Locating and Extracting Postsecondary Degree Information for the Survey of Doctorate Recipients

Carolina Milesi and Jake Bartolone

NORC developed and piloted a method for updating participant records in the Survey of Doctorate Recipients (SDR) to complete missing postsecondary degree information. The new method extracted degree information that was accurate when compared to existing SDR degree inventory records, which suggests that this method could be used to supplement records with missing postsecondary degree information.

The Survey of Doctorate Recipients (SDR), conducted by the National Science Foundation, includes a sample of 120,000 science and engineering doctorate recipients who earned their degrees from institutions within the United States. The SDR is the only source of data on the careers of science and engineering doctorate holders from U.S. institutions. The SDR regularly collects data from sample members on all postsecondary degrees attained. However, some survey participant records, particularly those recently selected for the 2015 SDR as a part of the expanded and redesigned sample, are missing information about the participant's most recent Master's degree and/or first Bachelor's degree. For analytical reasons, it is particularly important to have complete degree history information about the first Bachelor's degree.

NORC designed and piloted a method for locating and extracting information about participants' postsecondary degrees from publicly available sources for SDR sample members, including doctorate degree, most recent Master's degree, and first Bachelor's degree, and tested the accuracy of this method in three ways:

- To what extent is this method able to locate and extract this postsecondary degree information?
- Is this method able to fill in postsecondary degree information that is currently missing from SDR records?
- What is the accuracy of the postsecondary degree data collected compared to non-missing SDR records?

LOCATING AND EXTRACTING POSTSECONDARY DEGREE INFORMATION METHODOLOGY

During locating activities for the SDR, specialized NORC staff attempt to find sample members' contact information via Internet searches and other methods. During these Internet searches, staff often find more extensive information about sample members, including their Curriculum Vitae or other publicly available sources that include degree information. While the SDR contains near-complete information about sample members' doctorate degrees, some records are missing one or more data elements related to the participant's most recent Master's degree and/or first Bachelor's degree.

NORC selected 200 2013 SDR sample members to test the efficacy of this method for extracting degree information. To capture cases with a wide range of postsecondary trajectories, NORC divided the SDR sample frame into U.S. residents and those living abroad and then sampled within these two groups by race/ethnicity, sex, employment sector, and major occupation group of their current or most recent job.

For each of the selected 200 SDR sample members, NORC used standard locating procedures to find each sample member online. While NORC locating staff usually have access to all available degree history information when searching for individuals, only the doctorate degree information was used in this research task so the results could be

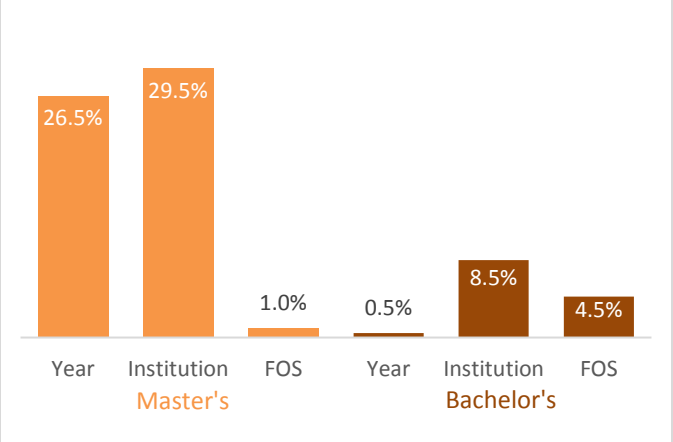
compared to non-missing SDR records for accuracy. For each found individual, staff extracted and recorded information about their doctorate degree, their most recent Master's degree and their first Bachelor's degree.

RESULTS

Rate of Missing Degree Information for the Included SDR Sample Members

Exhibit 1 shows the number of SDR sample members (out of the 200 included in this pilot) who were missing information about their most recent Master's degree or their first Bachelor's degree, including the year the degree was awarded, the degree-awarding institution, and the sample member's field of study (none of the 200 sample members were missing information about their doctorate).

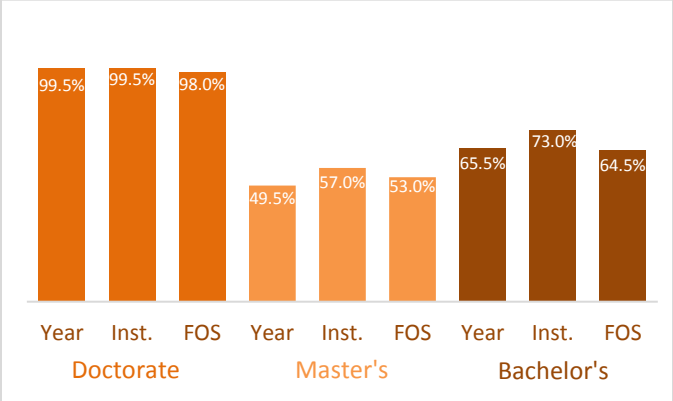
Exhibit 1. Percentage of selected SDR cases missing year, institution, or field of study degree information for most recent Master's degrees and first Bachelor's degrees



Extraction Rate

Overall, we were able to extract data for most of the selected SDR cases, finding information about doctorates for most of the selected cases (to verify that the correct sample member was located and compare against SDR records), most recent Master's degree information for a little more than half of the selected cases, and first Bachelor's degree information for between two-thirds and three-quarters of the selected cases. Exhibit 2 shows the successful extraction rate for each of the three elements of degree information (the year the degree was awarded, the degree-awarding institution, and the sample member's field of study) for each of the three degree types.

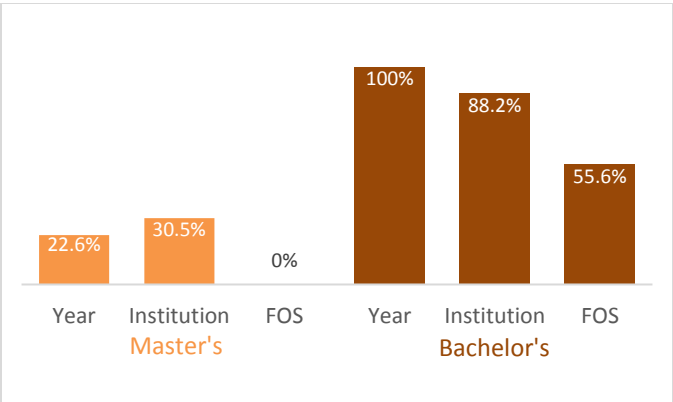
Exhibit 2. Percentage of selected SDR cases for which year, institution, or field of study degree information was extracted for doctorate degrees, most recent Master's degrees, and first Bachelor's degrees



Filling in Missing SDR Data

Exhibit 3 shows the proportion of cases that had missing data elements in the SDR records for which we successfully extracted data for their most recent Master's degree or their first Bachelor's degree. We had greater success extracting Bachelor's degree-related information than Master's degree-related information, although fewer cases were missing Bachelor's degree-related information in the SDR records. Please refer to Exhibit 1 above for the rates of missing data elements in the SDR records.

Exhibit 3. Percentage of selected SDR cases missing year, institution, or field of study degree information for which information was extracted for most recent Master's degrees and first Bachelor's degrees



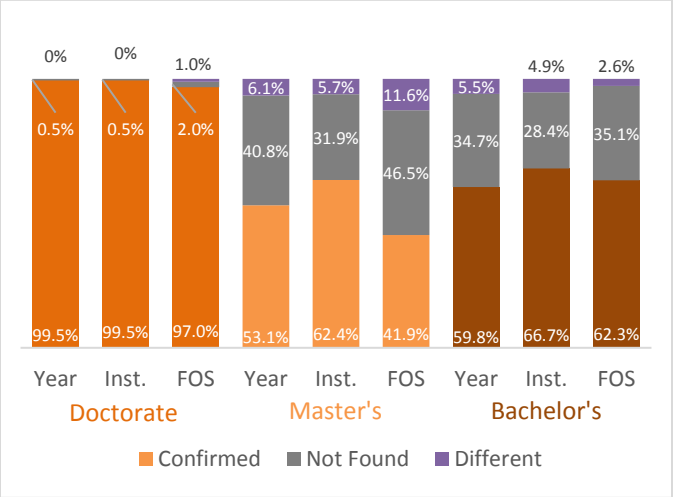
Comparison to Existing SDR Data

In addition to the rate at which we were able to fill in missing SDR data elements, we compared the data we extracted to the existing SDR records (that were not missing) to assess the accuracy of this method. Overall, the data that we extracted matched quite well with existing SDR records, particularly for doctorate-related information. In some cases, we extracted data that differed from non-missing SDR data elements for most recent Master's degrees, and to a lesser extent for first Bachelor's degrees.

Exhibit 4 shows how the extracted data compared to SDR records, with the bottom portion of each bar showing the match rate between the extracted data and non-missing SDR data elements, the middle portion showing the percentage of data elements for which no data were extracted, and the top portion showing the percentage of data elements for which the extracted data were different from the data elements present in the SDR records.

Please note that Exhibit 2 shows the extraction rate for each data element; unsuccessful extractions comprise the "Not Found" portion of Exhibit 4. Successful extractions contribute to the "Confirmed" and "Different" portions of Exhibit 4 only for those cases where the data element was not missing in the SDR records. Exhibit 1 shows the rate of missing data elements in the SDR records.

Exhibit 4. Percentage of selected SDR cases for which extracted data confirmed SDR records, for which no data were extracted, and for which extracted data differed from SDR records, for year, institution, and field of study degree information for doctorate degrees, most recent Master's degrees, and first Bachelor's degrees



NEXT STEPS

This method was designed to be conducted in tandem with locating efforts and with full use of currently available SDR degree information. Because this pilot was conducted to assess if missing information on most recent Master's degrees and first Bachelor's degrees could be reliably extracted from publicly available sources, the NORC staff involved in the pilot did not have access to all currently available SDR degree information. Because of this focus, and because the method used in this pilot was not integrated with SDR's typical locating procedures, this pilot did not accurately measure the cost (in labor) of the method. Future investigation might assess two related topics: the additional time needed to extract missing degree information while conducting the locating task, and the cost for extracting missing degree information for cases already found.

Additionally, this method could be extended to capture other degree information, including professional degrees earned after the doctorate degree.

ACKNOWLEDGEMENTS

NORC wishes to thank Wan-Ying Chang, Dan Foley, and Emilda Rivers from the National Center for Science and Engineering Statistics at the National Science Foundation for their endorsement of this research, and Karen Grigorian, Lance Selfa, and Michael Yang from NORC for their guidance and input. This work was funded by an NORC IdeaLabs Award.

ABOUT NORC

NORC at the University of Chicago is an independent research organization headquartered in downtown Chicago with additional offices on the University of Chicago's campus and in the DC Metro area, Atlanta, Boston, and San Francisco. NORC also supports a nationwide field staff as well as international research operations. With clients throughout the world, NORC collaborates with government agencies, foundations, educational institutions, nonprofit organizations, and businesses to provide data and analysis that support informed decision-making in key areas, including health care, education, economics, crime, justice, and energy. NORC's decades of leadership and experience in data collection, analysis, and dissemination—coupled with deep subject matter expertise—provide the foundation for effective solutions.