Comparison of Graduate Enrollment Data from the NSF-NIH Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) and the National Student Clearinghouse (NSC) data

Background

An initial comparison of the 2019 GSS data against NSC doctoral enrollment data found major discrepancies in terms of total counts, demographics, and field level distributions. In particular, the NSC data showed: fewer doctoral students in GSS eligible fields; relatively high levels of missing sex, citizenship, race/ethnicity data; significantly fewer foreign doctoral students; and higher numbers of doctoral students in fields where professional doctorates are common. As the initial NSC data were pulled for only five cohorts of graduate students (a cohort represents all the students who start their studies in a given academic year), the smaller number of doctorate students in the NSC data was thus reasonable. NCSES requested a review of additional NSC data to answer the following research questions:

1. Can we modify the data request to improve the estimates of doctoral students that can be obtained from the NSC data?
2. What coverage issues can be observed by comparing the NSC and GSS data?
    a. Do the NSC data highlight any coverage issues for the GSS at the school/institution level?
    b. Do the NSC data highlight any coverage issues for the GSS at the field or demographic level?
    c. Do the GSS data highlight any coverage issues for the NSC at the field or demographic level?
3. What are the strengths and weaknesses of using the NSC data as a frame for a survey of doctoral students?
4. What is the potential for using the NSC data for imputation of nonresponse in the GSS?

In brief, the answers to these questions are:

1. While the NSC estimates were substantially improved, the data remain substantively different from the GSS. The NSC data includes many non-research doctorates and the quality of the demographic data are much lower than that obtained through the GSS.
2. NSC was not able to provide their data at the institutional level. Without institution level data from NSC it is impossible to confirm whether the NSC data indicate any coverage issues for the GSS. In addition to the lack of institution level data, the following coverage issues for the NSC exist:
    a. There are many doctorates in non-GSS institutions in all fields, but most of these are in non-GSS eligible fields or degree programs Those within GSS-eligible fields are likely to be those in professional doctoral programs (e.g., PsyD, AuD, JD) that are intentionally excluded from the GSS, or non-doctoral students that are misclassified in the NSC data.
    b. Differences in the sex and citizenship, race, and ethnicity distributions between the NSC and GSS suggest that many of those missing these data in the NSC counts may be foreign doctoral students. However, unless all were foreign and male, the NSC data still seem to be missing a portion of the foreign doctoral population.
    c. Field level coverage is good, though there may be some undercoverage in biological and biomedical sciences, physical sciences, social sciences, and engineering. As noted above, it seems likely that the NSC data are missing a portion of the foreign doctoral population.

3. The NSC data provide much broader coverage of all doctoral students than the GSS as they include the humanities and all other non-GSS eligible fields. At the same time, the NSC data exclude some GSS institutions.
    a. However, the NSC degree data are messy, as they include non-doctoral students combined with professional and research doctoral students.
    b. The biggest potential weakness of using the NSC data as a frame for a survey of doctoral students is that it excludes many specialized research institutions that only offer graduate degrees, as well as nearly half of the institutions in Puerto Rico.
4. Given the inability to obtain institution level data, the substantially lower quality of the demographic data, and the inclusion of non-research-oriented degrees, using the NSC data would not improve current processes for imputing missing data within the GSS.

We provide more detailed discussion of the data request and results below.

Methods and initial limitations

RTI contacted NSC staff to request a series of custom tabulations. The requested tables were designed to minimize suppression of data by field and maximize comparability to GSS data by field and institution. Appendix 1 shows the data definitions used for the revised data request with changes highlighted in red text. The major changes included extending the number of academic cohorts from 5 (2016-20) to 10 (2013-2022) and expanding the request to all graduate students allowing for a comparison of NSC doctoral and master's enrollment data with GSS enrollment data and providing a more comprehensive picture of the NSC data should NCSES want to do a survey of all graduate students.

Because the original NSC estimates based on 5 cohorts undercounted doctoral enrollment in the GSS, we asked the NSC analysts about the best way to identify all potential graduate students in the NSC data and they suggested that we could use the three variables shown below. The first two variables suggested by NSC reflect the student's current enrollment status while the third is intended to capture the highest degree already earned by the student. Categories in black text indicate graduate student status.

Table 1. NSC variables related to graduate student status

Primary Program Credential Level:	Class Credential Level:	Level of Degree Awarded:
01 = Undergraduate Certificate or Diploma Program	A = Associate's	AD = Associate's Degree
02 = Associate Degree	C = Certificate (Undergraduate)	BD = Bachelor's Degree
03 = Bachelor's Degree	F = Freshman (Undergraduate)	CR = Credential
04 = Post Baccalaureate Certificate	S = Sophomore (Undergraduate)	DP = Doctoral-Professional
05 = Master's Degree	J = Junior (Undergraduate)	DR = Doctoral-Research
06 = Doctoral Degree	R = Senior (Undergraduate)	HR = Honors
07 = First Professional Degree	N = Unspecified (Undergraduate)	MD = Master's Degree
08 = Graduate / Professional Certificate	B = Bachelor's (Undergraduate)	NC = Not Classified
99 = Non-Credential Program (Preparatory Coursework/Teacher Certification)	T = Post Baccalaureate Certificate	PC = Post-Bachelor's Certificate
	M = Master's (Graduate)	PD = Postsecondary Diploma
	D = Doctoral (Graduate)	RN = Review Needed
	P = Postdoctorate (Graduate)	UC = Certificate
	L = First Professional (Graduate)	UN = Undetermined
	G = Unspecified (Graduate/Professional)	

Initial data runs based graduate status across any of the three variables yielded counts that were much higher than the initial NSC data based on the 2016-20 cohorts. While investigating these differences, the NSC analysts suggested limiting the analysis to graduate students identified through the primary program credential as this is the variable used for most of their reporting.

Table 2 shows a crosstab of the primary program credential with the level of degree awarded variable for all students enrolled in the fall 2022 that had level of degree information. Of note in this table are 1) that the total number of students with prior degree awarded data are much lower than the expected population of students, which makes sense as many undergraduates would not have any prior postsecondary degrees and 2) that the level of degree awarded data are inconsistent with the notion of progressing to higher degree levels. Over 40% of those with bachelor's degrees were enrolled in undergraduate degree programs and 70% of those with master's degrees were enrolled in undergraduate, post-baccalaureate certificate programs or master's level programs. Based on these inconsistencies, we agreed to simplify the data request to just use primary program credential level data.

<Table 2 (see attached excel file) >

Data quality issues are also raised when looking at the crosswalk NSC uses to code degree level based on degree strings provided by participating institutions (CREDENTIALLookup_20240418.xlsx). A cursory review makes it clear that the NSC research doctorate data include many clinical or professional doctorates that are explicitly excluded from the GSS as well as several non-doctoral level credentials. Table 3 provides several examples of non-GSS eligible degrees counted in the NSC data as research doctorates.

<Table 3 (see attached excel file) >

While it is likely that the vast majority of doctoral students in the NSC data are pursuing PhDs and other doctoral level degrees, the presence of so many non-research doctoral degrees limit the comparability of the NSC data to the GSS data.

The major data quality concern with the NSC data is the incompleteness of the data. As shown in the NSC data element documentation, sex was available for 64% of records, race and ethnicity were available for 62.8% of records, and citizenship available for only 27.2% of records in the 2022- 23 academic year (figures 8, 13, and 17). As shown by NSC, the completeness of these data has improved over time, but they are substantially less complete than the data provided to the GSS. The high missing rates of these data raise the specter of bias within estimates based on these data and severely impact the utility of these data to assist with GSS imputation.

A final major limitation for the proposed analysis was that NSC could not provide the requested data at the institutional level. The NSC custom research team initially thought that they could provide aggregate institution level data to NSF as a federal agency, however, the NSC legal team disagreed and indicated than NSC could not provide these data to RTI or NSF without explicit permission from each institution. This ruled out the institution level analyses we had planned to do and substantially limited our ability to look for coverage issues within the GSS. NCSES could look into obtaining permissions from GSS institutions for their NSC data as part of the GSS data collection, but this does not seem warranted at this time due to the limited expected value of the NSC data for GSS imputation and as the real benefit in terms of identifying coverage issues in the GSS comes from looking at institutions not already in the GSS universe.

Results

Coverage

NSC has excellent coverage of overall student enrollment in the US, reporting an average of 97.1% over the last 5 years (see Enrollment-Coverage-2017-2022.xlsx). Coverage of the GSS institutions as a whole is also very high as shown in Table 4. The 2021 cycle of the GSS surveyed 699 institutions that granted research doctorates or research-oriented master's degrees in eligible science, engineering, or health (SEH) fields. Of these, 670 (or 95.9%) could be matched by the NSC data team. The 670 institutions include the largest institutions and thus contain even higher percentages of graduate enrollments. If the NSC data provide complete coverage of graduate enrollment within the matched institutions, then coverage at the unit, doctoral and master's student levels would be around 99%.

<Table 4 (see attached excel file) >

Of the 29 institutions in the 2021 GSS but not in the NSC data, 17 were graduate only schools (i.e., they do not enroll undergraduate students) and 6 were from Puerto Rico (see Appendix 2 for a complete list of these institutions). Consequently, the NSC data does not provide good coverage of specialized, graduate only research institutions nor Puerto Rican universities. As noted by the grey text in table 4, the majority of the postdocs and NFRs are also found in the NSC institutions but the NSC only collects enrollment data, so provides no coverage of these data. The full NSC institution crosswalk can be accessed at: https://nscresearchcenter.org/wp-content/uploads/NSC_SCHOOL_CODE_TO_IPEDS_UNIT_ID_XWALK_APR-2023.xlsx

Cohort analysis

Table 5 provides NSC graduate enrollment counts by academic year (AY), degree level, and inclusion in the GSS universe. These data show that extending the time frame of the data request from 5 to 10 years was sufficient (very few doctorates were from the oldest cohorts) and that it substantially increased the number of doctoral students returned from the NSC data. The new request increased the number of doctoral students overall by 43.1% (from 554,577 to 793,542 students) and the number of doctoral students within GSS institutions by 31.8% (from 411,035 to 541,650 students). With the exception of AY 14-15, the cohort sizes declined over time as expected due to graduation and attrition. That more students from the AY 14-15 cohort were still enrolled in the fall of 2021 than from the AY 15-16 cohort suggests that there might be an issue with the data, but as the majority of the enrollments were from more recent cohorts, this is likely to have limited impact on the analysis.

<Table 5 (see attached excel file) >

Another oddity in the data is that 14.8% of master's students and 13.8% of those in other or unknown degree programs first enrolled more than 3 years prior. Since most master’s and other certificate programs can be completed in 1 or 2 years, this percentage seems high and suggests that some of these students may be pursuing multiple degrees or may be in a doctoral program that started as a master's which is not reflected in the NSC data.

NSC by field and degree level

Table 6 details NSC graduate enrollment in AY 21-22 by degree level and GSS field and institution level eligibility and shows that the NSC data are much broader than the GSS data. If NCSES wants to expand the fame, NSC data could be useful as a frame for a potential graduate student survey. Note that the overall number of graduate students in this table is slightly higher than in table 4, as it includes doctoral students enrolled after the fall—something RTI did not request. The failure to adhere to the specifications

of the data request represents a data quality issue and limits comparability to the GSS which includes only those enrolled in the fall.

<Table 6 (see attached excel file) >

In the NSC data, field of study is available for almost all graduate students. Most graduate students (68.5%) were pursuing degrees in non-GSS eligible fields, and just over half (2.5 million of the almost 5 million graduate students in AY21-22) were pursuing non-SEH related degrees. Among doctoral students, the majority (57.4%) were in GSS eligible SEH programs, but only 72.1% of these students were in GSS-eligible institutions. Among master's students, these statistics declined to 34.0% and 67.7%. While it would be useful to look at these data at the institutional level to determine if there are GSS eligible programs outside the GSS institutional universe, it is likely given the frame updating activities undertaken annually within the GSS contract that the vast majority of these doctoral and master's students are in professional or clinical doctoral programs that are not eligible for the GSS or are in other types of programs that are misclassified as research doctoral or master's programs (see table 3 above).

Comparison to GSS by broad field

Table 7 compares GSS and NSC doctoral and master's student enrollment counts within GSS eligible institutions and GSS eligible fields by broad field. This analysis uses the 2021 GSS data as they are most comparable to the NSC data as they reflect Fall 2021 enrollments. As noted above, the NSC enrollment data includes all enrollments in the 2021-22 AY so should be higher than the GSS counts. While this is true for master's level enrollment, it is only true for a few doctoral fields. That the NSC enrollment counts in psychology and health are so much higher than the GSS counts is primarily due to the inclusion of PsyDs and the clinical and professional health doctorates that are excluded from the GSS.

<Table 7 (see attached excel file) >

If NSC could exclude those degrees, the overall counts would be much more comparable but still lower than the GSS data. It could be that some of doctorates are outside of the 10-year window and expanding to 15 years would make the counts closer, but it seems equally likely that some of the students classified as master's students in the NSC are classified as doctoral students in the GSS data as hypothesized in the cohort level analysis above.

Comparison to GSS by demographics

Table 8 provides a comparison of the counts and distributions of the doctoral students by citizenship, race/ethnicity, and sex across the two NSC data requests and the 2021 GSS. This table shows that the new data request was successful in substantially reducing or eliminating the suppression issues in the first NSC data by requesting these data at more aggregate level (Appendix 3 provides the raw NSC counts by degree level within GSS institutions). However, the revisions to the data request did not address the issues that come from the incompleteness of the citizenship, race/ethnicity, and sex in the NSC data. More than a quarter (26.9%) of doctoral enrollment records from GSS institutions in the NSC data are missing citizenship and race/ethnicity in the NSC data; 8.0% are missing sex.

<Table 8 (see attached excel file) >

As noted in multiple data collections, high levels of temporary visa holders and higher proportions of men than women are long-standing characteristics of the doctoral population in the United States. This is reflected in the GSS data, but not the NSC data. If all the doctoral students missing these data in the NSC records were foreign males then these distributions would be similar, but this is implausible. As a result, it

seems likely that the NSC data are missing some of the foreign nationals studying in the United States and this accounts for some of the discrepancies between the GSS and NSC enrollment counts by within GSS eligible fields in GSS eligible institutions.

Appendix 4 expands table 7 to show enrollment by detailed field and adds columns for the percent female and percent foreign (temporary visa holders). What is striking about this analysis is that despite being overrepresented as among all doctoral students whole in the NSC data, female doctorates the proportions of female students are lower in the NSC data than the GSS in most fields (see pink highlights in columns M and Z) and that the overrepresentation of women in the NSC data is driven by the large numbers of doctoral and master's students in the health fields, where many clinical and professional doctoral and master's degrees are intentionally excluded from the GSS.

Summary

Based on the analysis above, NSC data are not directly comparable to the current GSS. As such, the NSC data do not provide much value for validating or imputing the GSS data. The utility of the NSC data is currently limited by the quality issues with the NSC data such as the inclusion of non-research doctorates within the NSC research doctorate enrollment counts as well as missing data rates. While there could be more value obtained from the NSC data if NCSES could get access to the NSC data at the institutional level, this would still be limited by the data quality issues. The likelihood of finding significant coverage issues in the GSS are low given the annual universe review conducted for the GSS using IPEDS and the quality control procedures implemented as part of the GSS data collection.