Imputation Procedures for the FY 2018 Higher Education Research and Development Survey

July 2019

prepared for

National Science Foundation
National Center for Science and Engineering Statistics

by

ICF • 530 GAITHER ROAD • ROCKVILLE, MD 20850

Contents

Introduction

This document details the procedures used for the imputation of missing values for the National Science Foundation (NSF) FY 2018 Higher Education Research and Development (HERD) Survey.

Overview

In 2010, the HERD Survey replaced NSF's survey of the R&D effort in the academic sector, the Survey of Research and Development Expenditures at Universities and Colleges (Academic R&D Expenditures Survey), which had been conducted annually since 1972. The FY 2018 survey was the ninth collection cycle completed with the redesigned survey. Questions included in the HERD Survey can be broadly divided into two groups: those similar in content to items in the Academic R&D Expenditures Survey and those that are new to the survey and were not asked of institutions prior to 2010.

Many of the data requested as part of the HERD Survey were identical to those requested by the FY 2009 Academic R&D Expenditures Survey but were included in questions that were expanded or restructured. The biggest change to most questions was the inclusion of non-science and engineering (S&E) fields in all R&D categories; most items in the Academic R&D Expenditures Survey asked for expenditures in only S&E fields. For example, Question 1 of the HERD Survey was very similar to Item 1 of the Academic R&D Expenditures Survey. Both asked for R&D expenditures by source of funds, but the FY 2010 survey asked for expenditures from S&E and non-S&E fields. The Academic R&D Expenditures Survey included one item that asked about non-S&E expenditures by field and source of funding (federal vs. nonfederal).

During the FY 2012 cycle, NSF introduced the HERD Short Form survey. This survey is sent to institutions in the HERD Survey population that reported less than $1 million in R&D expenditures in the previous fiscal year. The goal of the new instrument was to reduce the burden on smaller R&D-performing institutions that frequently had little or no expenditures in some categories. All variables in the Short Form HERD questionnaire are included in the standard HERD questionnaire. When applicable, data from both surveys were used to inform the imputation of a particular variable.

The FY 2018 survey was the seventh collection cycle completed with the inclusion of the HERD Short Form survey. Each year, there are institutions that move from the Short Form population in the previous year to the HERD standard form population in the current year. Procedures were added to address the imputation of missing HERD Survey data for an institution that completed the Short Form in the previous year. Variables that were included in both surveys were imputed using the existing methodologies. Variables that were not included in the HERD Short Form survey were imputed in one of two ways: using the most recent standard form survey data, whether FY 2011 or FY 2016, or using peer institution data. Throughout this document, we highlight how imputation procedures were altered to address missing FY 2018 Short Form data and missing FY 2018 standard form data when only the previous year's Short Form data were available. When not

specified, it should be assumed that we are referring to the imputation of data for the standard HERD Survey.

Prior to the start of the imputation process, the submitted data underwent a recoding process designed to address issues of logical imputation. Within the HERD Survey, the amount that can be reported for one question often is logically restricted by values reported for another question. For example, if Question 1, row a (federal R&D expenditures) was reported as zero and other questions asking for amounts that are a subset of federal R&D were left blank, the missing values were recoded as 0. If there were no federal expenditures reported in Question 1, federal expenditures were not imputed for any other part of the survey. During this recoding process, some values were changed to accurately reflect partial data provided by the institution. For example, respondents were asked to report total expenditures from federal sources in Questions 1, 6, and 9. For Question 1, they reported the total amount. For Question 6, they reported the amounts of federal expenditures for basic research, applied research, and experimental development, and the sum of these three values had to equal the value reported in Question 1, row a. For Question 9, respondents reported federally funded expenditures for R&D by agency and field of R&D. Again, the grand total for this question had to equal Question 1, row a. If a respondent could not report complete data for Question 6 or 9 (e.g., reported basic research expenditures but could not report applied research or experimental development), the total for the question, which was calculated automatically on the survey website, did not equal Question 1, row a. As part of the recoding effort, the value for total federal R&D expenditures for the questions with partial data was replaced with the correct value from Question 1. An additional logical imputation technique was implemented before the imputation of expenditures by field on Questions 9, 11, and 14. This is described in the procedures for Question 9.

Unless noted otherwise, the order of imputation described in this document depicts the order of imputation programming. At the end of the imputation process, all imputed data cells are flagged with an "i" in the database and in published tables.

General Procedures

Imputation techniques for variables can be broadly divided into two steps:

1. Using inflator/deflator factors to impute key variables based on the previous year's data for each nonresponding institution. Key variables are values identified as having high correlations across years and high correlations with other, smaller values within the current-year survey responses.
2. Using the relative percentages that were last reported by that institution or by peer institutions in the current year as a reference for the distribution of the key variables across detail fields. Imputed amounts were based on a mean value or mean proportion of a value within a group of institutions with similar characteristics, referred to as an imputation class.

In some circumstances, there was an intervening step. For questions for which the previous year's data could not be used as a basis for imputation, logistic regression was used to identify values that should be zero. There was a high prevalence of zero values for many variables in some questions (i.e., Questions 2, 4, 5, 6, 12, 15, 16). For these types of variables, it was efficient to first determine

whether the variable should take on a zero value before attempting to impute a nonzero value. Logistic models (SAS PROC LOGISTIC) were run for several variables. If the predicted value ($\hat{p}$) was less than 0.5, the variable in question was imputed with 0. Specific information about predictors and class variables is included in the descriptions below.

Because much current-year imputation is based on an institution's R&D expenditures from the previous year, alternative procedures were adopted for institutions that did not have FY 2017 data. Three short form institutions did not submit data for the FY 2018 survey and had no FY 2017 HERD Survey data. For these institutions, total R&D expenditures were set at the baseline for the short form survey ($150,000).

Questions 1.1, 10, and 13 were not imputed.

Determining Imputation Factors

The imputation process involves first determining imputation factors for certain key variables. Imputation factors are the ratio of current-year data to previous-year data for institutions that responded in both years (i.e., matched, clean data). These factors, when applied to institutions in a predefined group, reflect the average annual growth or decline in expenditures for reporting institutions in that group.

Imputation factors were derived for different groups of institutions based on the highest degree offered (HDO) and type of control (TOC). Factors were calculated separately for each key variable for each combination of HDO (PhD or no PhD) and TOC (public or private). These combinations are referred to as imputation classes.

All institutions in both the short form and standard form populations, including those that reported less than $150,000 in total R&D, could contribute to the imputation factors. Table 1 shows the number of institutions from the FY 2018 survey in each imputation class, including those that did not have matched, clean data for total R&D expenditures and were not used to derive imputation factors.

Table 1. Number of Institutions in the Population by Highest Degree Offered and Type of Control

HDO	TOC	
	Public	Private
PhD	354	204
No PhD	185	206

The imputation classes were further divided based on quartiles of total R&D expenditures within each class for some questions. This is noted in the description of each question.

Imputing Key Variables

Key variables are values identified as having high correlations across years and high correlations with other, smaller values within the current-year survey responses. Specific key variables are discussed, as applicable, for each survey question. All key variables were imputed for unit nonresponders; only missing key variables were imputed for partially nonrespondent institutions. The imputation technique used to calculate key variables is called ratio imputation and takes the following mathematical form:

Equation 1a: $$\widehat{\boldsymbol{y}}_{\boldsymbol{ik_t}} = \widehat{\boldsymbol{B}}_{\boldsymbol{k_t}} \boldsymbol{y}_{\boldsymbol{ik_{t-1}}}$$

where $\hat{y}_{ik_t}$ is the imputed value of key variable y_k for institution *i* for year *t,* and $\hat{B}_{k_t}$ is the inflator/deflator factor for key variable y_k, defined as

Equation 1b: $$\widehat{\boldsymbol{B}}_{\boldsymbol{k_t}} = \frac{\sum_{j=1}^{r} \boldsymbol{y}_{\boldsymbol{jk_t}}}{\sum_{j=1}^{r} \boldsymbol{y}_{\boldsymbol{jk_{t-1}}}}$$

where $y_{jk_{t-1}}$ is the value of key variable y_k for institution *j* for year *t*-1, and *r* is the set of institutions in the same degree level and institutional control peer group as institution *i* that provided key variable y_k both in years *t* and *t*-1.

If a key variable was imputed in the previous year, the factor was applied to the imputed value to derive the current year's value.

In some cases, the specific key variable from the past year was missing and not imputed. In these situations, a ratio of the missing key variable to a non-missing key variable for peer institutions that provided both values was used:

Equation 2a: $$\widehat{\boldsymbol{y}}_{\boldsymbol{il_t}} = \widehat{\boldsymbol{R}}_{\boldsymbol{lk_t}} \, \boldsymbol{y}_{\boldsymbol{ik_t}}$$

where $\hat{y}_{il_t}$ is the imputed value of key variable y_l for institution *i* for year *t,* and $\hat{R}_{lk_t}$ is the ratio of key variables y_l to y_k , defined as

Equation 2b: $$\widehat{\boldsymbol{R}}_{\boldsymbol{lk_t}} = \frac{\sum_{j=1}^{r} \boldsymbol{y}_{\boldsymbol{jl_t}}}{\sum_{j=1}^{r} \boldsymbol{y}_{\boldsymbol{jk_t}}}$$

where y_{jl_t} is the value of key variable y_l for institution *j* for year *t,* and where *r* is the set of institutions in the same imputation class as institution *i* that provided key variable y_l and y_k both in years *t.*

In the example where there is no previous year's value for R&D equipment expenditures, the imputed value would be the product of total R&D expenditures (imputed or reported) and the ratio of R&D equipment expenditures to total R&D expenditures for the imputation class.

Imputing Non-Key Variables

The ratio imputation technique described above was used to impute key variables. However, many HERD Survey variables are hierarchical, and each key variable has a number of lower-level, non-key detail variables associated with it. For example, the key variable Federally Funded R&D Expenditures has 326 lower-level, non-key variables associated with it in the standard form survey, such as federally funded R&D expenditures in astronomy, R&D expenditures funded by the U.S. Department of Health and Human Services (HHS), and R&D expenditures in chemistry funded by NSF. For nonresponding institutions, key variables (imputed or reported) were distributed across the associated non-key variables using the same relative percentages that were last reported by that institution. If some non-key fields were reported, the difference between the key variable and the reported non-key fields was distributed to the missing detailed fields using the same relative percentages last reported by that institution.

Non-key variables were derived from their associated key variables or higher-level, non-key variable using the following relation:

Equation 3:
$$\widehat{\boldsymbol{y}}_{\boldsymbol{in_t}} = \widehat{\boldsymbol{y}}_{\boldsymbol{ik_t}} \left(\frac{y_{in_{t-1}}}{y_{ik_{t-1}}} \right)$$

where $\hat{y}_{in_t}$ is the imputed value of non-key variable y_n for institution *i* for year *t,*
$\hat{y}_{ik_t}$ is the imputed value of key variable y_k for institution *i* for year *t,*
$y_{in_{t-1}}$ is the value of non-key variable y_n for institution *i* for year *t*-1, and
$y_{ik_{t-1}}$ is the value of key variable y_k for institution *i* for year *t*-1.

This was the same non-key variable imputation approach used for both unit nonresponders and those institutions that did not respond to individual non-key items. For example, if an institution reported federal R&D expenditures but did not provide the breakdown of those expenditures by field of study, the non-key values were imputed the same way; however, rather than using the imputed value of the key variable ($\hat{y}_{ik_t}$), the reported value of y_k was used.

If lower-level, non-key data were not available for a particular institution for the previous cycle, the key variables were distributed across detail fields based on the relative percentages for the institution's class. Non-key variables were derived from their associated key variables using the following relation:

Equation 4a:
$$\widehat{\boldsymbol{y}}_{\boldsymbol{in_t}} = \widehat{\boldsymbol{R}}_{\boldsymbol{nk_t}} \, \boldsymbol{y}_{\boldsymbol{ik_t}}$$

where $\hat{y}_{in_t}$ is the imputed value of non-key variable y_n for institution *i* for year *t,*
and $\hat{R}_{nk_t}$ is the ratio of non-key variables y_n to y_k defined as

Equation 4b:
$$\widehat{\boldsymbol{R}}_{\boldsymbol{nk_t}} = \frac{\sum_{j=1}^{r} y_{jn_t}}{\sum_{j=1}^{r} y_{jk_t}}$$

where y_{jn_t} is the value of key variable y_n for institution j for year t, and where r is the set of institutions in the same imputation class as institution i that provided variables y_n and y_k both in years t.

Procedures by Survey Question

Expenditures by Source of Funds (Question 1)

The imputation of missing values in Question 1 was completed only for unit nonresponders, which were defined as institutions in the population that did not report any data for FY 2018. The imputation of values for individual missing fields would necessarily impact the total R&D reported by the institution for Question 1, and it was decided that the total R&D reported by an institution would not be altered through imputation.

Question 1 Key Variables

There were two key variables imputed for Question 1: Federal R&D Expenditures and Total R&D Expenditures. Imputation factors for both key variables for each imputation class are listed in tables 2 and 3.

Table 2. Imputation Factors for Federal Expenditures by Class

HDO/TOC	n	Federal R&D
PhD		
Public	341	1.0446
Private	182	1.0394
No PhD		
Public	144	0.9314
Private	178	0.9613

n = number of institutions used to create the factor

Table 3. Imputation Factors for Total Expenditures by Class

HDO/TOC	n	Total R&D
PhD		
Public	341	1.0580
Private	182	1.0492
No PhD		
Public	144	0.9484
Private	178	0.9596

n = number of institutions used to create the factor

If an institution was missing a key variable from the previous year and that value was not imputed, the current-year value was based on the proportion for peer institutions of that key variable to a known value:

Equation 5: $$\widehat{y}_{il_t} = y_{ik_t} \frac{\sum_{j=1}^{r} y_{jl_t}}{\sum_{j=1}^{r} y_{jk_t}}$$

where $\hat{y}_{il_t}$ is the imputed value of federal or total R&D for institution *i* for year *t,* y_{ik_t} is the value of federal or total R&D for institution *i* for year *t,* *r* is the set of institutions in the same imputation class as institution *i* that provided variables y_l and y_k both in years *t.*

Question 1 Non-Key Variables

There are three hierarchical steps for the imputation of non-key variables in Question 1:

1. **Nonfederal R&D:** Total R&D minus Federal R&D
2. **Nonfederal Sources:** Nonfederal R&D expenditures were distributed across the associated nonfederal source variables (i.e., state and local government, business, nonprofit, institutional, and other) using the same relative percentages that were last reported by that institution.
3. **Institutional Sources:** The imputed value of institutionally funded expenditures was distributed across the three types of institution funds (institutionally financed organized research, cost sharing, and unrecovered indirect costs) using the same relative percentages that were last reported by that institution.

For each step in the imputation process, if the imputed details did not add to the total, the details were adjusted by adding 1 progressively until they totaled correctly. On the rare occasion that the sum of the details was more than the reported total, the analyst reduced the amount reported for the details by 1 until the values were equal. This same process was implemented for each stage of imputation of non-key variables for every question.

If a value in Question 1 from FY 2017 was missing and not imputed, which would happen only if the institution partially responded to Question 1 in 2017, it was considered unavailable in FY 2018. The other option was to impute as zero, but we consider that a misrepresentation of the previous year's data, which form the basis of current-year imputation.

Tables 4 and 5 provide summary data on imputed amounts and rates for imputation class and each Question 1 variable.

Table 4. Imputed and Aggregate Amounts for Total and Federal R&D by Class
(amounts are dollars in thousands)

HDO/TOC	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
PhD								
Public	4	41,516	25,703,791	0.16%	4	53,572	51,743,972	0.10%
Private	6	26,659	15,903,159	0.17%	6	39,205	26,922,442	0.15%
No PhD								
Public	7	16,784	263,624	6.37%	7	33,502	457,922	7.32%
Private	11	2,599	148,841	1.75%	11	7,074	314,658	2.25%

n = number of institutions with imputed values

Table 5. Imputed and Aggregate Amounts for Sources of Funds
(amounts are dollars in thousands)

Funding Source	n	Imputed	Total	% Imputed
Federal	28	87,558	42,019,415	0.21%
State/Local	28	8,302	4,321,480	0.19%
Business	28	4,375	4,723,897	0.09%
Nonprofit	28	6,730	5,452,898	0.12%
All Inst Funds	27	22,315	20,438,289	0.11%
Inst Financed Research	28	12,163	13,310,779	0.09%
Cost Sharing	28	1,350	1,589,047	0.08%
Unrecovered	26	8,802	5,538,463	0.16%
Other	28	4,073	2,483,015	0.16%
Total	28	133,353	79,438,994	0.17%

n = number of institutions with imputed values

Federal Expenditures by Field of R&D and Agency (HERD Question 9 and Short Form Question 2, Column 1)

As with Question 1, if an institution reported partial data for Question 9 of the HERD Survey or Question 2, column 1 of the HERD Short Form, and if the imputation of missing data would necessarily impact the federal R&D expenditures reported by the institution, it was decided that the federal R&D amount would not be altered and values would not be imputed for Question 9 for that institution. However, in most cases where some values in Question 9 were missing, all federal expenditures were reported on the survey, but the institution could not provide the level of detail required. For example, some institutions entered all engineering under Other Engineering and indicated that they could not break out these expenditures across the many detailed fields of engineering requested on the survey. In cases such as these, missing values were imputed.

Question 9 Key Variables

The Federal R&D Expenditures key variable was already imputed during the imputation process for Question 1.

Question 9 Non-Key Variables

For institutions where all or some of the information for Question 9 (or Question 2, column 1 of the short form) was missing and there were no reported past year's data for the missing values to refer to, an additional logical imputation technique was employed before proceeding with the imputation of non-key variables. Data collection staff reviewed the websites of institutions to determine which fields of R&D should be imputed as zero. The assumption was that if there were no degrees granted in an area or related area, no R&D was likely being performed. This approach was thought to be better than imputation based solely on imputation class, which typically resulted in expenditures being imputed in every field. For example, based solely on imputation class, a liberal arts college that specializes in social sciences and non-S&E programs would have expenditures imputed in engineering. By reviewing institution websites, we could avoid some of these obvious issues. The same imputation logic was applied to Question 11

Non-key variables in Question 9 were imputed in three hierarchical steps (see below). For HERD Short Form institutions, imputation for Question 2, column 1 ended after the first step. In each step, the target value was computed based on the ratio of the lower-level variable to the higher-level variable in the previous year's survey.

1. **Major Fields of R&D (e.g., engineering, physical sciences, life sciences):** Referring to equation 3, the total for each major field was a y_{in} variable, and Federal R&D Expenditures was the y_{ik} variable.
2. **Minor Fields of R&D (e.g., health sciences, economics, chemical engineering):** The detailed fields of R&D that contribute to subtotals were y_{in} variables. The major fields of R&D that are broken down into more detailed fields were y_{ik} variables. For institutions that were in the standard form population in FY 2018 but were in the short form population in FY 2017 and had standard form data for any years prior to FY 2017, the most recent data were distributed across detailed fields.
3. **Expenditures by Agency (e.g., NSF-funded expenditures in chemical engineering, HHS-funded expenditures in health sciences):** Each agency by lowest level of R&D field variable was a y_{in} variable, and the total federal expenditures for the corresponding fields were y_{ik} variables.

Detailed data were summed to provide the major field by agency total when major field subtotals by agency were needed.

If the past year's data were not available, key variables were distributed across associated non-key variables using the relative percentages reported by institutions in the same imputation class (equations 4a and 4b). If this was the case for major fields, standard form and short form institutions were used to derive relative percentages per class. Table 6 lists the imputed amount for federal R&D in each field and includes amounts for both the short form and the standard form. For this reason, the *n* for major fields is larger than for detailed fields.

Table 6. Imputed and Aggregate Amounts for Federal Expenditures by Field
(amounts are dollars in thousands)

Field of R&D	n	Imputed	Total	% Imputed
Computer and Information Sciences	34	2,335	1,635,198	0.14%
Engineering	34	7,763	7,099,651	0.11%
Aerospace, Aeronautical, and Astronautical	20	8	678,087	0.00%
Bioengineering and Biomedical	20	346	787,000	0.04%
Chemical	20	185	461,674	0.04%
Civil	20	890	592,396	0.15%
Electrical, Electronic, Communications	20	2,297	1,981,799	0.12%
Industrial and Manufacturing	20	5	306,696	0.00%
Mechanical	20	1,199	993,683	0.12%
Metallurgical and Materials	20	513	464,981	0.11%
Other	20	2,151	826,975	0.26%
Geosciences, Atmospheric, and Ocean sciences	34	423	2,054,549	0.02%
Atmospheric Sciences and Meteorology	20	192	485,737	0.04%
Geological and Earth Sciences	20	175	699,305	0.03%
Ocean Sciences and Marine Sciences	20	0	648,156	0.00%
Other	20	0	215,383	0.00%
Life Sciences	34	74,637	23,978,544	0.31%
Agricultural Sciences	20	9,820	956,060	1.03%
Biological and Biomedical Sciences	20	30,676	8,589,048	0.36%
Health Sciences	20	25,742	13,453,146	0.19%
Natural Resources and Conservation	21	2,684	314,948	0.85%
Other	21	29,487	633,813	4.65%
Mathematics and Statistics	34	2,985	459,454	0.65%
Physical Sciences	34	6,241	3,483,381	0.18%
Astronomy and Astrophysics	20	149	454,394	0.03%
Chemistry	20	2,011	1,136,317	0.18%
Materials Science	20	0	162,048	0.00%
Physics	20	2,805	1,564,866	0.18%
Other	20	835	154,035	0.54%
Psychology	34	1,135	764,434	0.15%
Social Sciences	34	7,310	947,919	0.77%
Anthropology	20	151	43,407	0.35%
Economics	20	87	102,053	0.09%
Political science and Government	20	129	98,089	0.13%
Sociology, Demography, and Population Studies	20	380	285,375	0.13%

Field of R&D	n	Imputed	Total	% Imputed
Other	20	6499	416,554	1.56%
Other Sciences	34	203	350,624	0.06%
Non-S&E Fields	34	3,642	1,245,661	0.29%
Business Management and Business Administration	20	412	71,706	0.57%
Communication and Communications Technologies	20	34	34,279	0.10%
Education	20	1,057	673,774	0.16%
Humanities	20	68	49,997	0.14%
Law	20	258	51,134	0.50%
Social work	20	107	114,073	0.09%
Visual and Performing Arts	20	16	10,829	0.15%
Other	20	1,431	234,242	0.61%

n = number of institutions with imputed values

Table 7 lists the imputed amount of federal R&D for each agency. Federal expenditures by agency are not collected on the short form; therefore, these amounts are for the standard form only.

Table 7. Imputed and Aggregate Amounts for Federal Expenditures by Agency
(amounts are dollars in thousands)

Agency	n	Imputed	Total	% Imputed
USDA	20	20,005	1,185,986	1.69%
DoD	20	8,240	5,900,829	0.14%
Energy	20	2,661	1,819,663	0.15%
HHS	20	51,987	22,922,192	0.23%
NASA	20	2,898	1,516,983	0.19%
NSF	20	11,310	5,273,511	0.21%
Other	20	6,635	3,325,918	0.20%

n = number of institutions with imputed values

Nonfederal Expenditures by Field of R&D and Source of Funds (HERD Question 11 and Short Form Question 2, Column 2)

Question 11 Key Variables

The key variable Nonfederal R&D Expenditures was already imputed during the imputation process for Question 1.

Question 11 Non-Key Variables

Non-key variables in Question 11 were imputed in three hierarchical steps (see below). For HERD Short Form institutions, imputation for Question 2, column 2 ended after the first step. In each step, the target value was computed based on the ratio of the lower-level variable to the higher-level variable in the previous year's survey.

1. **Major Fields of R&D (e.g., engineering, physical sciences, life sciences):** Referring to equation 3, the total for each major field was a y_{in} variable, and Nonfederal R&D Expenditures was the y_{ik} variable.
2. **Minor Fields of R&D (e.g., health sciences, economics, chemical engineering):** The detailed fields of R&D that contribute to subtotals were y_{in} variables. The major fields of R&D that were broken down into more detailed fields were y_{ik} variables. For institutions that were in the standard form population in FY 2018 but were in the short form population in FY 2017 and had standard form data for any years prior to FY 2017, the most recent data were distributed across detailed fields.
3. **Expenditures by Source (e.g., expenditures in chemical engineering sponsored by businesses, expenditures in health sciences funded by institutional funds):** Because total R&D funded by different nonfederal sources was already imputed for Question 1, there was no need to reference past-year or peer data to impute values for source by field cells. Each value was imputed as follows:

Q12rowXcolumnY = (column Y total / Total Nonfederal) * row X total

If the amount for a nonfederal source was missing in Question 1 and was not imputed because it would alter the reported total R&D expenditures, expenditures for R&D fields funded by that source also remained missing and un-imputed. Table 8 lists the imputed amount for nonfederal R&D in each field and includes amounts for both the short form and the standard form. For this reason, the *n* for major fields is larger than for detailed fields.

Table 8. Imputed and Aggregate Amounts for Nonfederal Expenditures by Field
(amounts are dollars in thousands)

Field of R&D	n	Imputed	Total	% Imputed
Computer and Information Sciences	41	21,805	772,598	2.82%
Engineering	40	70,767	5,287,268	1.34%
Aerospace, Aeronautical, and Astronautical	25	5,574	333,724	1.67%
Bioengineering and Biomedical	25	2,684	552,652	0.49%
Chemical	25	9,611	471,849	2.04%
Civil	25	4,017	767,778	0.52%
Electrical, Electronic, Communications	25	22,000	864,801	2.54%
Industrial and Manufacturing	25	3,281	208,338	1.57%
Mechanical	25	11,001	635,666	1.73%
Metallurgical and Materials	25	3,943	298,686	1.32%
Other	25	8,364	1,147,058	0.73%
Geosciences, Atmospheric, and Ocean sciences	41	8,819	1,117,398	0.79%
Atmospheric Sciences and Meteorology	25	3,972	122,342	3.25%
Geological and Earth Sciences	25	3,839	435,123	0.88%
Ocean Sciences and Marine Sciences	25	540	410,500	0.13%
Other	25	374	145,094	0.26%

Field of R&D	n	Imputed	Total	% Imputed
Life Sciences	41	244,822	21,922,366	1.12%
Agricultural Sciences	25	3,242	2,364,887	0.14%
Biological and Biomedical Sciences	25	36,108	5,965,052	0.61%
Health Sciences	25	190,900	12,485,116	1.53%
Natural Resources and Conservation	26	11,910	454,428	2.62%
Other	26	8,740	627,047	1.39%
Mathematics and Statistics	41	6,144	298,315	2.06%
Physical Sciences	41	24,025	1,773,027	1.36%
Astronomy and Astrophysics	25	4,997	212,709	2.35%
Chemistry	25	8,030	739,811	1.09%
Materials Science	25	0	93,882	0.00%
Physics	25	7,482	638,866	1.17%
Other	25	2,988	80,628	3.71%
Psychology	41	6,279	503,099	1.25%
Social Sciences	41	10,926	1,807,875	0.60%
Anthropology	25	424	77,605	0.55%
Economics	25	463	362,924	0.13%
Political science and Government	25	757	345,083	0.22%
Sociology, Demography, and Population Studies	25	6,674	321,908	2.07%
Other	25	2,215	693,567	0.32%
Other Sciences	41	1,307	540,229	0.24%
Non-S&E Fields	41	24,311	3,397,404	0.72%
Business Management and Business Administration	25	2,454	714,350	0.34%
Communication and Communications Technologies	25	1,087	137,035	0.79%
Education	25	2,045	813,614	0.25%
Humanities	25	1,638	463,088	0.35%
Law	25	1,071	217,196	0.49%
Social work	25	746	137,294	0.54%
Visual and Performing Arts	25	741	126,535	0.59%
Other	25	13,464	769,311	1.75%

n = number of institutions with imputed values

Equipment Expenditures by Field of R&D (Question 14)

Question 14 Key Variables

The Total R&D Equipment key variable was calculated in the same way as other key variables (equations 1a and 1b). The imputation factors for each class are listed in table 9. If there was no

value for Total R&D Equipment in the previous year, a ratio imputation technique was used (equations 2a and 2b). This was the procedure for institutions that were in the FY 2018 standard form population but had been in the FY 2017 short form population.

Table 9. Imputation Factors for Total Equipment Expenditures by Class

HDO/TOC	n	Total Equipment
PhD		
Public	307	0.9804
Private	132	0.9744
No PhD		
Public	70	0.8370
Private	81	0.9000

n = number of institutions used to create the factor

Question 14 Non-Key Variables

Non-key variables in Question 14 were imputed in three hierarchical steps:

1. **Federal and Nonfederal:** Total equipment expenditures were distributed based on the ratio of the current year's total federal to total nonfederal expenditures.
2. **Major Fields of R&D (e.g., engineering, physical sciences, life sciences, education):** Again, the ratios of field to total for federal expenditures or nonfederal expenditures were used to distribute equipment expenditures by major field.
3. **Minor Fields of R&D (e.g., health sciences, economics, chemical engineering):** The same process was used as for imputing major fields.

Table 10 provides summary data on imputed amounts and rates for each field of study included in Question 14.

Table 10. Imputed and Aggregate Amounts for Equipment Expenditures by Field
(amounts are dollars in thousands)

Field of R&D	n	Imputed	Total	% Imputed
Computer and Information Sciences	27	7,209	89,849	8.02%
Engineering	27	36,627	594,041	6.17%
Aerospace, Aeronautical, and Astronautical	27	4,140	31,210	13.26%
Bioengineering and Biomedical	27	713	67,052	1.06%
Chemical	27	332	42,024	0.79%
Civil	27	248	32,581	0.76%
Electrical, Electronic, Communications	27	19,158	125,712	15.24%
Industrial and Manufacturing	27	3,178	23,557	13.49%
Mechanical	27	5,943	85,668	6.94%
Metallurgical and Materials	27	510	51,129	1.00%
Other	27	2,405	135,108	1.78%
Geosciences, Atmospheric, and Ocean sciences	27	1,636	95,496	1.71%
Atmospheric Sciences and Meteorology	27	430	17,113	2.51%
Geological and Earth Sciences	27	460	39,840	1.15%

Ocean Sciences and Marine Sciences	27	451	33,037	1.37%
Other	27	295	5,506	5.36%
Life Sciences	27	8,356	874,888	0.96%
Agricultural Sciences	27	132	79,387	0.17%
Biological and Biomedical Sciences	27	6,340	394,038	1.61%
Health Sciences	27	1,783	357,822	0.50%
Natural Resources and Conservation	27	28	13,783	0.20%
Other	27	73	29,858	0.24%
Mathematics and Statistics	27	4,022	9,342	43.05%
Physical Sciences	27	10,389	383,982	2.71%
Astronomy and Astrophysics	27	1674	31,098	5.38%
Chemistry	27	762	120,695	0.63%
Materials Science	27	0	17,263	0.00%
Physics	27	5,308	191,497	2.77%
Other	27	2645	23,429	11.29%
Psychology	27	67	16,325	0.41%
Social Sciences	27	360	13,059	2.76%
Anthropology	27	2	1,886	0.11%
Economics	27	18	3,443	0.52%
Political science and Government	27	52	905	5.75%
Sociology, Demography, and Population Studies	27	196	1,302	15.05%
Other	27	92	5,523	1.67%
Other Sciences	27	1416	26,892	5.27%
Non-S&E Fields	27	578	41,721	1.39%
Business Management and Business Administration	27	59	5,962	0.99%
Communication and Communications Technologies	27	4	4,020	0.10%
Education	27	251	6,448	3.89%
Humanities	27	55	6,263	0.88%
Law	27	34	326	10.43%
Social work	27	8	217	3.69%
Visual and Performing Arts	27	6	1,195	0.50%
Other	27	161	17,290	0.93%

n = number of institutions with imputed values

Funds Received as a Subrecipient (HERD Question 7 and Short Form Question 3)

Question 7 Key Variables

Because of the inclusion of the short form survey, which requests subrecipient funds received only from higher education entities, it was necessary to have two key variables (i.e., Sub From Higher Education and Sub From Non-Higher Education). Institutions from the short form and long form

populations were used to calculate imputation factors for Sub From Higher Education, but only standard form institutions were included in the calculation of Sub From Non-Higher Education. In FY 2018 one institution, Roger Williams University, reported a $829,000 decrease in their total R&D expenditures received from non-higher education pass through entities. This change was unusually high compared to other institutions within the same imputation class (NoPhD/Private) who reported changes between $1,000 and $169,000 in the non-higher education data element. The inclusion of Roger Williams University in the calculation of the factor for Received from Non-Higher Education would have resulted in an unusually low number and so it was decided that the institution should be excluded as an outlier from that calculation. A similar decrease was not reported for the Higher Education factor and so the institution was included in that calculation. If there was no value for either key variable in the previous year, a ratio imputation technique was used (equations 2a and 2b). The imputation factors for each class and key variable are listed in table 11.

Table 11. Imputation Factors for Total Subrecipient Expenditures by Class

HDO/TOC	n	Sub From Higher Education	n	Sub From Non-Higher Education
PhD				
Public	297	1.0496	263	1.0460
Private	171	1.0899	126	1.0073
No PhD				
Public	142	0.9751	69	0.8337
Private	175	1.0050	79	0.8495

n = number of institutions used to create the factor

Question 7 Non-Key Variables

Sub From Higher Education was imputed in one hierarchical step, so step 1 below was the only step that applied to both short form and long form institutions. Sub From Non-Higher Education was imputed in two hierarchical steps:

1. **Source of Funds (federal or nonfederal)**
2. **Other Pass-Through Institutions:** Standard form institutions were asked to divide non-higher education pass-through sources into business, nonprofit, and other. If they were unable to report the non-higher education sources at this level of detail, they were asked to classify all expenditures as other and indicate that amounts from business and nonprofit sources were unavailable.

Distribution across categories was based on last year's response (equation 3) unless last year's data were missing, in which case distribution was based on current-year peer institutions (equations 4a and 4b).

Tables 12 and 13 provide summary data on federal and total imputed amounts and rates by imputation class and pass-through entity. Short form and standard form institutions are included in the summaries for table 13 but only standard form institutions are included in the summaries for Table 12.

Table 12. Imputed and Aggregate Amounts for Total and Federal R&D Received as a Subrecipient by Class
(amounts are dollars in thousands)

HDO/TOC	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
PhD								
Public	9	42,462	4,175,889	1.02%	45	956,430	5,078,949	18.83%
Private	9	20,836	1,998,498	1.04%	16	245,197	2,287,378	10.72%
No PhD								
Public	4	166	30,875	0.54%	5	1,645	36,855	4.46%
Private	5	418	22,024	1.90%	5	480	23,313	2.06%

n = number of institutions with imputed values

Table 13. Imputed and Aggregate Amounts for Total and Federal R&D Received as a Subrecipient by Pass-Through Entity
(amounts are dollars in thousands)

Pass-Through Entity	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
Higher Ed	41	21,938	3,192,875	0.69%	82	431,591	3,559,095	12.13%
Business	30	12,681	916,451	1.38%	72	202,871	1,180,215	17.19%
Nonprofit	30	15,200	1,182,214	1.29%	72	249,918	1,533,277	16.30%
Other	30	14,607	951,133	1.54%	72	195,633	1,170,563	16.71%

n = number of institutions with imputed values

Expenditures Passed Through to Other Institutions (HERD Question 8 and Short Form Question 4)

Question 8 Key Variables

Because of the inclusion of the short form survey, which requests subrecipient funds passed through only to higher education entities, it was necessary to have two key variables (i.e., Passed to Higher Education and Passed to Non-Higher Education). Institutions from the short form and standard form populations were used to calculate imputation factors for Passed to Higher Education, but only standard form institutions were included in the calculation of Passed to Non-Higher Education. In FY 2018 three institutions, California Polytechnic State University, San Luis Obispo, CUNY, Queens College, and CUNY, John Jay College of Criminal Justice, reported increases over $1,000,000 in their total R&D expenditures passed through to higher education pass through entities. This change was unusually high compared to other institutions within the same imputation class (NoPhD/Public) who reported changes between $1,000 and $522,000 in the

higher education data element. The inclusion of these institutions in the calculation of the factor for Passed to Higher Education would have resulted in an unusually high number and so it was decided that the institution should be excluded as an outlier from that calculation. Similiarly, CUNY, Queens College and two other institutions reported large decreases in their total R&D expenditures passed through to non-higher education pass through entities and were excluded as an outlier from that calculation. CUNY, Queens C. and Humboldt State University reported decreases over $1,000,000 while other institutions in the same imputation class (NoPhD/Public) reported changes between $1,000 and 201,000. Charles R. Drew University of Medicine and Science reported a decrease over $1,000,000 while other institutions in that imputation class (NoPhD/Private) reported changes between $1,000 and $408,000. If there was no value for either key variable in the previous year, a ratio imputation technique was used (equations 2a and 2b). The imputation factors for each class and key variable are listed in table 14.

The Total Pass-Through variable was reported in Question 12 as well as Question 8 on the standard form, and it was possible for the variable to be missing in one of the questions but reported in the other. There were three scenarios related to the imputation of Total Pass-Through:

1. If all variables in Question 8 were missing but Total Pass-Through was reported in Question 12, the pass-through value reported in Question 12 was used to impute detail values for Question 8.
2. If Total Pass-Through was missing in both questions but some partial data were included in Question 12, the variable was not imputed for either question and was left missing. The total value in Question 12 was Total R&D Expenditures, and it equated to the total in Question 1. As with Question 1, imputing an individual missing value in Question 12 would necessarily alter the value for Total R&D Expenditures reported by the institution.
3. When all Question 8 and Question 12 values were missing, the key variables Passed to Higher Education and Passed to Non-Higher Education were calculated with Total Pass-Through calculated as the sum of the two.

Table 14. Imputation Factors for Total Pass-Through Expenditures by Class

HDO/TOC	n	Passed to Higher Education	n	Passed to Non-Higher Education
PhD				
Public	317	1.0633	279	1.1136
Private	175	1.0471	132	1.0850
No PhD				
Public	139	1.0630	66	1.0567
Private	174	0.9251	79	0.9328

n = number of institutions used to create the factor

Question 8 Non-Key Variables

Passed to Higher Education was imputed in one hierarchical step, so step 1 was the only step that applied to both short form and standard form institutions. Passed to Non-Higher Education was imputed in two hierarchical steps:

1. **Source of Funds (federal or nonfederal)**
2. **Other Subrecipient Institutions:** Standard form institutions were asked to divide non-higher education pass-through into business, nonprofit, and other. If they were unable to report the non-higher education recipients at this level of detail, they were asked to classify all expenditures as other and indicate that amounts from business and nonprofit sources were unavailable.

Distribution across categories was based on last year's response (equation 3) unless last year's data were missing, in which case distribution was based on current-year peer institutions (equations 4a and 4b).

Tables 15 and 16 provide summary data on federal and total imputed amounts and rates by imputation class and subrecipient entity. Short form and standard form institutions are included in the summaries for Table 16, but only standard form institutions are included in the summaries for Table 15.

Table 15. Imputed and Aggregate Amounts for Total and Federal R&D Passed Through to a Subrecipient by Class
(amounts are dollars in thousands)

HDO/TOC	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
PhD								
Public	7	3,134	3,310,924	0.09%	7	3,354	3,919,750	0.09%
Private	9	17,466	1,956,436	0.89%	8	17,779	2,394,249	0.74%
No PhD								
Public	4	1254	15,380	8.15%	4	1368	19,076	7.17%
Private	4	83	9,967	0.83%	4	90	10,930	0.82%

n = number of institutions with imputed values

Table 16. Imputed and Aggregate Amounts for Total and Federal R&D Passed Through by Subrecipient Entity
(amounts are dollars in thousands)

Subrecipient Entity	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
Higher Ed	59	428,308	3,084,474	13.89%	59	526,242	3,540,969	14.86%
Business	52	144,520	825,140	17.51%	52	184,696	1,059,764	17.43%
Nonprofit	52	116,593	914,798	12.75%	52	150,423	1,102,866	13.64%
Other	52	93,694	470,707	19.90%	52	130,045	643,297	20.22%

n = number of institutions with imputed values

Foreign Funding for R&D (Question 2)

Prior to FY 2016 Question 2 had only one value to impute: Total R&D Funded by Foreign Sources (foreign_tot). In FY 2016 new variables were added to this question: Foreign Funding Received From Foreign Governments (foreign_gov), Foreign Funding Received from Foreign Businesses (foreign_bus), Foreign Funding Received from Foreign Nonprofit Organizations (foreign_np), Foreign Funding Received from Foreign Higher Education Institutions (foreign_ed), and Foreign Funding Received from Other Foreign Sources (foreign_oth).

Imputation of the Question 2 total expenditure value was performed first using the same methodology that has been applied in previous years and is described below. Imputation of the source categories was performed next based on last year's response, unless last year's data were missing, in which case distribution was based on current year peer institutions.

Total Foreign Funding

By definition, total expenditures from foreign sources must be equal to or less than the total expenditures from external, nongovernmental sources as reported in Question 1 (i.e., business sources + nonprofit sources + other sources). For the purposes of this calculation, external, nongovernmental funding is referred to as *T*. The value *T* was calculated during the recoding process prior to other imputation. If *T* was 0 or missing, Question 2 was imputed as 0.

If Question 2 was reported last year, this year's value was calculated by applying the same proportion reported last year (foreign_tot / T) to this year's reported or imputed value of *T*. For institutions that moved from the short form population to the standard form population, the proportion from the most recent standard form data (FYs 2011–16) was used, if reported.

If there were no reported data from last year, a logistic regression model was employed to identify cases in which Question 2 should be imputed as zero. PROC LOGISTIC was run separately for public and private institutions using the following predictors: the continuous variable *T,* HDO, and MedS. MedS is a variable indicating the inclusion of a medical school, derived from Question 4. If the predicted value ($\hat{p}$) was less than 0.5, the value for Question 2 was imputed as 0.

The next step was the imputation of the nonzero values for foreign-funded expenditures. For this step, the mean proportion of *T* ($\bar{p}$ = foreign_tot / T) was calculated for the nonzero values in imputation classes determined by TOC, HDO, and the quartiles of Total R&D Expenditures. The imputed value of Question 2 was then calculated as T * $\bar{p}$.

Foreign Funding by Source

If total foreign funded expenditures was imputed as zero in the first step than all sources were imputed as zero as well. The next step was the imputation of cases where nonzero data were reported or imputed for the total value.

Similar to the value for the total, expenditures from foreign businesses must be equal to or less than the total expenditures reported from businesses in Question 1(source_bus), expenditures from foreign nonprofit organizations must be equal to or less than the total expenditures reported from

total nonprofit organizations in Question 1 (source_np), and the total of expenditures from foreign governments, foreign higher education, and other foreign sources must be equal to or less than the total expenditures from all other sources in Question 1 (source_oth). This required the use of Question 1 variables when calculating proportions and means rather than using a simple ratio of each foreign source to the overall total. To accomplish this for those institutions that reported this distribution last year, last year's proportion of each source to the corresponding Question 1 source was calculated. For those where last year's distribution was not reported, the mean proportion of each source to the corresponding Question 1 source was calculated in the same imputation classes used for total foreign expenditures. For expenditures from foreign businesses and foreign nonprofit organizations the proportion was applied to the institution's corresponding Question 1 data: source_bus ($\bar{p}$= foreign_bus/source_bus) and source_np ($\bar{p}$= foreign_np/source_np)

A multiple step approach had to be used for the three foreign sources reported under all other sources in Question 1 (foreign_gov, foreign_ed, and foreign_oth). For the purposes of this calculation the sum of those three foreign sources is referred to as O.

For those institutions where last year's distribution was reported:
1. Last year's proportion of O to source_oth was calculated (O/source_oth)
2. Last year's proportion of each of those foreign sources to O was calculated:
    - foreign_gov/O
    - foreign_ed/O
    - foreign_oth/O
3. The proportions for each of the sources was applied to the current year value of O.

For those institutions where last year's distribution was not reported:
1. The mean proportion of O to source_oth was calculated source_oth ($\bar{p}_O$= (O/source_oth).
2. The mean proportion of each of those foreign sources to O was calculated:
    - $\bar{p}_f$= (foreign_gov/O)
    - $\bar{p}_e$= (foreign_ed/O)
    - $\bar{p}_t$(foreign_oth/O)
3. A total was computed as a sum of the three means calculated in the 2nd step.
4. A percentage of the total was computed for each variable (mean/total mean).
5. That percentage for each of the sources was applied to O.

For this question, there was an additional normalization step in the imputation procedures. The normalization step ensures that the five detail variables sum to the previously imputed or reported total.

- A total of the detail source data was calculated.
- A percentage of the summed total was computed for each variable (detail foreign source / sum of those imputed values).
- That percentage was applied to the previously imputed or reported total foreign expenditures to compute the imputed value.

Tables 17 lists summary data on foreign funded imputed amounts and rates by foreign source.

Table 17. Imputed and Aggregate Amounts for Total R&D Funded by Foreign Sources by Foreign Source
(amounts are dollars in thousands)

Foreign Funding Source	n	Imputed	Total	% Imputed
Foreign Governments	46	11,885	253,140	4.70%
Foreign Businesses	46	67,544	546,291	12.36%
Foreign Nonprofit Organizations	46	19,889	273,180	7.28%
Foreign Higher Education Institutions	46	13,150	117,876	11.16%
All Other Foreign Sources	46	9,306	67,907	13.70%
Total	26	62,434	1,258,394	4.96%

n = number of institutions with imputed values

R&D Contracts and Grants (Question 3)

Question 3 included three values: External Funding Received Through Contracts (external_contracts), External Funding Received Through Grants and Other Agreements (external_grants), and Total External Funding (external_tot). Total external funding was a known amount from Question 1, equivalent to total R&D (source_tot) minus institutionally funded expenditures (source_inst_tot). If external_tot was 0, contract and grant values were imputed as 0.

If Question 3 was reported last year, this year's value for contracts was calculated by applying the same proportion reported last year (external_contracts / (source_tot – source_inst_tot) to this year's reported or imputed value of Total External Funding. For institutions that moved from the short form population to the standard form population, the proportion from the most recent standard form data (FYs 2011–16) was used, if reported.

If there were no reported data from last year, the mean proportion of external_grants / external_tot was calculated for the non-missing values within imputation classes determined by TOC, HDO, the quartiles of Total R&D Expenditures, and the median value of Federal R&D Expenditures. The imputed values were calculated by applying the mean proportions to Total External Funding, either reported or imputed.

Table 18 lists summary data on externally funded imputed amounts and rates by type of agreement.

Table 18. Imputed and Aggregate Amounts for Total Externally Funded R&D Expenditures by Type of Agreement
(amounts are dollars in thousands)

Type of Agreement	n	Imputed	Total	% Imputed
Contracts	31	543,762	13,550,998	4.01%
Grants and Other Agreements	31	2,041,993	45,339,443	4.50%
Total	16	106,990	58,890,441	0.18%

n = number of institutions with imputed values

R&D Expenditures at Medical School (Question 4)

Question 4 included one expenditure amount, R&D Expenditures Within the Medical School (med_sch_tot), and a flag variable indicating that the institution did not have a medical school. The existence of a medical school was researched using online data sources.

If the institution was determined to have a medical school and if Question 4 was reported last year, this year's value was calculated by applying the same proportion reported last year (med_sch_tot / Total R&D Expenditures) to this year's reported or imputed value of Total R&D Expenditures. For institutions that moved from the short form population to the standard form population, the proportion from the most recent standard form data (FYs 2011–16) was used, if reported.

If there were no reported data from last year, a mean expenditure amount by imputation class was calculated for institutions reporting medical schools. Imputation class was determined by TOC, HDO, the quartiles of Total R&D Expenditures, and the median value of Federal R&D Expenditures. The imputed value was the calculated mean if the mean of that imputation class was less than the total reported in Total R&D Expenditures for that institution. If the calculated mean for the imputation class was greater than Total R&D Expenditures, the imputed value was assigned the value of the total.

Table 19 provides summary data on medical school imputed amount and rate.

Table 19. Imputed and Aggregate Amounts for R&D Expenditures Within a Medical School
(amounts are dollars in thousands)

	n	Imputed	Total	% Imputed
R&D Expenditures at Medical School	17	34,611	27,851,411	0.12%

n = number of institutions with imputed values

Clinical Trial Expenditures (Question 5)

Question 5 included three expenditure amounts (i.e., Federal Expenditures for Clinical Trials (trials_fed), Nonfederal Expenditures for Clinical Trials (trials_nonfed), Total Expenditures for Clinical Trials (trials_tot) and a flag variable indicating that the institution did not conduct clinical trials.

If Question 5 was reported last year, even partially, this year's value was calculated by applying the same proportion reported last year (trials_tot / source_tot) to this year's reported or imputed value of Total R&D Expenditures. The imputed amount for Total Expenditures for Clinical Trials was distributed across details based on the relative proportions reported last year. For institutions that moved from the short form population to the standard form population, the proportion from the most recent standard form data (FYs 2011–16) was used, if reported.

If there were no reported data from last year, a mean expenditure amount by imputation class was calculated for institutions reporting clinical trials. Imputation class was determined by TOC, HDO, the quartiles of Total R&D Expenditures, and MedS. This value was used to impute total clinical trials (trials_tot).

Federal and nonfederal amounts were then imputed using a proportion mean ($\bar{p}$). The imputed proportion was for expenditures for federal clinical trials (p1), while 1 - p1 was the proportion for nonfederal clinical trials (p2). The mean proportion of trials_fed / trials_tot was calculated for the non-missing values within imputation classes determined by TOC, HDO, the quartiles of Total R&D Expenditures, and MedS. The imputed values were calculated as trials_tot * p1 for federal clinical trials and trials_tot * p2 for nonfederal clinical trials.

Table 20 lists summary data on total and federally financed clinical trial imputed amounts and rates

Table 20. Imputed and Aggregate Amounts for Clinical Trial Expenditures
(amounts are dollars in thousands)

	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
Clinical Trial Expenditures	22	43,920	1,047,345	4.19%	21	97,505	2,975,165	3.28%

n = number of institutions with imputed values

Type of R&D (Basic, Applied, or Experimental Development) (Question 6)

Question 6 included 12 expenditure values: federal, nonfederal, and total amounts for basic research, applied research, experimental development, and overall R&D. Two cycles of imputation were performed for Question 6, one for the federal column and one for the nonfederal column. The totals for each column, Federal R&D Expenditures and Nonfederal R&D Expenditures, were known amounts from Question 1. If the total of either column was 0, the contributing values were imputed as 0.

Imputation was based on last year's data only for FY 2018 unit nonresponders that had reported data for Question 6 in FY 2017. In this case, the proportion of federal and nonfederal expenditures that were considered basic, applied, and experimental development were based on the relative proportions reported in FY 2017.

For institutions that were partial responders in FY 2018, after logical imputations were completed a logistic regression model was employed to identify cases where values should be imputed as zero. Logistic models were run for each of the Question 6 variables. PROC LOGISTIC was run separately for public and private institutions using the continuous variables Federal R&D Expenditures or Nonfederal R&D Expenditures, HDO, and MedS. If the predicted value ($\hat{p}$) was less than 0.5, the variable in question was imputed as 0.

The next step was the imputation of the nonzero values for basic, applied, and experimental development expenditures. For each variable, the mean expenditure was calculated for the non-missing values within each imputation class determined by TOC, HDO, and the quartiles of Total R&D Expenditures.

For this question, there was an additional normalization step in the imputation procedures. The normalization step ensures that the three variables in each column sum to the known column total.

If all three variables were missing:

- A total was computed as a sum of the class means for each variable.
- A percentage of the total was computed for each variable (mean / total).
- That percentage was applied to the known total (Federal R&D Expenditures or Nonfederal R&D Expenditures) to compute the imputed value.

If only two variables were missing (e.g., applied and experimental development):

- A total was computed as a sum of the class means for each variable.
- A percentage of the total was computed for each variable (mean / total).
- That percentage was applied to the known total (Federal R&D Expenditures or Nonfederal R&D Expenditures minus the reported value, usually basic research expenditures) to compute the imputed value.

Table 21 lists summary data on federal and total imputed amounts and rates by type of R&D conducted.

Table 21. Imputed and Aggregate Amounts for Total and Federal R&D by Type of R&D Conducted
(amounts are dollars in thousands)

Type of R&D	Federal R&D				Total R&D			
	n	Imputed	Total	% Imputed	n	Imputed	Total	% Imputed
Basic Research	69	6,114,888	26,799,164	22.82%	70	9,635,138	49,391,250	19.51%
Applied Research	70	2,577,707	11,963,001	21.55%	71	4,294,099	22,200,867	19.34%
Experimental Development	67	635,313	3,182,917	19.96%	67	1,236,897	7,693,749	16.08%

n = number of institutions with imputed values

Cost Elements of R&D (Question 12)

Question 12 had eight different variables that sum to the known value of total R&D expenditures (source_tot). In addition to total value, three of the variables were known from other questions: Unrecovered Indirect Cost (Question 1), Total Pass-Through (Question 8), and Total Capitalized Equipment (Question 14). If all of Question 12 was missing, the values for these three variables were taken from the corresponding variables in the other questions. In many cases, those values were also missing. For example, if unrecovered indirect cost from Question 1 was missing, it must also be a missing value for Question 12.

As with Question 1, values in Question 12 can only be imputed if the entire question is missing. The imputation of values for individual missing fields would necessarily impact the total R&D reported by the institution, and it was decided that the total R&D reported by an institution would not be altered through imputation.

If an institution reported data for Question 12 in FY 2017, imputation was based on last year's data. Values that were not already imputed as part of other questions were based on the relative proportion of Total R&D Expenditures reported in FY 2017. For institutions that moved from the short form population to the standard form population, the proportion from the most recent standard form data (FYs 2011–16) was used, if reported.

If there were no reported data from last year, a logistic regression model was employed to identify cases where values should be imputed as zero. Logistic models were run for each of the unknown Question 12 variables. PROC LOGISTIC was run separately for public and private institutions using the continuous variables Federal R&D Expenditures, HDO, and MedS. If the predicted value ($\hat{p}$) was less than 0.5, the variable in question was imputed as 0.

The next step was the imputation of the nonzero values for unknown expenditures. For each variable, the mean expenditure was calculated for the non-missing values within each imputation class determined by TOC, HDO, and the quartiles of Total R&D Expenditures.

For this question, there was an additional normalization step in the imputation procedures (see below). The normalization step ensures that the variables in each column sum to the known total.

- A total was computed as a sum of the class means for each variable plus the values of the known variables.
- A percentage of the total was computed for each variable being imputed from the class mean (i.e., not the known values) (mean / total).
- That percentage was applied to the known total minus the known values to compute the imputed value.

Table 22 lists summary data on total imputed amounts and rates by type of cost.

Table 22. Imputed and Aggregate Amounts for Total R&D by Type of Cost
(amounts are dollars in thousands)

Type of Cost	n	Imputed	Total	% Imputed
Wages, Salaries, Fringe Benefits	27	344,725	34,766,504	0.99%
Noncapitalized Software	27	3,143	111,024	2.83%
Capitalized Software	28	244	11,451	2.13%
Capitalized Equipment	25	7,574	2,145,595	0.35%
Passed through	23	22,591	6,344,005	0.36%
Other Direct Costs	28	360,561	17,607,097	2.05%
Recovered Indirect	27	112,711	12,764,810	0.88%
Unrecovered Indirect	14	8,671	5,535,380	0.16%
Total Indirect	27	137,299	18,300,190	0.75%

n = number of institutions with imputed values

Headcount for R&D Personnel (Question 15)

Question 15 had three different variables: R&D Principal Investigators (personnel_pi_count), Other R&D Personnel (personnel_oth_count), and Total Personnel (personnel_tot_count). Questions 15 is the only item in the survey that does not request expenditures. Alternative

procedures were developed because the procedures applied to the imputation of expenditure values could not be used accurately here.

If values for this question were reported last year, the same values were pulled forward and flagged as imputed for FY 2018. If there were no reported data from last year, the imputations of personnel_pi_count, personnel_oth_count, and personnel_tot_count were performed in a stepwise manner. We first imputed personnel_pi_count and personnel_oth_count, then personnel_tot_count was computed from the two imputed values.

For personnel_pi_count (principal investigators), we developed regression models separately for public and private institutions using PROC REG with the independent variables Total R&D Expenditures, HDO, and q12blank (a dichotomous variable based on the completion of Question 12). Predicted values were applied as follows to impute missing personnel_pi_count: if the predicted value is less than 0, personnel_pi_count = 0; otherwise, personnel_pi_count = predicted value rounded to the nearest integer.

Following the imputation of personnel_pi_count, we then modeled personnel_oth_count (other personnel) using the independent variables Total R&D Expenditures, HDO, q12blank, and personnel_pi_count.

The final steps consisted of rounding each component and summing them to obtain personnel_tot_count.

Table 23 lists summary data on total imputed amounts and rates by personnel type.

Table 23. Imputed and Aggregate Personnel Headcounts by Personnel Type

Personnel Type	n	Imputed	Total	% Imputed
PIs	40	9,191	163,638	5.62%
Other Personnel	55	65,296	784,005	8.33%
Total	55	79,729	947,643	8.41%

n = number of institutions with imputed values

Retro-imputation

The last step in the imputation process is performing a backcasting, or retro-imputation, of previous years' imputed data. If an institution reports expenditures after 1 year or more of nonresponse, the current year's data are used to re-impute previous years' data. Retro-imputation is conducted for both unit and item nonresponses. Beginning with the FY 2013 cycle, data were not retro-imputed prior to FY 2010. (It was determined that the possible changes to any imputed values prior to FY 2010 would be too minor to justify the additional effort.) Although values imputed prior to FY 2010 were no longer retro-imputed in FY 2013, reported values from those cycles continued to be used to retro-impute imputed values for FYs 2010–12. Beginning with the FY 2014 cycle, reported values from survey cycles prior to FY 2010 were no longer used during

retro-imputation in any way. All institutions that have been part of the population since FY 2009 have reported more recent data.

During the recoding process occurring prior to imputation, some institutions or their imputed data were removed from past-year records based on additional information collected during the current cycle. The mostly likely source of this information was the population review. Institutions are sent a screener asking about their R&D expenditures in the previous fiscal year. The FY 2018 population review screener asked institutions to categorize their FY 2017 R&D expenditures as one of the following: no expenditures, less than $150,000, between $150,000 and $999,999, or $1 million or more. Four institutions that had been imputed as unit nonresponders during the FY 2017 cycle responded to the screener sent prior to the FY 2018 cycle to say that their FY 2017 expenditures were less than $150,000. Because this new information negated the numbers imputed in FY 2017, the FY 2017 imputed values were removed, and the institutions were excluded from the FY 2017 totals and population counts.

Imputing Back to a Reported Year

Retro-imputation is applied when data are reported following a period of nonresponse. For example, if data were reported for FY 2010 and FY 2018 but not for the intervening years, the difference between the reported figures for each item total would be calculated and evenly distributed across the intervening years (FYs 2011–17) as follows:

Equation 6: $$\hat{\hat{\boldsymbol{y}}}_{\boldsymbol{k_v}} = \boldsymbol{y}_{\boldsymbol{k_u}} + \frac{\boldsymbol{v-u}}{\boldsymbol{t-u}}(\boldsymbol{y}_{\boldsymbol{k_t}} - \boldsymbol{y}_{\boldsymbol{k_u}})$$

where $\hat{\hat{y}}_{k_v}$ is the calculated value of imputed variable $\hat{y}_{k_v}$ for year *v,*
y_{k_u} is the reported value for variable y_k for earlier year *u,*
y_{k_t} is the reported value for variable y_k for current year *t,* and
$t > v > u$.

The highest-level value for each question, which is typically a key value, is imputed for missing years. The new figures are then spread across the lower-level detail figures on the basis of the most recent reporting pattern. This is similar to equation 3, except that the ratio of detail data to key data for the current year is being used to impute past years.

Retro-imputing When There Is No Previously Reported Year

If an institution reports after a period of nonresponse but there was no previous reported year, we apply the reverse of the relevant imputation factor for that variable and year:

Equation 7: $$\hat{\boldsymbol{y}}_{\boldsymbol{ik_{t-1}}} = (\boldsymbol{1} - \hat{\boldsymbol{B}}_{\boldsymbol{k_t}})\boldsymbol{y}_{\boldsymbol{ik_t}}$$

where $\hat{y}_{ik_{t-1}}$ is the imputed value of key variable y_k for institution *i* for year *t-1,*
and $\hat{B}_{k_t}$ is the inflator/deflator factor for key variable y_k in year *t* (see equation 1b).

This approach applies only to key variables, the ones imputed based on imputation factors. To retro-impute lower-level values, we apply the ratio of detail data to key data for the current year.

All questions except Questions 1.1, 10, 13, the question asking for ARRA expenditures (removed during the FY 2015 cycle), and the one asking for a headcount of postdocs (removed during FY 2016 cycle), which are not imputed, are retro-imputed. Question 15, which was not reported on an institution level prior to FY 2012, is retro-imputed back to FY 2012 only. Because Question 15 is not imputed using inflator/deflator factors or as a proportion of a reported expenditure amount, past-year values are retro-imputed with the values reported in the current year.