# Attachment 14. Imputation of GSS 2021 Data

The 2021 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 6.4%. The survey imputed all missing data.

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. Inflation factors were calculated for 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.

In rare instances where neither current-year totals nor data from a prior year were available, a method called *adjusted enrollment* (*AE*) was used for imputation of graduate student totals. Unlike the CF and NN methods, which use only GSS data, the AE method uses Integrated Postsecondary Education Data System (IPEDS) data to estimate the graduate student totals. IPEDS is built around a series of interrelated surveys to collect institution-level data in nine major topical areas, such as institutional characteristics, enrollments, program completion, graduation rates and outcomes, admissions, student financial aid, human resources, finance, and academic libraries. Because the IPEDS data do not include counts of PDs or NFRs, the unit's graduate student enrollment counts in IPEDS were used to identify an NN donor from the pool of GSS units, and the donor's PD and NFR key totals and details were assigned to the imputee. The AE method was not needed for 2021 processing.

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.