2018

Using National Student Clearinghouse Data for Measuring Public Postsecondary Outcomes: Washington Case Study

ERDC
Education Research & Data Center

AUTHOR

Karen Pyle

Education Research and Data Center

ABOUT THE ERDC

The research presented here uses data from the Education Research and Data Center, located in the Washington Office of Financial Management. ERDC works with partner agencies to conduct powerful analyses of learning that can help inform the decisionmaking of Washington legislators, parents, and education providers. ERDC's data system is a statewide longitudinal data system that includes de-identified data about people's preschool, educational and workforce experiences.

ADDRESS

Education Research and Data Center
106 11th Ave SW, Suite 2200
PO Box 43124
Olympia, WA 98504-3113

PHONE

360-902-0599

FAX

360-725-5174

EMAIL

erdc@ofm.wa.gov

Introduction

The postsecondary outcomes of high school graduates is a metric often used in policy research and reports in Washington state. In our research, ERDC uses the P20W data warehouse as the source of enrollments in Washington public postsecondary institutions. Like many other states, ERDC uses the matching services of the National Student Clearinghouse (NSC) to find enrollments in in-state private and all out-of-state institutions. Concerns about the accuracy and completeness of the NSC matching have been raised by researchers. Goldrick-Rab and Harris cautioned researchers not to assume that NSC matching captures all enrollment.1 And in the frequently cited "Missing Manual" study, differences in NSC coverage among institution types (highest for public 4-years and lowest in the for-profit sector) and geographical regions were found. The authors urged researchers to consider these potential sources of measurement error when using NSC data in studies using estimates of program impacts on postsecondary outcomes.2 Since ERDC and our customers frequently use NSC data for just that purpose, we decided to explore possible problems with the quality of NSC matching.

To do so, we looked at how the NSC Student Tracker system performs in matching Washington (WA) public high school graduates to in-state public 2 year and 4 year postsecondary institution enrollments. The central questions asked by this brief are:

- Do the NSC matching results include enrollments that ERDC finds in their matched WA public 2 year and 4 year data?
- Are the match results comparable across all institutions?
- Is there any evidence of bias?

Background

Nearly all postsecondary institutions that participate in federal financial aid submit their student enrollment records to the National Student Clearinghouse. Using this data, NSC provides a student tracking service to high schools who want to track postsecondary enrollment outcomes of their graduates. High schools and other research organizations submit the names and birthdates of students to the student tracking service. According

1 Goldrick-Rab, S. and Harris, D. H. 2010. "Letter to Colleagues: Observations on the Use of NSC Data for Research Purposes." (http://www.finaidstudy.org/documents/NSC%20Dear%20colleagues%20letter.pdf).

2 Dynarski, S. M., Hemelt, S. W., and Hyman, J. M. 2013. "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes." National Bureau of Economic Research working paper 19552. .(http://www.nber.org/papers/w19552).

to NSC, 96 percent of all students enrolled in degree-granting institutions attend schools that regularly submit their student data. The highest participation rate is among public institutions.[3]

The data includes all students enrolled in credit-bearing courses, including non-degree seeking students. The only students not included are those in institutions that do not participate in the Clearinghouse. These include US military academies, very small private career schools that do not participate in federal financial aid programs, and some private degree-granting colleges. Some students are excluded from match results because they have chosen to block the release of their personal information. The average block rate overall is 5%.[4] Given these factors, we would expect NSC matching to pick up a very high proportion of WA public postsecondary enrollment of WA high school graduates.

Method

To test the completeness and accuracy of the NSC Student Tracker system matching, we compared the results of an NSC match of 2014-15 WA high school graduates to the results of an ERDC match of the same students to the WA P20W data warehouse. We looked at enrollments for the first year after high school graduation (2015-16 academic year, summer through spring terms) at WA public 2 year and 4 year institutions.

To prepare the records for matching with NSC, names and birthdates of the graduates were extracted from the P20W identity matching system, where personal identifying information from all data sources, including K-12 and postsecondary, are stored. Names were standardized to remove all extraneous characters and punctuation and put all letters in upper case. All variations of names associated with each individual were sent to NSC in November 2017, over a year after the students had graduated from high school. Enrollment records were found for 55,692 of the 123,807 records submitted. Of records found, only 513 were blocked by the school or the student. The returned records were filtered to include only the enrollments that fell within the date spans covered by WA public 2 and 4 year institutions for the selected academic year (2015-16), and then aggregated to one record per student per institution per year.

The same high school graduates were run through ERDC's master data management system and matched to WA public postsecondary data in our P20W data warehouse. The data comes from the Public Centralized Higher Education Enrollment System

3 National Student Clearinghouse Research Center. 2014. "Using NSC StudentTracker for High School Reports: Considerations for Measuring the College Enrollment Rates of High School Graduates." (https://nscresearchcenter.org/wp-content/uploads/Considerations-in-Using-NSC-STHS-Reports.pdf.)

4 Ibid.

(PCHEES)[5] and the State Board for Community and Technical College's student data warehouse[6]. (See ERDC's website for more information about the identity matching process). For the purpose of this study, we consider the results of our identity matching to be the source of record for WA public 2 year and 4 year enrollments to which we can compare the NSC match results. The matched records from P20W were filtered to include only records where students were enrolled for credit during the 2015-16 academic year and, like the NSC match results, were aggregated to one record per student per institution. The file from NSC was joined to the P20W match on a person identifier and on institution code, to determine whether or not students were found at the same institution in the same year in both data sets.

Findings

Through matching the 67,808 high school graduates from 2014-15 with P20W data, ERDC found 32,828 enrollments in WA public 2 year or 4 year institutions during the 2015-16 academic year. Of those enrollments, 29,251, or 89.1 percent, were also found by NSC as being enrolled in the same institution during the same year. Looking at it from the opposite perspective, NSC found 30,394 enrollments and, of those, only 1,143 (4 percent) were not found by ERDC. This high amount of overlap is encouraging in terms of NSC overall matching quality.

The 11 percent not found in NSC seems reasonable. The lower match rate for NSC is likely due, in part, to a more conservative matching process that uses less information compared to ERDC's. According to their documentation, NSC goes beyond exact matching of names and birthdates, using algorithms that take into account common misspellings, data entry errors, nicknames and so on, within a small tolerance level. ERDC takes these same steps, but then performs manual review of the remaining potential matches, picking up additional matches that could not be found using

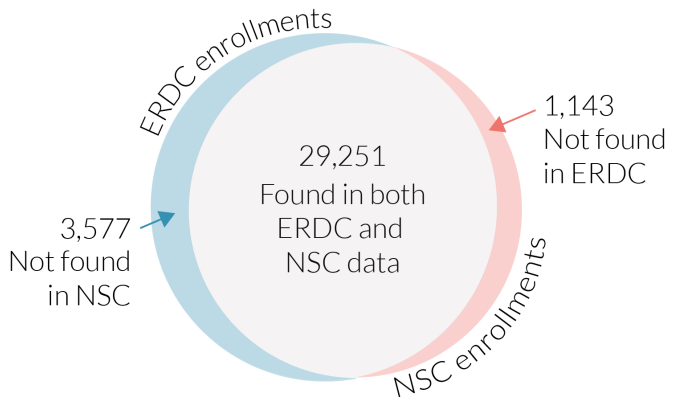

Figure 1. Overlap between ERDC and NSC data (see also Table A1).

5 The PCHEES system, operated by ERDC, collects student data from the six Washington public 4-year institutions.

6 SBCTC collects student data from the state's 34 community and technical colleges and loads it into a data warehouse. Extracts from the data warehouse are provided to ERDC.

automated processes and algorithms. ERDC also has personally identifiable information records on students from many other sources over several years, including marriage and driver's license data, that are used to confirm or negate matches.

Another possibility is that institutions may be reporting more completely to ERDC's sources than they are to NSC. If there is variation in NSC match rates among the institutions, we would be concerned that is the case. To look into this, we calculated the proportion of students found in ERDC data and not found in NSC, for each postsecondary institution.

Figure 2 shows the percentage of WA public postsecondary enrollments found in NSC for each of the WA public postsecondary institutions against the total high school grad enrollment counts at each institution, to give a sense of the impact of missed NSC hits. There are seven 2-year colleges with rates of 80% percent or less. One, albeit small, community college did not have any NSC matches at all. For state-level analysis, these missed NSC hits will not make much impact, as these seven colleges are small and together include about only about 18 percent of all community and technical enrollments of high school grads. However, this raises concern that some community and technical colleges may not be reporting all students to NSC. This could result in underestimates if the school districts that feed into those colleges are using NSC to determine the college-going rate for their high school graduates.

Another concern for researchers is that the matches may be missing disproportionately among students with certain characteristics. Depending on the research design, this could bias postsecondary outcomes measures. To examine this, we analyzed the student characteristics of the cohort of high school graduates, focusing on the proportion not found in the NSC data. Results are shown in Table 1. Asian and Hispanic students and students receiving Migrant Education and English language services have the highest percentages of students whose enrollments were not found in NSC. This is likely due to the structural differences in the names and the use of nicknames, that most student

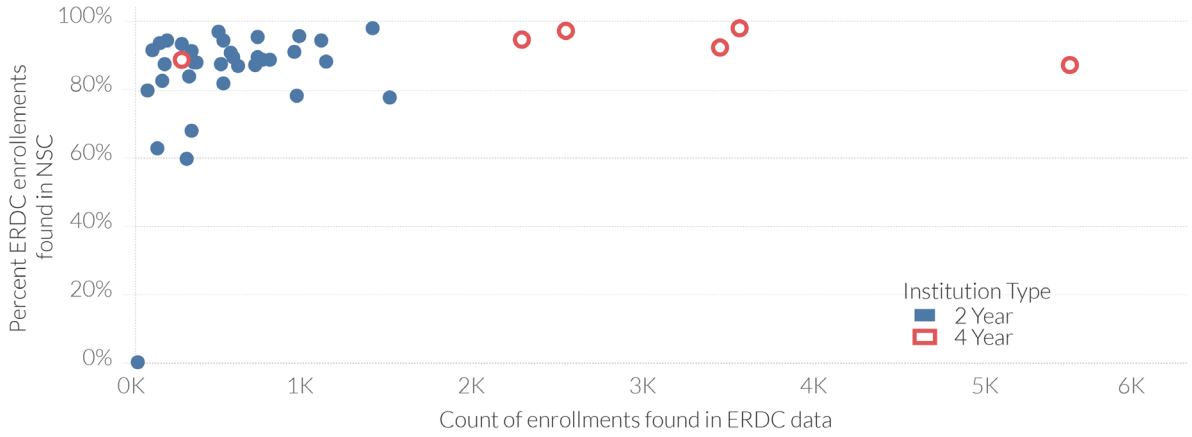

Figure 2. Percent of ERDC enrollments also found in NSC data, by institution (see also Table A1).

data systems are not designed to accommodate[7]. Such differences are difficult to catch in algorithms, which NSC uses, but can be dealt with through ERDC's manual review process. As a result, ERDC matching can better handle matches for non-English names.

Table 1. Selected Characteristics of 2015 High School Grads Found in ERDC Public Postsecondary Data: Proportion Not Found in NSC Data

	Count of 2015 high school grads found in WA public postsecondary data	Proportion not found in NSC data
Gender		
Female	16,835	10.5%
Male	14,874	11.7%
Race		
American Indian or Alaska Native	266	10.9%
Asian	3,795	13.5%
Black or African American	1,515	10.5%
Caucasian or White	19,121	10.2%
Hispanic or Latino	4,934	13.8%
Native Hawaiian or Other Pacific Islander	190	7.9%
Of more than one race or Multiracial	1,888	8.7%
Program participation		
Free and Reduced Price Lunch	10,920	11.1%
English Language Learners	680	14.9%
TitleI Migrant Education Program	464	16.2%

If the colleges with the lowest NSC hit rates are also most likely to have more students with non-English names, that could account for the problem. Looking at race and ethnicity distributions among the colleges does not suggest this is the case. Figure 3 shows that some of the best hit rates are among colleges with the highest proportion of Hispanic or Asian students. However, these are broad race and ethnicity categories and a closer look may show that name structures are indeed at work here as opposed to problems with college reporting students to NSC.

7 Examples of structural differences and use of alternative names among non-English speakers: Chinese, Korean, and Vietnamese names usually list the family name first, followed by the given name. Or the student may have both an English name and a given name, entered in the student record interchangeably. Spanish first and last names often consist of two parts, and sometimes the first part is in the middle name field. For more information, see: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=REL2016158.

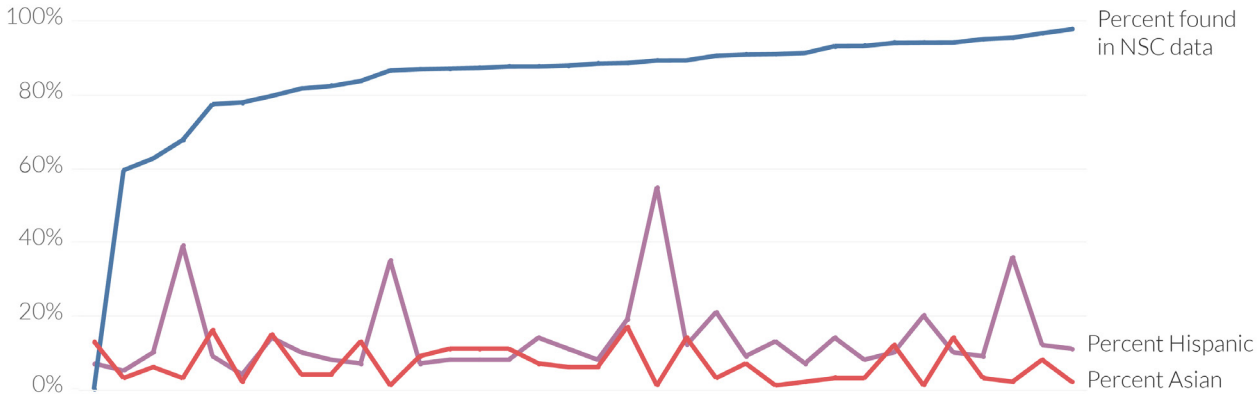

Figure 2. Enrollments Found in NSC by Percent Hispanic and Percent Asian Enrollment at WA Public 2 Year Insitutions, 2015-16 (see also Table A1).

Conclusion

The purpose of this analysis was to examine the accuracy and completeness of the National Student Clearinghouse's Student Tracker matching of high school graduates to postsecondary enrollment data. We matched a cohort of public high school graduates against the ERDC P20W system to find enrollments in WA public postsecondary institutions in the year after graduation. We then looked to see if the enrollments were also found in NSC, in the same institutions during the same year. We were pleased to discover that 89 percent of the enrollments we found in P20W were also found by NSC.

Analysis at the institution level, however, showed very poor match rates for several community and technical colleges. We also found that NSC matching was not as robust for Asian and Hispanic students as it was for other student groups, likely due to issues with names. Some of the missing matches in the colleges could be related to the race and ethnicity of the students in those colleges. But probably not all of it, which means that some colleges may be under-reporting their students to NSC. Overall, the differences we found are of enough magnitude to raise concerns about biases in the NSC match results, especially when breaking out postsecondary outcomes by race and ethnicity or by geographic areas such as school districts. Researchers should be aware of the potential biases and note it in their findings.

Appendix: Tables

Table A1. Counts and percentages of enrollments found in ERDC and NSC data.

Institution Type	Cout of enroll-ments found in ERDC data	Cout of enrollments not found	Percent not found	Percent Found	Percent Hispanic	Percent Asian
2 Year	151	10	7%	93%	8%	3%
2 Year	285	19	7%	93%	14%	3%
2 Year	796	91	11%	89%	8%	6%
2 Year	571	53	9%	91%	21%	3%
2 Year	1137	136	12%	88%	11%	6%
2 Year	334	54	16%	84%	7%	13%
2 Year	188	24	13%	87%	8%	11%
2 Year	376	46	12%	88%	8%	11%
2 Year	22	22	100%	0%	7%	13%
2 Year	516	65	13%	87%	8%	11%
2 Year	1505	338	22%	78%	9%	16%
2 Year	755	85	11%	89%	19%	17%
2 Year	947	85	9%	91%	9%	7%
2 Year	352	43	12%	88%	14%	7%
2 Year	339	30	9%	91%	13%	1%
2 Year	1399	29	2%	98%	11%	2%
2 Year	522	95	18%	82%	10%	4%
2 Year	617	82	13%	87%	35%	1%
2 Year	728	77	11%	89%	55%	1%
2 Year	964	212	22%	78%	4%	2%
2 Year	314	127	40%	60%	5%	3%
2 Year	978	43	4%	96%	36%	2%
2 Year	348	112	32%	68%	39%	3%
2 Year	524	30	6%	94%	20%	1%
2 Year	732	35	5%	95%	9%	3%
2 Year	710	92	13%	87%	7%	9%
2 Year	579	61	11%	89%	12%	14%
2 Year	145	54	37%	63%	10%	6%
2 Year	117	10	9%	91%	7%	2%
2 Year	193	11	6%	94%	10%	14%
2 Year	89	18	20%	80%	14%	15%
2 Year	171	30	18%	82%	8%	4%
2 Year	506	16	3%	97%	12%	8%
2 Year	1105	64	6%	94%	10%	12%
4 Year	3549	72	2%	98%		
4 Year	289	33	11%	89%		
4 Year	2539	71	3%	97%		
4 Year	3437	267	8%	92%		
4 Year	5487	712	13%	87%		
4 Year	2277	123	5%	95%		

Trusted. Accurate. Objective.