Attachment H – ECDS Sample Design

ECDS Sample Design Plan

The Early Career Doctorates Survey (ECDS) plans to collect data from about 18,000 early career doctorates (ECD). The sample design will be a two-stage stratified sample of U.S. academic institutions, federally funded research and development centers (FFRDCs), and the National Institutes of Health (NIH) Intramural Research Programs (IRPs), and individuals working at these institutions. At the first stage, we will select approximately 350 institutions with selection probability proportional to their size (PPS), where the size measure is described later in this document. We expect 300 of these institutions will participate in the survey and provide lists of ECD working at their institution. These lists will then be used as the sampling frame for the second, individual stage, of the data collection. At the second stage, individual sample members will be selected within institutions such that their overall (unconditional) selection probabilities are equal across sample members within each of the following domains of analysis: employment sector (institution type), postdoc status, sex, citizenship, and race/ethnicity.

This self-weighted sample design is also known as an equal probability of selection method (*epsem*) sample. The sampling weight is calculated as the inverse of probability of selection. When the probability of selection is equal for all sampling units, their sampling weights are also equal (constant). When the sampling weights vary across units, this variability would increase variance of the estimate. Thus, a sample with equal selection probability will be more efficient than that with unequal selection probability.

The ECDS has several domains of interest, some of which are relatively small in the population. A completely proportional sample allocation would achieve equal probabilities of selection across all domains and strata, but to achieve adequate precision across all of the domains with a proportionate allocation would require an extremely large sample size. Therefore, to achieve adequate precision within and across domains while controlling the total sample size across domains, sampling rates will be allowed to vary across the domains and strata. The domains of analysis, specification of sample size, and selection of institutions (first-stage sampling) and ECD (second-stage sampling), treatment of missing frame variables, and sample release strategy are discussed in detail below.

In summary, the steps in sample selection are done as follows:

(a) Step 1: Determine domain minimum effective sample size based on pre-specified values of coefficient of variations (CVs) by domain of analysis to allocate sample of ECD by domain.

(b) Step 2: Determine the sample size of institutions (first-stage sampling) by sampling stratum based on expected respondents of ECD per institution.

(c) Step 3: Calculate composite measure of size for the first-stage sampling for each institution in the frame, determine certainty institutions, and draw sample of non-certainty institutions.

(d) Step 4: Collect list of potential ECD from each sampled institution. Impute missing values in the sampling variables as necessary, and evaluate imputation results.

(e) Step 5: Calculate the second-stage sampling rate for each institution by sampling stratum and domain, and draw samples of ECD.

A. Domain of Analysis and Sample Size

The first exercise is to allocate a sample size of approximately 16,750 ECD in the U.S. academic institutions,[1] 850 ECD in the FFRDCs, and 400 ECD in the NIH IRPs into each domain of analysis.[2] Note that these are respondent sample sizes and will need to be inflated by the anticipated response and eligibility rates. The allocation to the analytic domains is determined based on the level of detail needed in the ECDS tables and the information available on the frame, and the sample sizes are determined so as to produce estimates with specific precision defined by the desired CV within each domain. Because it is important that the analyses that are produced for these domains are supported by adequate sample sizes, the domain population size information (when available) can be used to allocate the sample across domains; this gives the flexibility to over- or under-sample certain domains.[3] Therefore, based on the specific precision requirement, a threshold or minimum effective sample size should be determined for each domain.

For the U.S. academic institutions, the domains of interest are cells defined by the Institution Type, Postdoc Status, Sex, and Citizenship-Race-Ethnicity variables as shown in Table 1. A priori counts of ECD by gender, citizenship, and race/ethnicity are not available for the FFRDCs and NIH IRP, and as a result the same strata as in the U.S. academic institutions (GSS Substrata) cannot be constructed. Instead, the composite size measures for these two strata will only include overall size and postdoc status.

Allocating the sample to the domains proportionally means that larger domains would get larger sample sizes and smaller domains would get smaller sample sizes; this allocation would provide smaller variances and would be efficient for estimation that cuts across domains. However, this sample size allocation might end up with some small domains with too small of a sample size for analysis of interest, and hence would not meet the pre-specified precision requirements. An alternative option is to allocate the samples equally across domains. This equal sample allocation across domains usually has an advantage of higher statistical power for tests for comparisons. However, this comes with a price that the variance used in the analysis might be larger due to variation in the weights resulting from oversampling or undersampling some domains. Therefore, the allocation for the sample size of ECDS started with an approximate proportional sample allocation, but iterated in order to satisfy a minimum sample size threshold for domain level based on the required precision of analysis. This results in an allocation that satisfies precision constrains for multiple domains but is no longer exactly proportional. For example, small domains that are of interest for analysis are sampled at higher rates compared to domains that are not as rare.

---

[1] The sampling frame for U.S. academic institutions can be developed from the National Science Foundation – National Institutes of Health Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) data.

[2] Ideally the sample should be allocated proportionally to the U.S. academic institution, FFDC and NIH-IRP. In doing so, however, the FFRDC and NIH-IRP will have small sample size compared to that sample allocated to the U.S. academic institutions. In such case, the comparison across institution type will not be optimal (may not detect a meaningful difference for a given power of the test, or for a specified minimum detectable difference the power of the test is low). This allocation is subject to change after discussion with the National Science Foundation (NSF).

[3] When information on domain size is not available and sample size allocation across domains may not be controlled, a random sample might produce proportional sample size across domains but not guarantee. When the sample is proportional, small domain will receive smaller sample size, which could lead to an estimation issue such as issue of reliability of the estimate.

Table 1. Domains of interest, expected coefficient of variation (CV) and associated minimum sample sizes needed for a total sample size of 18,000 ECDS from U.S. Academic Institutions (GSS Institutions), FFRDCs, and NIH IRP's

Domain level	Category	Minimum sample size[a]	Expected CV[b]
Combined GSS, FFRDC, and NIH IRP			
Postdoc Status	Non-Postdoc	6,336	0.03
	Postdoc	5,979	0.03
GSS Institutions			
GSS, total		11,445	0.01
Sex	Female	4,237	0.02
	Male	4,259	0.02
Postdoc Status	Non-Postdoc	4,815	0.03
	Postdoc	4,588	0.03
Citizenship-Race-Ethnicity	Non-U.S. citizen	1,527	0.04
	U.S. citizen–White	5,981	0.02
	U.S. citizen–Asian	1,329	0.04
	U.S. citizen–Minority	1,622	0.04
GSS Substrata	Medical schools and centers	2,487	0.03
	Very high research activity universities	3,052	0.03
	High research activity universities	2,853	0.03
	All other colleges and universities	3,708	0.03
GSS substrata × Postdoc Status	Medical schools and centers; Postdoc	1,557	0.04
	Very high research activity university; Postdoc.	2,508	0.04
	High research activity university; Postdoc	383	0.10
	All other colleges and universities; Postdoc	270	0.10
	Medical schools and centers; Non-Postdoc	2,214	0.04
	Very high research activity university; Non-Postdoc.	2,253	0.04
	High research activity university; Non-Postdoc	2,730	0.04
	All other colleges and universities; Non-Postdoc	3,267	0.03
Postdoc Status × Sex × Citizenship-Race-Ethnicity	Postdoc; Non-U.S. citizen; Female	921	0.04
	Postdoc; Non-U.S. citizen; Male	1,592	0.04
	Postdoc; U.S. citizen–White; Female	459	0.05
	Postdoc; U.S. citizen–White; Male	771	0.05
	Postdoc; U.S. citizen–Asian; Female	254	0.09
	Postdoc; U.S. citizen–Asian; Male	210	0.09
	Postdoc; U.S. citizen–Minority; Female	236	0.08
	Postdoc; U.S. citizen–Minority; Male	165	0.08
	Non-Postdoc; Non-U.S. citizen; Female	711	0.05
	Non-Postdoc; Non-U.S. citizen; Male	495	0.05
	Non-Postdoc; U.S. citizen–White; Female	999	0.05
	Non-Postdoc; U.S. citizen–White; Male	2,564	0.03
	Non-Postdoc; U.S. citizen–Asian; Female	1,071	0.05
	Non-Postdoc; U.S. citizen–Asian; Male	616	0.05
	Non-Postdoc; U.S. citizen–Minority; Female	765	0.05
	Non-Postdoc; U.S. citizen–Minority; Male	600	0.05

Domain level	Category	Minimum sample size[a]	Expected CV[b]
GSS substrata × Postdoc Status × Sex × Citizenship-Race-Ethnicity for first 2 strata (Medical schools/ centers, and Very high research activity)[c]	Med-schools; Postdoc; Non-U.S. citizen; Female	313	0.06
	Med-schools; Postdoc; Non-U.S. citizen; Male	505	0.06
	Med-schools; Postdoc; U.S. citizen–White; Female	222	0.08
	Med-schools; Postdoc; U.S. citizen–White; Male	223	0.08
	Med-schools; Postdoc; U.S. citizen–Asian; Female	109	0.10
	Med-schools; Postdoc; U.S. citizen–Asian; Male	110	0.10
	Med-schools; Postdoc; U.S. citizen–Minority; Female	104	0.10
	Med-schools; Postdoc; U.S. citizen–Minority; Male	102	0.10
	Med-schools; Non-Postdoc; Non-U.S. citizen; Female	77	0.12
	Med-schools; Non-Postdoc; Non-U.S. citizen; Male	102	0.12
	Med-schools; Non-Postdoc; U.S. citizen–White; Female	429	0.08
	Med-schools; Non-Postdoc; U.S. citizen–White; Male	408	0.08
	Med-schools; Non-Postdoc; U.S. citizen–Asian; Female	178	0.08
	Med-schools; Non-Postdoc; U.S. citizen–Asian; Male	180	0.08
	Med-schools; Non-Postdoc; U.S. citizen–Minority; Female	143	0.10
	Med-schools; Non-Postdoc; U.S. citizen–Minority; Male	114	0.10
	Very-High-Research; Postdoc; Non-U.S. citizen; Female	341	0.07
	Very-High-Research; Postdoc; Non-U.S. citizen; Male	536	0.07
	Very-High-Research; Postdoc; U.S. citizen–White; Female	393	0.07
	Very-High-Research; Postdoc; U.S. citizen–White; Male	401	0.07
	Very-High-Research; Postdoc; U.S. citizen–Asian; Female	133	0.12
	Very-High-Research; Postdoc; U.S. citizen–Asian; Male	165	0.10
	Very-High-Research; Postdoc; U.S. citizen–Minority; Female	113	0.12
	Very-High-Research; Postdoc; U.S. citizen–Minority; Male	108	0.10
	Very-High-Research; Non-Postdoc; Non-U.S. citizen; Female	173	0.12
	Very-High-Research; Non-Postdoc; Non-U.S. citizen; Male	244	0.08
	Very-High-Research; Non-Postdoc; U.S. citizen–White; Female	1,067	0.05
	Very-High-Research; Non-Postdoc; U.S. citizen–White; Male	970	0.05
	Very-High-Research; Non-Postdoc; U.S. citizen–Asian; Female	239	0.12
	Very-High-Research; Non-Postdoc; U.S. citizen–Asian; Male	300	0.08
	Very-High-Research; Non-Postdoc; U.S. citizen–Minority; Female	178	0.12
	Very-High-Research; Non-Postdoc; U.S. citizen–Minority; Male	177	0.12
FFRDC			
Postdoc Status	Non-Postdoc	438	0.05
	Postdoc	406	0.05
NIH IRP			
Postdoc Status	Non-Postdoc	123	0.12
	Postdoc	287	0.08

[a] The minimum sample size in this column is the sample size threshold that is set to ensure that all domains would have effective sample sizes larger than or equal to the threshold sample sizes. In this exercise, the minimum sample size is calculated based on the pre-specified expected CV under the conservative calculation using proportion of 0.5 and the design effect calculated from Pilot ECDS data.

[b] Expected (or desired) CVs were provided by the NSF. The expected CVs are developed based on reviewing analytical goals and the estimated CVs achievable under the full sample size of 18,000.

[c] Constraints were not set for the domains Postdoc Status x Sex x Citizenship-Race-Ethnicity in the "GSS high research activity and "GSS All other colleges and universities." The population sizes are so small in these domains that achieving adequate precision would require selecting a very high proportion of the ECD.

In table 1, we present the list of domains of interest for analyses and tabulations for the U.S. academic institutions. We proposed the minimum sample size by inflating the effective sample size by the design effect due to unequal weight variation.[4] Precision in this table is expressed as the coefficient of variation for estimating a proportion of an ECD characteristic within domain, where the proportion is set at 0.5. The computation of the minimum e sample size in Table 1 is described below.

Suppose we want to estimate a proportion (or mean) of a certain characteristic for ECD within a certain domain d, for example, a proportion of respondents who expressed a change in their career track interest within U.S. citizen early career doctorates. Let P_d denotes the proportion of U.S. citizen early career doctorates who expressed change in their career track interest. The estimate of P_d, denoted by p_d is calculated based on sample of size n_d, and this estimate has variance $Var(p_d)$. For the purpose of calculating the sample size, this variance can be expressed as:

$$Var(p_d) = \left(1 - \frac{n_d}{N_d}\right)\frac{P_d(1-P_d)}{n_d} Deff_d$$

where N_d denotes the population size, and $Deff_d$ is the overall design effect due to unequal weight variation and from clustering as a result of the two-stage sampling. The design effect estimate $Deff_d$ used in the allocation was obtained from the Pilot ECDS. This formula can be inverted for sample size calculation:

$$n_d = \frac{P_d(1-P_d)Deff_d}{Var(p_d) + \frac{P_d(1-P_d)}{N_d} Deff_d}.$$

Replacing the variance with $CV^2(p_d) = Var(p_d)/P_d^2$, and using $P_d = 0.5$, this formula can be simplified as:

$$n_d = \frac{Deff_d}{CV^2(p_d) + \frac{Deff_d}{N_d}}.$$

To achieve the required minimum effective sample size above, the sample of 18,000 ECD needs to be allocated to the domains by accounting for the increase in design effect due to weight variation within domains. This is done iteratively in the following steps:

- Allocate 18,000 proportionally to all domains,
- adjusting for effective sample size in FFRDC and NIH IRP domains x Postdoc Status,
- adjusting for effective sample size in domains defined by GSS stratum (only for GSS medical schools and centers, and GSS very high research activity university, ignoring GSS high research activity and all other GSS colleges and university) x Postdoc x Gender x Race-Foreign,
- adjusting for effective sample size in domains defined by Postdoc x Gender x Race-Foreign,
- adjusting for effective sample size in domains defined by GSS substrata x Postdoc,
- adjusting for effective sample size in domains defined by GSS substrata,
- adjusting for effective sample size in domains defined by postdoc,

[4] An effective sample size can be defined as a ratio of actual sample size to the design effect due to unequal weight variation: $n_{eff} = n/deff_w$, where $deff_w = n\sum_i w_i^2/(\sum_i w_i)^2$. When there is no variability in the weights, $n_{eff} = n$. The effective sample size is used here instead of just the sample size because when the weights vary within domain, this weight variation will increase the variance of estimates.). The effective sample size has taken into account such weight variation.

- adjusting for effective sample size in domains defined by citizenship-race-ethnicity,
- adjusting for effective sample size in domains defined by gender, and
- adjusting for effective sample size in overall domain.

In each step above, the sample size allocation takes into account the design effect due to unequal weights variation, to ensure that the minimum effective sample size would produce precision that meets the pre-specified CV. The adjustments are carried out as follows:

(a) Proportionally allocate the sample size of 18,000 to the 68 domains defined by 64 domains of GSS Institution Type × Postdoc Status × Sex × Citizenship-Race-Ethnicity (the lowest domain level) and the 4 domains of FFRDC/NIH × Postdoc Status.

(b) Calculate the design effects and the effective sample sizes (at the first cycle, the design effect is 1 because sample sizes are proportional). Check if any of the domains above has the effective sample size less than that specified in the above table.

(c) For a given level of domains, adjust the sample size in domains where the effective sample size is less than that specified as follows. Suppose in a specific level of domains, there are *d1* domains (*d1*>0) where their sample size is less than specified. For these *d1* domains, calculate the adjustment factor as:

$$af = \frac{n_{min}}{n_{sampcell}},$$

where n_{min} and $n_{sampcell}$ are, respectively, the threshold/minimum effective sample size and the original sample size in the sampling cell. Inflate the original sample size in the sampling cell in these *d1* domains by multiplying it by af; that is, $n_{adjust} = n_{sampcell} \times af$. For the remaining domains, recalculate the sample size by allocating the remaining total sample size proportionally to the remaining domains.

(d) For the next domain level, calculate the design effect and the effective sample size within each domain. Check if any of these domains has sample size less than specified. Suppose there are *d2* domains (*d2*>0) where their sample size is less than specified. For *d2* domains where the effective sample size is less than specified, calculate the adjustment factor af and inflate the original sample size in the sampling cell in these *d2* domains by multiplying it by af. For the remaining domains, recalculate the sample size by allocating the remaining sample size, proportionally to the remaining domains, while also keeping the minimum sample size assigned in the previous iteration. That is, when allocating the sample proportionally, maintain the minimum sample size requirement; this is done by using distribution of sample allocation in the previous step (that meet the minimum sample size threshold).

(e) Repeat these processes for all other levels of domain. In each level of domain, check if any of domains has the sample size below threshold, and then adjust as in (c) or (d).

Note that if similar variables (Postdoc Status, Sex, and Citizenship-Race-Ethnicity) are available to construct domains in the FFRDCs and NIH IRPs as in the GSS data, the population counts in these two strata can be combined into the above exercise to allocate a total of 18,000 sampled ECD.

Table 2 shows the resulting numbers of responding, eligible ECD that are needed to satisfy the precision. **Table 3** gives there sampling rates. Note that the sampling rates within the domains and first stage institution strata are not all constant; this variability in the sampling rates is a consequence of allocating the fixed sample size of 18,000 to the strata and domains in order to satisfy multiple variance constraints. Under this allocation, the design effect due to weight variation for the GSS is 1.09, and for the entire

sample (GSS, FFRDC, and NIH IRP combined) is 1.11. Note also that these are numbers of responding, eligible ECD; the actual number to be sampled will be obtained by inflating by the anticipated response and eligibility rates.

Table 2. Sample size allocation of 18,000 ECD by sampling domains

Stratum	Non-Postdoc								Total Non-Postdoc
	Foreign		White		Asian		Minority		
	Female	Male	Female	Male	Female	Male	Female	Male	
GSS									
Medical schools and centers	124	102	468	520	302	164	143	114	1,937
Very high research activity	281	244	1,068	972	343	300	178	177	3,563
High research activity	182	163	800	835	175	165	181	142	2,643
All other colleges and univ.	189	114	1,194	1,014	297	145	298	218	3,469
FFRDC									439
NIH IRP									123
Total									12,174

Stratum	Postdoc								Total Postdoc	Stratum Total
	Foreign		White		Asian		Minority			
	Female	Male	Female	Male	Female	Male	Female	Male		
GSS										
Medical schools and centers	367	522	256	233	104	104	112	101	1,799	3,736
Very high research activity	432	845	442	446	126	156	122	107	2,676	6,239
High research activity	63	136	56	60	10	18	11	10	364	3,007
All other colleges and univ.	59	114	35	46	9	14	7	10	294	3,763
FFRDC									406	845
NIH IRP									287	410
Total									5,826	18,000

Table 3. Sampling Rates for the second stage strata for the allocation shown in Table 2

Stratum	Non-Postdoc								Total Non-Postdoc
	Foreign		White		Asian		Minority		
	Female	Male	Female	Male	Female	Male	Female	Male	
GSS									
Medical schools and centers	13.7%	6.1%	4.0%	4.0%	8.8%	4.0%	5.3%	5.4%	4.9%
Very high research activity	12.8%	5.5%	6.5%	5.0%	8.8%	5.9%	5.4%	6.0%	6.2%
High research activity	17.8%	9.4%	9.4%	9.4%	14.5%	9.5%	9.8%	9.8%	10.0%
All other colleges and univ.	12.4%	5.3%	5.4%	5.4%	10.1%	5.4%	5.6%	5.6%	5.8%
FFRDC									8.9%
NIH IRP									19.9%

Stratum	Postdoc								Total Postdoc
	Foreign		White		Asian		Minority		
	Female	Male	Female	Male	Female	Male	Female	Male	
GSS									
Medical schools and centers	7.5%	6.7%	7.8%	6.8%	8.9%	8.5%	15.2%	15.6%	7.8%
Very high research activity	7.4%	6.7%	9.0%	6.8%	8.9%	8.4%	12.8%	11.5%	7.6%
High research activity	11.8%	10.8%	12.0%	10.6%	9.3%	9.0%	12.2%	10.2%	11.0%
All other colleges and univ.	16.4%	14.4%	16.1%	14.6%	11.3%	11.9%	14.9%	12.7%	14.6%
FFRDC									15.6%
NIH IRP									16.4%

B. First-stage Sampling: Selection of Institutions

A total of approximately 300 responding institutions will be included in this survey. The sample of institutions will be selected through a probability proportional to size (PPS) sampling. First, the type of institutions (U.S. academic institution, FFRDC, and NIH IRP) serves as sampling strata in this first-stage of sampling (Primary Sampling Unit/PSU strata). The selection of GSS institutions will be independent of the selection of the FFRDS institutions and NIH programs. All NIH IRPs (25 programs) will be selected with certainty, while the institutions in the other strata will be sampled. The first-stage sampling strata that will also the base for domain of analysis, and the population of institutions by stratum is given in **Table 4** below:

Table 4. Institution count in the population by stratum

Stratum number	Description of type of institutions	Number of institutions in the population	Expected number of responding institutions in the sample
1	GSS Medical schools and centers	172	53
2	GSS Very high research activity universities	109	76
3	GSS High research activity universities	98	54
4	GSS All other colleges and universities	461	67
5	FFRDC	43	25
6	NIH IRP	25	25
Total		908	300

For the purposes of this sampling plan, *h*, *i*, *j*, and *k*, respectively, indicate indexes for stratum, institution, domain, and ECD as follows:

h = index for the first-stage sampling stratum; $h = 1, \ldots, 6$ (U.S. academic institution, FFRDC, NIH IRP)

i = index for institution; $i = 1, \ldots, I_h$, where I_h = the total number of eligible institutions in stratum h in the frame

j = index for domain; $j = 1, \ldots, J$, where J is the number of domains of interest

k = index for ECD.

Under the PPS sampling, the measure of size for each eligible institution i within stratum h in the frame will be determined as a composite measure of size S_{hi} as follows (see Folsom, Potter, and Williams, 1987 for more details on composite size measures):

$$S_{hi} = \sum_{j=1}^{J} n_{hj} \frac{N_{hij}}{N_{hj}} = \sum_{j=1}^{J} f_{hj} N_{hij} \qquad (1)$$

where

f_{hj} = the sample fraction of ECD for domain j in PSU stratum h; $f_{hj} = n_{hj}/N_{hj}$

N_{hij}= the total number of ECD for domain j in institution i within PSU stratum h

n_{hj} = the sample size of ECD allocated for domain j in PSU stratum h

N_{hj}= the total number of ECD for domain j in PSU stratum h; $N_{hj} = \sum_{i=1}^{I_h} N_{hij}$.

Note that a composite of size S_{hi} is a summation of measure of size across J domains; that is, $S_{hi} = \sum_{j=1}^{J} S_{hij}$, where

$$S_{hij} = \frac{n_{hj} N_{hij}}{N_{hj}}. \qquad (2)$$

In addition, the sum of composite measure of sizes across all institutions in the GSS frame constitutes the total sample size of ECD in the first four strata (GSS strata), which is *n* = 16,748:

$$\sum_{h=1}^{4}\sum_{i=1}^{I_h} S_{hi} = \sum_{h=1}^{4}\sum_{i=1}^{I_h}\sum_{j=1}^{J} n_{hj}\frac{N_{hij}}{N_{hj}} = \sum_{h=1}^{4}\sum_{j=1}^{J} n_{hj}\frac{\sum_{i=1}^{I_h} N_{hij}}{N_{hj}} = \sum_{h=1}^{4}\sum_{j=1}^{J} n_{hj} = \sum_{h=1}^{4} n_h = n. \qquad (3)$$

Similarly, the sums for the last two strata (FFRDC and NIH IRP) are 844 and 410, respectively.

The sample size of ECD allocated for each domain, n_{hj}, needs to be determined prior to sample selection (done in the previous section), and the domain size N_{hij} needs to be available.

Since all programs in NIH IRP will be selected, we do not need to calculate the selection probabilities as will done for the other strata as follows. Given the composite measure of size S_{hi} above, the probability selection for each institution in the first five PSU strata can be determined as follows:

$$\pi_{hi} = m_h \frac{S_{hi}}{\sum_{i=1}^{I_h} S_{hi}}. \qquad (4)$$

where

m_h = the sample size of institutions (PSUs) allocated for stratum *h*.

For large institutions, the value of selection probability above may be greater than 1. Such institutions will be selected with certainty. We will identify the institutions selected with certainty in strata 1-5 iteratively. That is,

(a) The first round of iteration is calculating selection probabilities as in formula (4)

(b) Identify certainty institutions based on selection probabilities calculated in (a), and set aside these certainty institutions from the frame. So we have m_h^{C1} and m_h^{NC1}, respectively, denotes the sample size of certainty institutions and non-certainty institutions identified at the first round of iteration, where $m_h = m_h^{C1} + m_h^{NC1}$. (Note: superscripts *C1* indicates certainty in the first round and *NC1* indicates Non-Certainty in the first round.)

(c) After dropping the certainty institutions from the frame, recalculate the selection probability for the non-certainty institutions:

$$\pi_{hi} = (m_h - m_h^{C1}) \times \frac{S_{hi}}{\sum_{i=1}^{(I_h - m_h^{C1})} S_{hi}} \qquad (5)$$

(d) Continue with second round of iteration, that is to identify new certainty institutions m_h^{C2} based on selection probability in (5), and recalculate the selection probability under the new sample size.

(e) Repeat the process of calculating selection probability and identifying the certainty institutions until there are no more certainty institutions identified in the frame.

Suppose $m_h^C = m_h^{C1} + m_h^{C2} + \cdots,$ and m_h^{NC}, respectively, denotes the final sample size of certainty institutions and non-certainty institutions, where $m_h = m_h^C + m_h^{NC}$. Among the remaining non-certainty institutions in the frame, we draw a sample of institutions in each stratum, with size m_h^{NC} institutions. At the end of this process, the probability of selection is determined as:

Certainty U.S. academic institutions in stratum *h*: $\pi_{hi} = 1$ (6)

Certainty FFRDC: $\pi_{5i} = 1$

Certainty NIH IRP: $\pi_{6i} = 1$

Non-certainty U.S. academic institutions in stratum *h*: $\pi_{hi} = m_h^{NC} \times \frac{S_{hi}}{\sum_{i=1}^{(I_h - m_h^C)} S_{hi}}$

Non-certainty FFRDC: $\pi_{5i} = m_5^{NC} \times \frac{S_5}{\sum_{i=1}^{(I_5 - m_5^C)} S_{5i}}$

C. Second-Stage Sampling: Selection of ECD

1. Sample Allocation

In this second-stage sample selection, we will select a total of 16,748 ECD from the U.S. academic institutions, 844 ECD from the FFRDC, and 410 ECD from the NIH IRPs. The sample allocation for each domain n_j has been determined earlier (table 2). Now, the goal in this stage is to, first, allocate n_j to each sampled institutions so that this allocation will result in a self-weighting sample within domain. That is, at the end of sampling process, the unconditional selection probability of ECD is the same across ECD within domain. Second, we will determine a sampling method for selecting ECD within sampled institutions.

The following sample size allocation is exercised for the initial calculation:

- **Initial sample size within institution**:

  To achieve self-weighting sample within domain, the sample size in the certainty institutions should be allocated proportionally based on the composite measure of size, while the sample size in the non-certainty institutions should be allocated equally across non-certainty institutions as follows:

  Certainty U.S. academic institutions: $n_{hi} = n_h \frac{S_{hi}}{\sum_{i=1}^{I_h} S_{hi}} = S_{hi}$ (7)

  Certainty FFRDC: $n_{5i} = n_5 \frac{S_{5i}}{\sum_{i=1}^{I_5} S_{5i}} = S_{5i}$

  Certainty NIH IRP: $n_{6i} = n_6 \frac{S_{6i}}{\sum_{i=1}^{I_6} S_{6i}} = S_{6i}$

  Non-certainty U.S. academic institutions: $n_{hi} = \frac{\sum_{i=1}^{(I_h - m_h^C)} S_{hi}}{m_h^{NC}}$

  Non-certainty FFRDC: $n_{5i} = \frac{\sum_{i=1}^{(I_5 - m_5^C)} S_{5i}}{m_5^{NC}}$

- **Sample size within institution and domain**:

  The allocation of sample size within institution to each domain (within institution) is:

$$n_{hij} = n_{hi}\frac{S_{hij}}{S_{hi}}. \qquad (8)$$

  The following expressions are obtained by substituting n_{hi} and S_{hij} in (8) with that in (7) and (2), respectively:

  Certainty U.S. academic institutions: $n_{hij} = S_{hi}\frac{S_{hij}}{S_{hi}} = S_{hij} = \frac{n_{hj}N_{hij}}{N_{hj}} = N_{hij}\times f_{hj}$

  Certainty FFRDC: $n_5 = S_{5i}\frac{S_{5ij}}{S_{5i}} = S_{5ij} = n_{5j}N_{5ij}/N_{5j} = N_{5ij}\times f_{5j}$

  Certainty NIH IRP: $n_{6ij} = S_{6i}\frac{S_{6ij}}{S_{6i}} = S_{6ij} = n_{6j}N_{6ij}/N_{6j} = N_{6ij}\times f_{6j}$

  Non-certainty U.S. academic institutions:

$$n_{hij} = \frac{\sum_{i=1}^{(I_h - m_h^C)} S_{hi}}{m_h^{NC}} \times \frac{1}{S_{hi}} \times \frac{n_{hj}N_{hij}}{N_{hj}} = \frac{\sum_{i=1}^{(I_h - m_h^C)} S_{1i}}{m_h^{NC}} \times \frac{N_{hij}}{S_{hi}} \times f_{hj}$$

  Non-certainty FFRDC:

$$n_{5ij} = \frac{\sum_{i=1}^{(I_5 - m_5^C)} S_{5i}}{m_5^{NC}} \times \frac{1}{S_{5i}} \times \frac{n_5 N_{5ij}}{N_{5j}} = \frac{\sum_{i=1}^{(I_5 - m_5^C)} S_{1i}}{m_5^{NC}} \times \frac{N_{5ij}}{S_{5i}} \times f_{5j}$$

To see whether the above sample allocations produce self-weighting sample, we can calculate the unconditional selection probability of ECD. The unconditional probability of ECD selection is a multiplication of institution selection probability and the conditional ECD selection probability within institution, where the conditional probability in the second stage is calculated as:

$$\pi_{hjk|i} = \frac{n_{hij}}{N_{hij}}. \qquad (9)$$

Therefore, the unconditional selection probability of ECD k in domain j in institution i and stratum h can be calculated as follows:

$$\pi_{hijk} = \pi_{hi}\times\pi_{hjk|i} \qquad (10)$$

The following expressions are obtained by substituting π_{hi}, $\pi_{hjk|i}$, and n_{hij} in (10), with that in (6), (9), and (8), respectively:

Certainty U.S. academic institutions: $\pi_{hijk} = \pi_{hi} \times \pi_{hjk|i} = 1 \times \frac{1}{N_{hij}} N_{hij} f_{hj} = f_{hj}$

Certainty FFRDC: $\pi_{5ijk} = \pi_{5i} \times \pi_{5jk|i} = 1 \times \frac{1}{N_{5ij}} N_{5ij} f_{5j} = f_{5j}$

Certainty NIH IRP: $\pi_{6ijk} = \pi_{6i} \times \pi_{6jk|i} = 1 \times \frac{1}{N_{6ij}} N_{6ij} f_{6j} = f_{6j}$

Non-certainty U.S. academic institutions:

$$\pi_{hijk} = \pi_{hi} \times \pi_{hjk|i} = \left(m_h - m_h^C\right) \frac{S_{hi}}{\sum_{i=1}^{(I_h - m_h^C)} S_{hi}} \times \frac{1}{N_{hij}} \frac{\sum_{i=1}^{(I_h - m_h^C)} S_{hi}}{m_h^{NC}} \frac{N_{hij}}{S_{hi}} f_{hj} = f_{hj}$$

Non-certainty FFRDC:

$$\pi_{5ijk} = \pi_{5i} \times \pi_{5jk|i} = \left(m_5 - m_5^C\right) \frac{S_{5i}}{\sum_{i=1}^{(I_5 - m_5^C)} S_{5i}} \times \frac{1}{N_{5ij}} \frac{\sum_{i=1}^{(I_5 - m_5^C)} S_{5i}}{m_5^{NC}} \times \frac{N_{5ij}}{S_{5i}} \times f_{5j} = f_{5j}$$

We can see that within domain j, the allocation in (8) results in equal selection probability within the stratum but not across the strata. This is because the institutional strata are also analytic domains and higher sampling rates are needed in some of the strata in order to satisfy the precision requirements.

The allocation in (8) can be adjusted to result in equal selection probability across strata as follows:

Certainty U.S. academic institutions: $n_{1ij} = N_{1ij} \times f_j$ (11)

Certainty FFRDC: $n_{2ij} = N_{2ij} \times f_j$

Certainty NIH IRP: $n_{3ij} = N_{3ij} \times f_j$

Non-certainty U.S. academic institutions:

$$n_{1ij} = \frac{\sum_{i=1}^{(I_1 - m_1^C)} S_{1i}}{m_1^{NC}} \times \frac{N_{1ij}}{S_{1i}} \times f_j$$

Non-certainty FFRDC:

$$n_{2ij} = \frac{\sum_{i=1}^{(I_2 - m_2^C)} S_{2i}}{m_2^{NC}} \times \frac{N_{2ij}}{S_{2i}} \times f_j$$

where f_j is an overall sample fraction for domain j calculated across all strata. (Note that the sample allocation (11) may produce non-integer sample size. We will come back to this issue later.)

Now, if we substitute the sample allocation in (11) into (10), the resulting unconditional ECD selection probability within domain are all equal to f_j as shown below:

Certainty U.S. academic institutions: $\pi_{1ijk} = \pi_{1i} \times \pi_{1jk|i} = 1 \times \frac{1}{N_{1ij}} N_{1ij} f_j = f_j$ (12)

Certainty FFRDC: $\pi_{2ijk} = \pi_{2i} \times \pi_{2jk|i} = 1 \times \frac{1}{N_{2ij}} N_{2ij} f_j = f_j$

Certainty NIH IRP: $\pi_{3ijk} = \pi_{3i} \times \pi_{3jk|i} = 1 \times \frac{1}{N_{3ij}} N_{3ij} f_j = f_j$

Non-certainty FFRDC:

$$\pi_{1ijk} = \pi_{1i} \times \pi_{1jk|i} = \left(240 - m_1^C\right) \frac{S_{1i}}{\sum_{i=1}^{(I_1 - m_1^C)} S_{1i}} \times \frac{1}{N_{1ij}} \frac{\sum_{i=1}^{(I_1 - m_1^C)} S_{1i}}{m_1^{NC}} \frac{N_{1ij}}{S_{1i}} f_j = f_j$$

Non-certainty U.S. academic institutions:

$$\pi_{2ijk} = \pi_{2i} \times \pi_{2jk|i} = \left(40 - m_2^C\right) \frac{S_{2i}}{\sum_{i=1}^{(I_2 - m_2^C)} S_{2i}} \times \frac{1}{N_{2ij}} \frac{\sum_{i=1}^{(I_2 - m_2^C)} S_{1i}}{m_2^{NC}} \times \frac{N_{2ij}}{S_{2i}} \times f_j = f_j$$

2. Sample Selection

Equation (11) gives the sample allocation n_{hij} for selecting the ECD within sampled institutions, however, these numbers are not integer. One may round these number to integer and use them as the sample size with rounding. However, the original sampling rate:

$$\pi_{hj|i} = \frac{n_{hij}}{N_{hij}}, \qquad (13)$$

where the numerator is unrounded institutional-level domain sample size, would not be retained.

To overcome this, we can implement a PPS sampling with the sampling rate (13) used as the measure of size. If these sampling rates are used as the measure of size in PPS sampling when selecting ECD, the selection will result in random rounding that result in integer sample size. The PPS sequential sampling where the frame is sorted by PSU strata, institution, and domain variables, can be used to select the 18,000 ECD.

D. Treatment of Missing Variables for Defining Domains

Selection of sampled ECD in the second stage of sampling will use stratification based on the following variables:

- Postdoc Status (2 levels): Postdoc, Non-Postdoc
- Sex (2 levels): Male, Female
- Citizenship-Race-Ethnicity (4 levels): Non-U.S. citizen, U.S. citizen–White, U.S. citizen–Asian, U.S. citizen–Other

We will request these information to be included on the ECD lists from the institutions sampled in the first stage of sampling. We expect to be able to get complete information on postdoc status, however some institutions may not provide this information. Sex and Citizenship-Race-Ethnicity will be missing entirely from some lists and for a subset of individuals on other lists. This section describes procedures for imputing missing values prior to the selection of the second stage sample members.

1. Imputation for Postdoc Status

We anticipate that almost all institutions will be able to provide the postdoc status for the individuals on their ECD list (or tell us which job titles represent postdoc positions). In the pilot ECDS, only one

FFRDC did not provide this information. We will use the job titles and pilot ECDS responses to impute postdoc status where missing in the frame.

2. Imputation for Sex

We anticipate that most institutions will be able to provide sex information for most individuals on their ECD list, but as many as 10% of list members may be missing this information. Any missing sex data will be imputed using several external databases. First, we will attempt to match the list member to the Survey of Earned Doctorates (SED) and, if we are able to link to these data sets, use the sex of the individual from the SED to fill in the missing ECD frame data. The linking process will be explained in a separate section later.

For all remaining cases where we have a name, we propose to use the database of names by sex maintained by the Social Security Administration (SSA) to impute the missing sex data. These databases provide a list of first names for individuals born in the U.S. in a given year, along with the count by sex. The databases include all name/sex combinations that occurred at least 5 times in a given year. A description of the database is at http://www.ssa.gov/oact/babynames/limits.html, and the national-level data is at http://www.ssa.gov/oact/babynames/names.zip. We would pool the names and counts to arrive at percentages that are male and female for each name. We would start by using first name. If the percentage of times a name is a given sex is very high (for example, greater than 90 percent), then any individuals with missing sex and that first name will be assigned to that sex.

Next, the middle name will be examined for any ECD that still do not have sex assigned, and any of ECD whose middle name is in the list with a high percentage being one sex (for example, greater than 90 percent) will be assigned to that sex.

After this step, we will randomly assign sex to any remaining using the database of names. For each ECD with missing sex whose first name appears on the list, we would generate a uniform random number, compare this random number to the distribution, and impute the sex. For example, if 40 percent of a given name is Male and 60 percent is Female, and the generated random number is 0.40 or less, then the sex would be imputed as "Male," and random numbers greater than 0.40 would impute the sex to "Female." Any cases with names that are not in the SSA Names by Sex database will be examined and assigned manually. Some of these may be foreign names with entries in similar name by sex databases focusing on names from other countries.

For cases without names, we will randomly assign sex based on the distribution of individuals by sex within the institution. As with the name based imputation, if 65 percent of the predicted number of ECD at the focal institution were men based on the combined GSS and IPEDS data, then a random number is 0.65 or less, then the sex would be imputed as "Male," and random numbers greater than 0.65 would impute the sex to "Female."

3. Imputation for Race/Ethnicity

Although most institutions track by race/ethnicity, some institutions may not be willing or able to provide it for many individuals on their ECD lists. When race/ethnicity is missing, we suggest using a combination of logical editing and imputation to fill in the missing values. As with sex, we will attempt to match the case to the SED and fill in missing race/ethnicity for the ECDS frames in an individual match can be found.

For the remaining cases that are missing race/ethnicity but include last name, we will use the U.S. Census database of surnames (http://www2.census.gov/topics/genealogy/2000surnames/names.zip) that gives the percentage of times each of the surnames that is white, black, Asian/Pacific Islander, American Indian, or Hispanic. A description of the database is located here: http://www.census.gov/topics/population/genealogy/data/2000_surnames.html. In reviewing this database, we see that some surnames fall almost exclusively into only one of the race or ethnic groups. We would extract names that are highly likely to be of one particular race or ethnicity (for example more than 80 percent Asian/Pacific Islander), and assign any missing ECD with that last name to that race/ethnicity.

All surnames with missing race/ethnicity that are not in the Census database will be manually reviewed in conjunction with the first and middle names to see if a logical assignment can be made (e.g., Hispanic or Asian/Pacifica Islander). Finally, a random assignment using the database of surnames would be used to fill in any missing data that remain within the cases with surnames. For a given name, we will use the percentage provided in the census database for that name to randomly assign the name to the race/ethnicity. For example if the percentage for a particular name were as follows:

Race/ethnicity	Percentage	Cumulative Percentage
White	73.35	73.35
Black or African American	22.22	95.57
Asian or Pacific Islander	0.40	95.97
American Indian or Alaskan Native	0.85	96.82
Two or more races	1.63	98.45
Hispanic	1.55	100.00

Then, a missing race with random number 0.8 will be imputed with Black or African American (since 0.8 = 80 percent is between 73.35 and 95.57 percent).

4. Imputation for Citizenship

Based on the results of the pilot survey, not all institutions will provide an indicator for whether an individual is a U.S. citizen. SED collects data on the citizenship status of doctorates in two variables: (1) citizenship at birth, and (2) citizenship at doctoral graduation. Though there could be citizenship status change between the time of graduation and time of survey, this data can be used to impute missing citizenship status in the list from institutions.[5] For any ECD with missing citizenship status and can be linked to SED, we will use the two variables of citizenship status for imputation as follows:

SED citizenship at birth	SED citizenship at graduation	Imputed citizenship status
U.S. citizen	U.S. citizen	U.S. citizen
Non-U.S. citizen	U.S. citizen	U.S. citizen
U.S. citizen	Non-U.S. citizen	Non-U.S. citizen
Non-U.S. citizen	Non-U.S. citizen	Number of years since graduated: < 5 years: Non-U.S. citizen ≥ 5 years: random imputation (below)

[5] Citizenship status will be collected during the ECDS survey, so we can assess the magnitude of misclassification error that may occur during the sampling.

For cases where citizenship at birth and at graduation are both non-U.S. citizen, and the number of years since graduated is greater than or equal to 5 years, random imputation will be based on the number of years since graduated, assuming the longer the years the more likely to change the citizenship status. For example, we would assign cases with number of years since graduated 5, 6, 7 years probability of U.S. citizen 0.4, and cases with number of years since graduated 8, 9, 10 years probability of U.S. citizen 0.6..

For cases that cannot be linked with SED, we will use any indication of non-U.S. origin of doctoral degree provided on the frame to impute missing citizenship status to Non-U.S. citizen. Then we will use the name-race/ethnicity database and impute missing citizenship status to U.S. citizen when the race is White with high percentage (greater than or equal to 90 percent). After that, any remaining cases with missing citizenship status will be reviewed manually with the help of information available from the list.

For cases missing name and citizenship-race-ethnicity, citizenship-race-ethnicity will be imputed randomly using the institution level percentages derived from the GSS and IPEDS data when developing the composite ECD size measures for each institution.

5. Linking Institution's List of ECD with the SED Data

For ECD with earned doctorate degrees from U.S. institutions, cases with missing sex, race/ethnicity, or citizenship status in ECDS lists will be linked to SED based on several key variables such as academic institution of doctorate, doctorate degree year, last name, first name, birth year (if available), and sampling variables. Combinations of these key variables will be used to maximize the linking rates. For example to get sex from SED for missing sex in the institution list, first we will link the SED and ECD list using the most variables that are available, for example:

- academic institution of doctorate, degree year, last name, first name, race/ethnicity, and birth year.

Remaining un-linked cases will be linked sequentially using less number of key variables as follows:

- academic institution of doctorate, degree year, last name, race/ethnicity, birth year,
- academic institution of doctorate, degree year, last name, first name, race/ethnicity,
- academic institution of doctorate, degree year, last name, race/ethnicity,
- academic institution of doctorate, degree year, last name, first name,
- academic institution of doctorate, degree year, last name,
- academic institution of doctorate, last name, first name, race/ethnicity,
- academic institution of doctorate, last name, race/ethnicity,
- academic institution of doctorate, last name, first name,

Similarly, for linking SED and the institution's list to obtain race/ethnicity, we can use combinations of sex, academic institution of doctorate, degree year, last name, first name, and birth year as key variables for linking. To obtain citizenship status, we can use combinations of sex, race/ethnicity, academic institution of doctorate, degree year, last name, first name, and birth year as key variables for linking.

6. Evaluation of the Imputation for Frame Data

The level of missing data in the frame variables is not known at this time for all of the variables because they were not requested on the institution lists in the pilot survey. It will be important to evaluate the

imputation procedures and improve on it if possible. We suggest the following tabulations and analyses of the missing frame data and to evaluate the success of the imputation:

- tabulate counts and rates of missing data as each institution's frame is received,
- tabulate counts of matches to the SED and Census data bases,
- compare demographic distributions (including the imputed data) to distributions from IPEDs and the GSS,
- compare the data from the two sources using statistics such as Cohen's kappa or the intraclass correlation when the variables race, sex, etc. are provided on the institution frames and the ECD matches to the SED or census data base (e.g. match to the SED or the name matches a census data base for the variables being imputed). This should give an idea of how well the procedure works when we don't have frame data. When we have frame data for groups that may be difficult to impute, such as potential foreign doctorates, we can implement the imputation procedures for those where the information is known as well as unknown to get an early look at how the procedures are working. That is, we'd follow the same imputation procedures when do have frame data; we would use the imputed values for evaluation of the procedures (but for sampling we'd use the actual frame data).
- compare frame, imputed, and data collected in the survey for the variables of interest after data collection is complete.

E. Inflating Sample Size to Account for Survey Nonresponse

The sample sizes given in the previous sections are the numbers of target completes; that is, the expected numbers of eligible survey respondents. During fielding of the survey, however, we expect to have nonrespondents and that not all of the sampled individuals are eligible. Therefore, the respondent sample sizes n_{hij} need to be inflated by dividing by the expected response and eligibility rates. That is, the initial sample size n_{hij}^{init} for stratum h, institution i, and domain j is $n_{hij}^{init} = n_{hij}/R_{hij}$, where R_{hij} is the overall expected response and eligibility rate in stratum h, institution i, and domain j. **Table 5** presents the expected nonresponse and ineligibly rates for both stages 1 and 2 of sampling.

Table 5. Assumptions: Nonparticipation, Nonresponse, and Ineligibility

	Stage 1 Sampling	Stage 2 Sampling	
Sampling stratum	% Non-participating	% Ineligible	% Non-responding
GSS			
Medical schools and centers	15.0	3.0	25.0
Very high research activity universities	10.0	2.0	17.5
High research activity universities	15.0	1.0	17.5
All other colleges and universities	20.0	1.0	20.0
FFRDC	10.0	1.0	12.5
NIH IRP	0.0	1.0	30.0

It is expected that some sampled institutions will not respond to our request to provide list of ECD (we call this as stage-1 institution nonresponse). When the institution response rate is known or can be

estimated, the number of institutions to be sampled can be calculated as the target completes (i.e., number of institutions providing lists) divided by the institution response rate. For example, the initial institutional sample size for GSS medical schools and centers is 63 institutions because the target completes (institutions providing lists) from this stratum is 53 institutions and the estimated institutional response rate is 85 percent ($53 / 0.85 \approx 63$).

The actual response rate during fielding may be lower or higher than estimated. When the actual response rate is lower than estimated, then the target completes will not be achieved. On the other hand, when it is higher than estimated, we might obtain many more completes than desired. To account for this institution nonresponse, especially when the response rate is lower than expected, RTI will draw a larger initial sample of institutions. These extra institutions will serve as reserve samples which may or may not be released depending on the need. **Table 6** shows the numbers of institutions and ECD initially sampled in order to obtain the desired numbers of institutions providing lists and the desired numbers of responding, eligible ECD.

Table 6. Example of initial and desired sample sizes for the academic institutions, FFRDC, and NIH IRP (computed using the rates in Table 4)

	Institution sample		ECD sample in Institutions that Provide Lists	
Stratum	Initial sample	Target completes (i.e. providing lists)	Initial sample	Target completes
1 - GSS Medical schools and centers	63	53	5,136	3,736
2 - GSS Very high research activity universities	85	76	7,717	6,239
3 - GSS High research activity universities	64	54	3,682	3,007
4 - GSS All other colleges and universities	84	67	4,752	3,763
Total GSS	**296**	**250**	**21,287**	**16,745**
5 = FFRDC	28	25	976	845
6 = NIH IRP	25	25	592	410
OVERALL TOTAL	**349**	**300**	**22,855**	**18,000**

As discussed in this document, some large institutions will be sampled with certainty. These certainty institutions will be put into a separate strata and a proportional sample selected from within each institution. Table 7 shows an example of the number of institutions providing lists (target completes) and the number of eligible and responding and eligible ECD for each of the GSS institution strata to demonstrate the resultant certainty institution samples, and sample size of ECD within non-certainty samples.

Table 7. Sample size allocation for first-stage sampling, and estimate sampled ECD by certainty/non-certainty institution

	Population		Sample of Institutions			Sample ECD					
							Certainty institution			Non-Certainty institution	
Stratum	Institution	ECD	Total	Certainty	Non-Certainty	Total	Sample size	Max per inst	Min per inst	Sample size	Sample size per inst
1 - Medical schools and centers	172	62,854	53	16	37	3,736	1,397	128	66	2,339	63
2 - Very high research activity universities	109	92,847	76	28	48	6,239	2,841	164	72	3,398	71
3 - High research activity universities	98	29,713	54	5	49	3,007	348	96	58	2,659	54
4 - All other colleges and universities	461	61,637	67	1	66	3,763	107	107	107	3,656	55
Total GSS	**840**	**247,051**	**250**	**50**	**200**	**16,745**	**4,693**			**12,052**	
5 = FFRDC	43	7,520	25			845					34
6 = NIH IRP	25	2,368	25			410					16
OVERALL TOTAL	**908**	**256,939**	**300**			**18,000**					

We will likely set an upper bound on the number of ECD selected from the certainty institutions in order to control the burden. After NSF and RTI have finalized the precision and sample sizes, we will be able to identify the certainty institutions and can work with NSF to determine how many ECD to include from each.

The institution response rates (i.e. the proportion that provide lists) may vary from those given in **Table 5**. Rather than initially fielding all of the institutions shown in the first column of **Table 6**, we will first select a larger sample of institutions; in this larger sample, the desired ECD will also be inflated for purposes of calculation of the composite size measure so that the domain by stratum sampling rates are the same as intended. Next, we will randomly partition the initial sampled institutions into a set of mini samples called replicates (or waves) for sample release, so that each mini-sample or sample replicate is a random subset of the initial selected institutions. Under this approach, one typically releases several of the replicates at the start of the data collection period; the number initially released is selected based on an optimistic level of response so that the release would be expected to yield a respondent sample that fall short of the desired respondent quotas.

Fielding the institutions in waves will help control the number that we contact and the number from which we obtain lists. The sample will be monitored, and once a better understanding of the realized response rate is obtained we can estimate the additional sample size needed to reach the target number of institutions that provide lists. Then, the number of replicates needed to reach the additional sample size requirements is released at a subsequent point in the field collection. This process may occur in several iterations until the end of the stage-1 survey when the desired number of institutions providing lists are achieved. Waves will be maintained and released separately for each of the institution strata and certainty strata in order to have better control over the number of institutions from which we obtain lists. Because the certainty institutions are so large and important to the survey, we may choose to release all of them at the beginning of data collection.

Sampled institutions would be randomly assigned to replicates within a stratum. The number of institutions in a replicate should be small enough to provide control over the sample size of institutions

that provide lists; for example 5 to 10 institutions per wave might be a reasonable number for the ECDS. We'd prefer that the replicates within each stratum be close to equal. For example, GSS stratum 3 calls for 54 institutions to provide lists (Table 3), and an initial sample of 64 institutions need to obtain this number given the response rate assumptions. Here, we might sample 70 institutions (so that we are covered in case the response rates is less than expected), randomly divide the sample into 14 replicates of 5 institutions each, and initially field 11 replicates (55 institutions). This would leave 3 replicates, and one or more could be released if needed.

Suppose there are a total of M_h sample institutions across all replicates in stratum h, and in the replicates that are released there are m_h sample institutions. Also suppose there are R_h replicates and r_h are released.

Replicates or waves that are not released are not treated as nonrespondents for either response calculations or weighting; they are treated the same as if they had not been sampled. Suppose there are a total of M_h sample institutions across all replicates in stratum h, and in the replicates that are released there are m_h sample institutions. Weights for institutions would first be adjusted to account for the sample release, by multiplying by the factor M_h/m_h. Alternately, if all of the replicates in a stratum contain the same number of institutions, the factor could be R_h/r_h. This will be followed by an adjustment for institution nonresponse. Then, the response adjusted institution weights will be calibrated to the total number of institutions on the frame within each of the first stage strata.

As with any sampling scheme that inflates the number of units selected in order to account for nonresponse and eligibility, the implementation of institution sample waves and release will change the selection probabilities from those that are designed. However, if the expected and actual response rates are similar, the nonresponse adjustments to the weights that are made after data collection is complete should help to restore the weights so that the design effect due to unequal weighting for the respondents is close to that anticipated in the sample design.

F. Adjusting the Sample Size Allocation for Discrepancies in Counts of ECD

There may be differences between the counts of ECD counts used during the institution sampling (first stage sample selection) and those counts used during the ECD sampling (second stage sample selection). During the list collection for the second-stage sampling frame construction, we will receive list of ECD with sampling variables from the sampled institutions. This will provide a more accurate counts, while the counts used for the first stage sampling are estimates. To maintain the goal of *epsem* sample when the actual count based on the institution-provided ECD list available, we can adjust the sample size n_{hij} as follows. Suppose $\widetilde{N}_{hij}$ denotes the count of ECD provided by the institution i for domain j in stratum h. The institution-level domain-specific sample size may be recalculated as follows:

Certainty U.S. academic institutions: $\tilde{n}_{hij} = \left(\widetilde{N}_{hij}/N_{hij}\right)\times n_{hij} = \widetilde{N}_{hij}\times f_{hj}$ (12)

Certainty FFRDC: $\tilde{n}_{5ij} = \left(\widetilde{N}_{5ij}/N_{5ij}\right)\times n_{5ij} = \widetilde{N}_{5ij}\times f_{5j}$

Certainty NIH IRP: $\tilde{n}_{6ij} = \left(\widetilde{N}_{6ij}/N_{6ij}\right)\times n_{6ij} = \widetilde{N}_{6ij}\times f_{6j}$

Non-certainty U.S. academic institutions:

$$\tilde{n}_{hij} = \left(\tilde{N}_{hij}/N_{hij}\right)\times\frac{\sum_{i=1}^{(I_h-m_h^C)} S_{hi}}{m_h^{NC}}\times\frac{N_{hij}}{S_{hi}}\times f_{hj} = \frac{\sum_{i=1}^{(I_h-m_h^C)} S_{1i}}{m_h^{NC}}\times\frac{\tilde{N}_{hij}}{S_{hi}}\times f_{hj}$$

Non-certainty FFRDC:

$$\tilde{n}_{5ij} = \left(\tilde{N}_{5ij}/N_{5ij}\right)\times\frac{\sum_{i=1}^{(I_5-m_5^C)} S_{5i}}{m_5^{NC}}\times\frac{N_{5ij}}{S_{5i}}\times f_{5j} = \frac{\sum_{i=1}^{(I_5-m_5^C)} S_{9i}}{m_5^{NC}}\times\frac{\tilde{N}_{5ij}}{S_{5i}}\times f_{5j}$$

Under the condition that $\tilde{n}_{hij} \le \tilde{N}_{hij}$ for all domains and institutions, we will achieve equal weights within each stratum by domain:

Certainty U.S. academic institutions: $\tilde{\pi}_{hjk} = \pi_{hi}\times\tilde{\pi}_{hjk|i} = 1\times\left(\tilde{n}_{hij}/\tilde{N}_{hij}\right) = \frac{\tilde{N}_{hij}\times f_j}{\tilde{N}_{hij}} = f_{hj}$

Certainty FFRDC: $\tilde{\pi}_{5ijk} = \pi_{5i}\times\tilde{\pi}_{5jk|i} = 1\times\left(\tilde{n}_{5ij}/\tilde{N}_{5ij}\right) = \frac{\tilde{N}_{5ij}\times f_j}{\tilde{N}_{5ij}} = f_{5j}$

Certainty NIH IRP: $\tilde{\pi}_{6ijk} = \pi_{6i}\times\tilde{\pi}_{6jk|i} = 1\times\left(\tilde{n}_{6ij}/\tilde{N}_{6ij}\right) = \frac{\tilde{N}_{6ij}\times f_j}{\tilde{N}_{6ij}} = f_{6j}$

Non-certainty U.S. academic institutions:

$$\tilde{\pi}_{hijk} = \pi_{hi}\times\tilde{\pi}_{hjk|i} = \left(m_h - m_h^C\right)\frac{S_{hi}}{\sum_{i=1}^{(I_h-m_h^C)} S_{hi}}\times\frac{1}{\tilde{N}_{hij}}\frac{\sum_{i=1}^{(I_h-m_h^C)} S_{hi}}{m_h^{NC}}\times\frac{\tilde{N}_{hij}}{S_{hi}}\times f_{hj} = f_{hj}$$

Non-certainty FFRDC:

$$\tilde{\pi}_{5ijk} = \pi_{5i}\times\tilde{\pi}_{5jk|i} = \left(m_5 - m_5^C\right)\frac{S_{5i}}{\sum_{i=1}^{(I_5-m_5^C)} S_{5i}}\times\frac{1}{\tilde{N}_{5ij}}\frac{\sum_{i=1}^{(I_5-m_5^C)} S_{1i}}{m_5^{NC}}\times\frac{N_{5ij}}{S_{5i}}\times f_{5j} = f_{5j}$$

The second-stage sampling will take place in rolling basis. That is, once the list of ECD is received from a sampled institution, we will draw the sample of ECD within that institution. If we keep the sampling rate f_{hj} as in formula (12) fixed during the second-stage sampling, a consequence is that the total number of sampled ECD may not be exactly 18,000; that is, the total number of sampled ECD can be less or more than 18,000 depending on the number of institutions responding to the survey. We will monitor the numbers of ECD sampled from each institution, stratum, and domain and may adjust the sample sizes (after discussion with NSF staff) if it appears that the counts of ECD on the institution lists or the number sampled will differ greatly from the sampling plan.

G. Small Institutions

Some of the GSS and FFRDC institutions are too small to support the average sample sizes called for in **Table 6** (especially after inflating by the expected ECD response and eligibility rates). RTI and NSF reviewed frame coverage and examined domain distributions when dropping the smallest institutions from the frame, and NSF determined that institutions with fewer than 50 ECD could be dropped without substantial loss of coverages. Others that are too small to support the full sample will be combined with another in the same institution strata (ideally in the same GSS stratum and state) for purposes of forming institution PSUs.

Combining institutions that are too small to support the full sample with other institutions to form PSUs would be done prior to selection of the first stage sample. Assuming that all institutions in a PSU provide lists, the minimum number of ECD needed in a PSU would be the counts shown in the last column of Table 7 inflated by the ECD eligibility and response rate (Table 5). In the first two GSS strata, we plan to combine any institution with fewer than 100 ECD with a larger institution, and in the last two GS strata we plan to combine any institutions with fewer than 75 ECD with a larger institution. If one or both of the institutions does not provide a list then we may consider adding an additional PSU in that strata as part of the wave release. In any case, if only one of the set of institutions in a PSU provides a list, then we will likely still select ECD from the cooperative institution.

H. Institutions that Cannot Identify their ECD

As in the pilot survey, not all institutions will be able to identify their ECD and will provide a list with variables such as year of degree and job type. We plan to classify individuals on these lists according to the likelihood of being an ECD ("not likely," "somewhat likely," "likely"). Those that are "not likely" to be an ECD will not be a part of the sample; we learned in the pilot survey that every few of these were actually ECD. Those that are "somewhat likely" or "likely" will be sampled, with the sampling rates set lower for those that are "somewhat likely" to be an ECD and higher for those that are "likely" to be an ECD. A higher sampling rates for those that are "likely" to be an ECD means that a larger number will be eligible and actually ECD. This will increase the design effect due to unequal weighting, but will also increase the proportion of the sampled individuals that are actually ECD and therefore eligible for the survey.