ATTACHMENT 14: IMPUTATION SECTION FROM THE 2024 GSS METHODOLOGY REPORT

9.a Describe Imputation Methods Used

The 2024 GSS collected 543 data items related to enrollment and financial support for full-time and part-time master's and doctoral students, postdocs, and NFRs. Of the 543 data items collected in the GSS, the item imputation rates ranged from 1.7% to 7.3%. The survey imputed all missing data. The *item imputation rate* is a measure of the amount of missing data for each key total and grid detail variable collected on the GSS. For all items imputed, the mean item imputation rate was 4.2%, where 186 items had imputation rates between 1% and 3%, 157 items had rates between 3% and 5%, 193 items had rates between 5% and 7%, and 7 items had rates between 7% and 9%.. Table 9-1 presents a summary of the proportion of imputed data for full-time and part-time master's students, full-time and part-time doctoral students, postdocs, and NFRs.

Table 9-1
Proportion imputed for part-time and full-time graduate students, by degree type, postdoctorates, and nonfaculty researchers: 2024
(Number and percent)

Personnel type	Total	Number reported	Number imputed	Percent imputed
Master's part-time students	183,893	180,560	3,333	1.8
Master's full-time students	322,037	316,225	5,812	1.8
Doctoral part-time students	37,547	36,930	617	1.6
Doctoral full-time students	274,601	272,937	1,664	0.6
Postdoctorates	69,877	68,009	1,868	2.7
Nonfaculty researchers	35,142	34,377	765	2.2

Note(s):
Detail does not add to total due to rounding.

Source(s):
National Center for Science and Engineering Statistics, Survey of Graduate Students and Postdoctorates in Science and Engineering, 2024.

9.a.1 Imputation Methodology

Different imputation techniques were used for units with and for those without comparable historical data. For units missing a key total (total full-time master's, full-time

doctoral, part-time master's, and part-time doctoral students, total postdocs, or total NFRs) with at least 1 year of qualified historical data, a carry-forward (CF) imputation method was used. The CF method matched the imputee record to its most recent eligible historical record, designated as the *base record*. GSS data from three years prior were used as *base periods* for graduate students, PDs, and NFRs. Once the base records were identified from past GSS data, inflation factors based on the ratio of the current year total to the prior year total were calculated for each of the six key totals to account for year-to-year change. The previous year's key totals were carried forward as the imputed values for the current year's key totals and imputed according to the previous year's proportions.

For units that reported totals but no details, the details were imputed according to the prior distribution if qualified historical details were available. Otherwise, the survey used a nearest-neighbor imputation method. In this method, a donor unit that was "nearest" to the unit whose data were being imputed (imputee) was identified among all responding units having similar characteristics as the imputee (such as having the same GSS code for program fields and offering a doctoral degree). When the survey imputed graduate student details, the selected nearest neighbor was the one that had full-time and part-time graduate enrollments that were most similar to the imputee's enrollments by degree type. The imputed values were calculated by adjusting the donor's values to account for the difference in full-time and part-time enrollment totals within degree type between the two units.

Similarly, when the survey imputed postdoc or NFR details, the total number of postdocs or NFRs, respectively, was used to choose the nearest neighbor. If the postdoc or NFR total was missing, the graduate student totals were used to select the nearest neighbor to impute the postdoc or NFR variables. If either the postdoc or NFR key total (or both) was missing, other

available key totals were used to select the nearest neighbor to impute the data. The same donor was then used to impute the details corresponding to the imputed key totals. Occasionally, institutions are not able to provide complete data at the unit level and provide partial data with instructions on how to use the data. These units are marked as special imputation. The most frequent type of special imputation is where institutions provide key totals at the institution or school level and then these totals needed to be spread to the units.

9.a.2 Results of the Imputation

Table 9-2 shows the distribution of imputation methods for key totals (master's students, doctoral students, postdocs, and NFRs) for the 2024 GSS. At least 93% of the key totals did not require imputation, as shown in the row labeled "No imputation." The most frequently applied imputation method was CF for full-time and part-time graduate students by degree type, postdocs, and NFRs. For NFRs, the second most frequently applied imputation method was nearest neighbor. The *2024 GSS Imputation Report* (Ault et al. 2025) provides additional details about the imputation methods.

Table 9-2
Key totals, by imputation methods: 2024
(Number and percent)

	Master's part-time graduate students		Master's full-time graduate students		Doctoral part-time graduate students		Doctoral full-time graduate students		Postdoctorates		Nonfaculty researchers	
Imputation method	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total	23,121	100.0	23,121	100.0	23,121	100.0	23,121	100.0	23,121	100.0	23,121	100.0
No imputation	22,696	98.2	22,695	98.2	22,722	98.3	22,721	98.3	22,224	96.1	21,699	93.8
Carry forward	397	1.7	397	1.7	379	1.6	379	1.6	638	2.8	803	3.5
Nearest neighbor	0	0.0	0	0.0	0	0.0	1	0.0	259	1.1	619	2.7
Adjusted enrollment	28	0.1	29	0.1	20	0.1	20	0.1	0	0.0	0	0.0

Source(s):
National Center for Science and Engineering Statistics, Survey of Graduate Students and Postdoctorates in Science and Engineering, 2024.

9.b Total Nonresponse Adjustments

For institutions or schools that did not respond, all data at the unit level were imputed. These are *total institution nonrespondents or total school nonrespondents*. For these institutions or schools, if prior unit-level data were available, counts were carried forward; if no prior data were available, then the nearest-neighbor method was used.