is also based on the success of the Survey of Earned Doctorates (SED), which has used this strategy for several decades. It will increase the likelihood of finding institutional contacts knowledgeable about ECDs and should minimize the difficulty encountered in prior ECD activities. Therefore, it is necessary for NCSES to conduct this

research prior to implementing this methodology. NCSES plans to use these results to design a full scale pilot study.

The primary goals of this research are (1) to test methodologies for building a sample frame of ECDs (stage 1 of data collection), (2) to contact a sample of ECDs to complete a Web survey (stage 2 of data collection), and (3) to assess strategies for recruiting ECDs. Experience with other NSF studies suggests that institutional involvement in the data collection effort may result in better response rates than a survey contractor would receive working independently.

A secondary goal is to test the ECD questionnaire (Attachment C) which was revised based on results from the 2009 questionnaire pretest and feedback from stakeholders.

NCSES will use the results of this methodological research to design a full scale pilot study.

### **Proposed Methodology**

Similar to the SED data collection methodology, this research will test strategies for implementing a two-stage data collection. The first stage is to build a sampling frame of ECDs using institutional contacts. The second stage is to sample and survey the ECDs identified through the first stage.

#### *1. Institutional Sampling and Frame Creation (Stage 1 of Data Collection)*

For the first stage, 81 institutions will be selected from: U.S. academic institutions from the Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS), FFRDCs, and NIH Intramural Research Programs. Table 1 shows the proposed sample allocation: 64 GSS institutions, 16 FFRDCs, and NIH Intramural Research Programs. Institutions that do not agree to participate will be replaced in the sample with institutions with similar characteristics. The allocation in Table 1 oversamples the FFRDC institutions to include more of them in the research given there are only 37 FFRDCs. Of the 64 GSS institutions, 16 institutions from four strata will be allocated to ensure the selection of smaller institutions. NIH Office of Training and Education will provide a list of ECDs, therefore subsampling will not occur. Each of the sampled institutions will be asked to provide a list of their ECDs and the following information: name and contact information (such as email address, mailing address, and telephone number), year of doctorate, job title at the institution, gender, date of birth, field of doctorate, and field of employment.

#### *2. ECD Sampling and Recruitment for ECD Survey (Stage 2 of Data Collection)*

For the second stage, a sample of ECDs will be selected as follows: 40 from each of the 64 GSS institutions, 40 from each of the 16 FFRDCs, and 80 from the NIH Intramural Research Programs. Table 1 also summarizes the ECD sampling design. The proposed design will yield 3,280 potential ECDs across 81 institutions to ensure that a diverse set of institutions and ECDs are sampled. The ECDs will be invited to participate in a 30-minute web-based questionnaire.

This proposed sampling strategy is similar to the design of the National Center for Education Statistics' National Study of Postsecondary Faculty. That study sampled approximately 40 faculty members per institution, and had an average of approximately 30 eligible faculty respondents per institution. NSF is limiting the number of institutions given the more qualitative than quantitative aspects of this research. However, with 40 potential respondents per institution, the number of ECDs will be large enough to provide acceptable power for comparisons. Assuming a within

institution correlation of 0.01, a difference in the ECD response rate of 5% can be detected with 80% power at alpha = 0.05 when the response rate is 55% in one of the experiment groups and 60% in the other experimental group.

Table 1. Sampling Strata and Sample Sizes

Institutional strata	Institutions on frame	Estimated Number of ECDs <sup>1</sup>	Number of institutions sampled	ECDs in sample	ECDs in each experimental group
GSS institutions	572	225,731	64	2,560	1,280
• <i>medical schools</i>	135	135,630	16	640	320
• <i>highly active research institutions</i>	113	48,318	16	640	320
• <i>doctorate-granting institutions</i>	130	19,892	16	640	320
• <i>non-doctorate granting institutions</i>	194	21,891	16	640	320
FFRDCs	37	2,449	16	640	320
• <i>reporting postdocs in the 2010 survey</i>	23	2,449	8	320	160
• <i>all other FFRDCs</i>	14	0	8	320	160
NIH Intramural Research Programs	25	3,650	1 <sup>2</sup>	80	40
Total	660	232,549	81	3,280	1,640

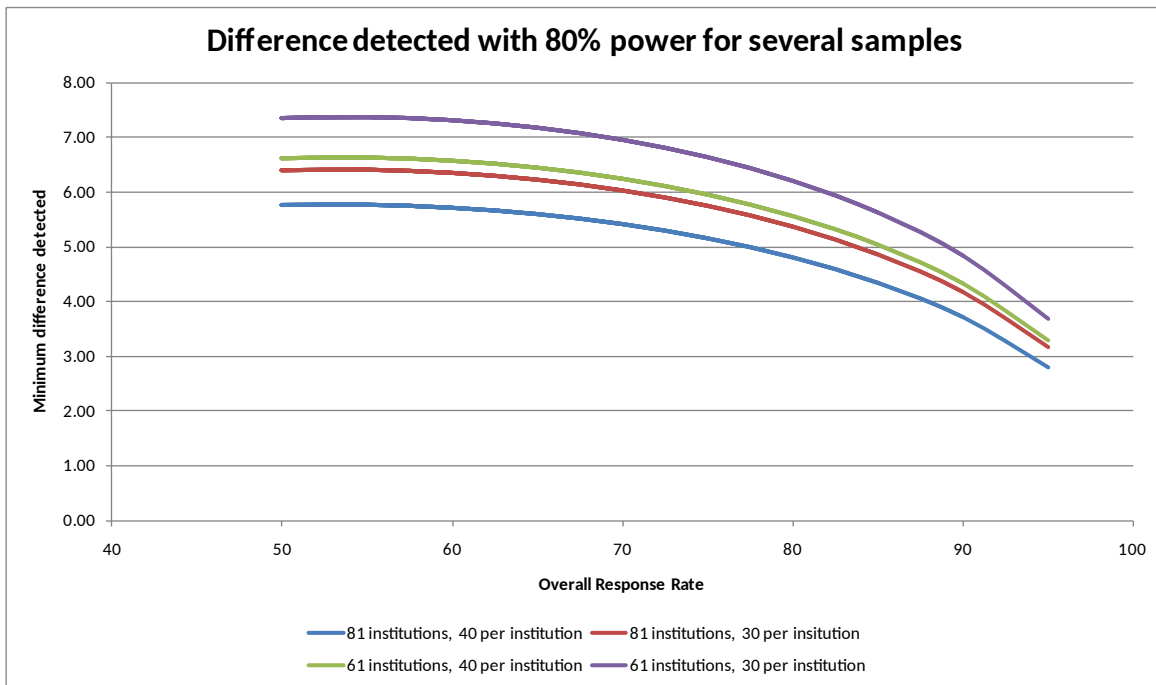
<sup>1</sup> The number of ECDs is not known. The estimated number is derived from the number of postdocs and NFRs reported in the GSS and FFRDC surveys and the number of instructional and research staff reported in IPEDS. NOTE: FFRDC and NIH IRP counts represent postdocs only and will be adjusted to better estimate all ECDs.

<sup>2</sup> The NIH Office of Intramural Training & Education will provide a single list for all 25 intramural research programs.

The minimum detectable effect is 5% at 80% power, an alpha level of 0.05, and an overall response rate between 75% and 80%. As seen in the table and graph below, we considered other configurations of institutions and ECDs per institution, but several factors moved us toward the current design, of which the 5% discernible effect is just one, albeit an important one. First, as the goal of the methodological pilot is to test our frame creation and contact strategies, it is imperative that we sample a wide variety of institutions. Second, for experimental validity, it is important that the response burden for each institution in the pilot reflect the likely burden in the future full-scale data collection. Third, for the experiment to be meaningful, we need enough ECDs to have a reasonable chance at detecting a difference in response rate within each experimental group. Sampling 80 institutions rather than 60 will ensure that we get sufficient variation across the institutional respondents, and sampling 40 ECDs per institution will ensure experimental validity. The overall sample size of 3,280 yields a minimum detectable difference

close to 5% starting at an overall response rate of 75% and provides the best opportunity to achieve a meaningful and reasonable result, namely that if the difference in response rates is less than 5% the ECD contact strategies will be deemed equivalent.

Possible sampling designs	Selected	Alternate 1	Alternate 2	Alternate 3
Number of GSS + FFRDC Institutions	80	80	60	60
Number of ECDs per GSS + FFRDC Institution	40	30	40	30
Number of ECDs from the NIH IRP	80	60	80	60
Total Institutions	81	81	61	61
Total ECDs	3,280	2,460	2,480	1,860
Detectable Difference in Response Rates	5.0%	5.5%	5.5%	6.0%



### Proposed Protocol

Institutional contact materials for stage one (Attachment A) of this study are similar to those used by the GSS and FFRDC surveys. Data collection will begin with NSF contacting a high authority (HA) at each institution, typically the president, to inform them of the project, to seek their participation, and ask for a point of contact (POC). For many institutions NSF will have identified a likely POC. HAs will be asked to confirm previously identified POCs or to suggest POCs when one has not been identified.

Once a POC is identified, NSF will work with them to develop the ECD frame and to identify the information available on the ECDs. These data fields will be used to recruit respondents (i.e., ECDs) in stage two. ECD contacting materials for stage 2 of the data collections are provided in Attachment B.

### Confidentiality

At the outset of the interview, respondents will be informed of their privacy and confidentiality rights, including the right to decline participation altogether and the right to refuse any individual question item in the interview. The paragraphs below convey the confidentiality information that will be provided to all respondents:

This information is solicited under the authority of the National Science Foundation Act of 1950 as amended, and the Confidential Information Protection and Statistical Efficiency Act of 2002. The information that you provide will be used for statistical purposes only. Your responses will be kept confidential. Your response is voluntary and failure to provide some or all of the requested information will not in any way adversely affect you.

The average time to complete this survey is about 30 minutes. Please send any comments on the time required to complete this survey to the National Science Foundation, 4201 Wilson Blvd, Suite 295, Arlington VA 22230, Attn: NSF Reports Clearance Officer.

**Burden Information**

The total estimated burden for this research project is 1,754 hours: 729 hours for stage one and 1,025 hours for stage two. The stage one burden estimate assumes an average of 6 hours per POC and 1 POC per institution. The stage two burden estimate assumes a 60% response rate and 30 minutes to access and complete the questionnaire (3,280 x 0.60 x 0.5 hour per respondent). In the prior research, the average time to complete the questionnaire was 26 minutes.

Table 2 provides details for the estimated burden.

**Table 2. Project Burden by Stage and Respondent Type**

Respondent Type	Hours Per Respondent	Number of Respondents	Estimated Total Hours
<i>Stage 1: Frame Creation</i>			
HA	1.0	81	81
POC	6.0	108	648
Subtotal			729
<i>Stage 2: ECD Survey</i>			
HA or designee	1.0	41	41
ECD	0.5	3,280	984*
Subtotal			1,025
Total		3,510	1,754

\* Assumes a 60% response rate

Note: HA = High Authority, POC = Point of Contact, ECD = Early Career Doctorate

Actual burden will be collected at the end of each stage for all participants in the research project.

**Incentive Payments**

None.

## **Contact Information**

The contact person for questions regarding this research is:

Kelly S. Phou  
National Center for Science and Engineering Statistics  
National Science Foundation  
4201 Wilson Boulevard, Suite 965  
Arlington, VA 22230  
703-292-4722  
kphou@nsf.gov

### Attachments

- Attachment A: Institutional Contact Materials
- Attachment B: ECD Contact Materials
- Attachment C: ECD Questionnaire Specifications

cc: Frederic J. Wendling  
Stephen H. Cohen  
Emilda B. Rivers  
Kelly S. Phou  
Darius Singpurwalla