Connecticut State Colleges & Universities
Board of Regents for Higher Education

39 Woodland Street
Hartford, CT 06105-2337
860-493-0000
www.ctregents.org

Report for Data Request: SDE-BOR Test Evaluation
P20 WIN_1305_2_0002[1]

Introduction

This report includes analysis of data linked between Connecticut State Department of Education (SDE) and the Board of Regents for Higher Education (BOR) as approved by data request number P20 WIN_1305_2_0002. The report demonstrates the utility of cross-agency data connections and provides information about the validity of the data matching algorithm that is used to match data sets between agencies participating in the Preschool through Twenty and Workforce Information Network (P20 WIN).

The evaluation tests the P20 WIN matching tool by examining how data matched through P20 WIN compared to data matched through a separate, trusted process. Information about the enrollment of Connecticut public high school graduates was matched with postsecondary enrollment data from the Connecticut Colleges and state Universities (CSCU) under the administration of the Board of Regents for Higher Education (BOR). The same data set was also matched with postsecondary enrollment data from the National Student Clearinghouse, and the results of both matches were compared.

It is beneficial to the State to assess P20 WIN data matching functionality and identify areas in need of improvement. In order to maximize the investment being made in the system, the Participating Agencies and stakeholders need to trust the data matching process and the quality of the data output. Conducting this evaluation provides an understanding of system functionality necessary for conducting future audits and evaluations.

Process

P20 WIN utilizes a software package called Data Ladder for conducting matches between participating agency data where there is no unique identifier for linking the data. As a commercial product, Data Ladder has proven utility to its customers; however, there is a need to establish confidence in the probabilistic matching algorithm that is being employed for P20 WIN with the Participating Agencies.

In order to evaluate this tool, SDE developed a file of student data that was matched to data from the National Student Clearinghouse (NSC) using NSC's proprietary algorithm. The same data set was also matched to data from the BOR about students in CSCU institutions using Data

[1] This is a P20 WIN log number that identifies the specific data request and accompanying data sharing agreements.

Ladder for the matching process. For both approaches, a 'match' represents a high school graduate that enrolled in a postsecondary institution. Since the NSC data can be filtered to count only matches for CT Colleges and Universities, the number of matches obtained by the two approaches can be compared directly.

The first part of the process was to match SDE student data to NSC data. The student data file included information about students from the freshman cohort of 2005-6 who graduated from any Connecticut public high school in the 2009-10 academic year. The public high school student file was uploaded to the National Student Clearinghouse (NSC) to obtain the number of students identified by the NSC that enrolled in an institution of higher education anywhere in the United States. NSC data includes information about enrollments at both public and private institutions, so SDE filtered the results to identify the number of students who were reported as enrolled in a 2- or 4-year public institution in Connecticut. SDE also excluded enrollment information from the University of Connecticut which is not a CSCU institution.

The second phase of the process was to match the same SDE student data file to enrollment data from the Board of Regents for Higher Education (BOR). The BOR data file included data about students enrolled in any of the Connecticut Community Colleges or the Connecticut State Universities (CSCU) with birthdates more recent than 12/31/1987. This date was selected so that the output file would contain individuals who were 21 years or younger during the 2009-2010 academic year. The age of twenty-one was used as a boundary for the data set since this is the oldest a student can be and still be served by the public school system. SDE and BOR sent data files with information used to conduct the match to DOL via secure FTP, and DOL conducted the matching process using Data Ladder.[2]

Results

The NSC matching process resulted in 15,570 matches. This means that of the 38,426 students from the 2009-2010 four-year high school graduate cohort, the NSC has a record of 15,570 enrolling in a Connecticut public institution of higher education after high school graduation. Given that the match was conducted by the NSC, we do not know if there were records that could not be matched and thus were left out of the resulting data set. Likewise, we do not know what NSC's match rate is – although we do expect that it is very high.

The matching process with BOR enrollment data resulted in 16,600 matches. This means that the process of matching data from BOR through Data Ladder found 1,030 additional high school graduates enrolled in a Connecticut public postsecondary institution than were found by the NSC. Using the Data Ladder tool, analysts from DOL and BOR were able to review the output based on the thresholds for each matching criteria. Filtering the entire output data set for low threshold scores did not reveal any matches that looked incorrect upon visual inspection – producing an estimated 100% match rate.[3]

[2] Both SDE and BOR have a Memoranda of Agreement (MOA) with DOL that enables them to send data to DOL to conduct these matches. Copies of these MOA can be found at: http://www.ct.edu/initiatives/p20win#approach.

[3] Liam McGucken, Connecticut Department of Labor (e-mail communication, July 11, 2014)

Such a high estimated match rate is an encouraging outcome; however, one must remember that because the matching process employs both deterministic and probabilistic methodology to identify linkages. Without a shared unique identifier, there is an inherent probability that both false positives and false negatives might have occurred, but that we are unable to detect them despite analysis and inspection. With this in mind, the analyst also looked at the match rate from a prior match with a slightly smaller data set.

A prior match using a slightly smaller data set from BOR which produced a total of 16,493 matches contained 2 matches that looked inaccurate with visual inspection. Additional detailed manual review of 1,000 records found 3 more records that were 'interesting cases' meaning it was not clear based upon the available data if the records were correctly matched or not. Attachment A provides details about the business rules for matching and the review process. If one extrapolates this pattern and assumes there could be approximately 5 invalid or questionable matches for every thousand matches, then the match rate would be approximately 99.5%.

Results of Match 1: 2009-10 CT public high school graduates linked to BOR enrollments with cumulative credits of 12 or more.

$$\text{Estimated Match Rate} = \frac{16{,}493 - (\text{\# questionable matches * \# of records in thousands})}{16{,}493} \times 100$$

	Number of CT public HS graduates with PS enrollment information	Number of invalid or questionable matches	Estimated match Rate
NSC	15,570	unknown	Unknown
BOR	16,493	5/1000	99.5%*
Difference	923		

Results of Match 2: 2009-10 CT public high school graduates linked to BOR enrollment independent of credits obtained

	Number of CT public HS graduates with PS enrollment information	Number of invalid or questionable matches	Estimated match Rate
NSC	15,570	unknown	Unknown
BOR	16,600	0	100%
Difference	1030		

Discussion and Conclusions

There are two components to consider when assessing the utility of the data matching process. One should first consider the source of the data for the match and then the data matching algorithm itself.

NSC and BOR are data sources with considerable differences. NSC is a national repository, and BOR is a state-level data source. The NSC has been gathering enrollment and degree data from a large proportion of institutions in the nation for over 20 years, so the NSC has amassed a large volume of data about student participation in postsecondary institution from every state. Despite some limitations and anomalies with data from NSC, they provide an irreplaceable source of information about CT students that attend a campus outside CT. This data is important for auditing, evaluating and improving CT education programs. However, as important as NSC data is for understanding post-secondary enrollment, persistence and completion information, the NSC does not have data from all institutions in CT, nor does it have in-depth information about the experience of students at any institution. The BOR can provide information not only about student enrollment, but also about course placement and course taking behavior. For example, with BOR data, one can connect information from SDE about students' course taking behavior in high school to the same students' course placement information in a Connecticut public college or university. These types of data connections are highly valuable for improving educational programming and the level of students' college readiness.

The value of any data matching process depends on the quality of the algorithm or tool used to conduct the data match itself. While we do not know what the NSC's match rate is, we do know that they are a trusted and respected organization with considerable experience matching student data. Because of this experience, this evaluation used NSC data to produce a benchmark for the number of matches that should be expected when linking SDE and BOR data through Data Ladder. Business rules for the matching process were defined within the Data Ladder software with the goal of approximating the expected number of matches from the NSC comparison. Comparing the number of matches obtained by linking with BOR to the number of matches from the NSC provided insight into the quality of the Data Ladder software.

The results of this comparison found that using Data Ladder to match SDE with BOR produced 1030 more matches than the data match with NSC data. An initial question was whether Data Ladder was identifying accurate matches or creating matches between records for different people (false positives); however, the review of Data Ladder matches did not reveal any pattern or significant number of invalid links. Without evidence of inaccurate matches, the Data Ladder matching process appears valid.

The 'extra' matches obtained by linking directly to BOR enrollment data are attributed to the fact that the BOR has more recent information about college enrollments in Connecticut Community Colleges and State Universities (CSCU) than are present in the NSC data store. CSCU institutions typically upload enrollment data to NSC once a term. Because students can enroll and even complete additional coursework in between data uploads, the BOR should have more enrollment data than is available through NSC.

This assessment shows that linking data between SDE and BOR for postsecondary outcomes directly provides additional value to data obtained from the NSC. This assessment also shows that the P20 WIN tool for conducting matches, Data Ladder, produces trustworthy results.

National Student Clearinghouse matching methodology[4]

The vast majority of matches made by NSC are exact matches pulled from their data base. When an exact match is not possible, the matching algorithm considers the factors to create a hierarchy that determines overall match confidence. Factors utilized by the matching algorithm include: student's name, date of birth, ACT high school code, geographical location and timing of the enrollment. Elements are weighted to distinguish between close matches or to confirm a match.

Name variations account for the largest number of matching challenges, so the NSC has developed algorithms to handle situations such as common misspellings, hyphenated names, titles, changed names, shortened names, compound names and typical data entry errors. NSC does not use the Soundex function because it does not meet their standards for reliability. Instead they have a proprietary logic for matching when names are different by a small tolerance level. Their logic incorporates their experience working with the historical data in their repository. When a definitive match cannot be made with the algorithm alone, NSC analysts assess the data manually to ensure accuracy in the final determination or match.

Limitations

A. The approved data request specifies that there will be no public report utilizing student data obtained from this request. Rather information obtained about the quality of the data matches will be used by the P20 WIN participating agencies to inform future data requests for additional audits or evaluations.

B. It is impossible to know with certainty whether the matches between data sets are accurate without a trusted unique identifier utilized by both systems and for all students.

C. While the National Student Clearinghouse has a deep repository of enrollment data from most institutions in the nation, they do not have a complete data set.[5] In early 2014, the NSC estimates that they have enrollment data on 96% of all enrollments at Title IV, degree granting institutions in the United States. In Connecticut, for the fall of 2010, NSC estimates that they have 93.3% of the enrollment data from all CT institutions both public and private: 99.2% of enrollment data from all public institutions, 89.7% from private, not-for-profit 4-year institutions and 100% from private, not-for-profit, 2-year institutions. As of 2011, the NSC did not have data on students enrolled at Goodwin College or Trinity College which combined enrolled several thousand students. The combined 12 month enrollment of Goodwin College and Trinity College in the 2011-12 academic year was 7,067.[6]

[4] Diana Gillum, National Student Clearinghouse (e-mail communication, April 21, 2014)

[5] See Attachment B for a list of data anomalies identified in 2011 for the BOR's production of the College and Career Readiness Workbook and high school feedback reports.

[6] Source: U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS). http://nces.ed.gov/ipeds/datacenter/InstitutionProfile.aspx?unitId=acadb4acb0af

Table 1: National Student Clearinghouse Enrollment Coverage[7]

Connecticut Sector Description	Fall_2010	Fall_2011	Fall_2012	Fall_2013**
Overall	93.4%	93.9%	94.4%	94.4%
All 4-year Institutions	90.8%	91.6%	92.3%	92.3%
All 2-year Institutions	99.5%	99.4%	99.5%	99.5%
All Public Institutions	99.2%	99.2%	99.2%	99.2%
All Private, not-for-profit Institutions	89.8%	94.1%	97.8%	97.8%
All Private, for-profit Institutions	0.0%	0.0%	0.0%	0.0%
Public, 4-year	98.5%	98.5%	98.6%	98.6%
Private, not-for-profit, 4-year	89.7%	94.1%	97.8%	97.8%
Private, for-profit, 4-year	0.0%	0.0%	0.0%	0.0%
Public, 2-year	100.0%	100.0%	100.0%	100.0%
Private, not-for-profit, 2-year	100.0%	100.0%	.	.
Private, for-profit, 2-year	0.0%	0.0%	0.0%	0.0%

[7] National Student Clearinghouse (2014), Enrollment Coverage workbook (data file). Retrieved from http://nscresearchcenter.org/workingwithourdata/

Attachment A

The business rules for defining how elements are used to conduct matches within Data Ladder are set by creating ‘definitions’, and the user can create as many data matching definitions as necessary to produce the best match rate. For this evaluation which was conducted in preparation for the completion of data request P20 WIN 1305_1_0001 for the State Department of Education, three data matching definitions were defined.

Definition 1	<u>match if all elements fit this pattern</u>
First Name	match based on fuzzy matching rules[8] at 80% level
Middle Name	match based on fuzzy matching rules at 80% level
Gender	match based on fuzzy matching rules at 80% level
High School Code	exactly the same
Date of Birth	exactly the same
Definition 2	
SASID[9]	exactly the same
Definition 3	
First Name	exactly the same
Last Name	exactly the same
Date of Birth	exactly the same

<u>Validity Checks for Match 1</u>

1) SASID Matches = all fine
2) Filter on 100% match (score 2) visual search from beginning looking for changes in H.S. codes and/or gender = all fine. Stopped checking at group 612
3) score 0 <> score 2
   - and score 2 = 66.67% = all matches fine
   - and score 2 = 33.33% = 1 bad match
4) score 0 <75% and score 2 = 66.6; gives 3 cases which are all fine
5) score 0 <85% and score 2 = 66.6%; gives 11 cases which are all fine
6) score 2 = 33.3 ; only 1 bad match
7) score 0 <90 %, visually scanned 1,000 records; all matches appeared fine

[8] Data Ladder’s matching algorithm for probabilistic (a.k.a fuzzy) matching is based on the Jaro-Winkler approach.
[9] The SASID is the State Assigned School Identifier

Attachment B

NSC Data Anomalies

When conducting analysis on the NSC data, the following points of context may be important.

- **Not all CT schools send data to NSC:**
  NSC provides attendance and completion data from over 3,300 institutions representing approximately 98% of the national postsecondary enrollment. Below are the institutions in CT with enrollments of 1,000 or more who *did not* report attendance and completion data to NSC as of 2010.
  - Goodwin (enrollment = 1,589)
  - Trinity (enrollment = 2,504)
  - Post (enrollment = 1,687)

  Other institutions which do not report to the NSC include the following:

  The Graduate Institute
  Hartford Seminary
  Holy Apostles College and Seminary
  Legion of Christ College of Humanities
  Lyme Academy College of Fine Arts
  Mitchell College
  Paier College of Art
  Rensselaer at Hartford
  Sanford-Brown College

- **Data for students may not show the full academic history:**
  - Graduation data may be present without any corresponding enrollment data. This would occur if the enrollment occurred before the institution began sending data to the NSC. For example, Albertus Magnus began sending data in July of 2007, so there are graduation records dated July, 2007, but there are no prior enrollment records.
  - **Some institutions block the student enrollment data which creates a difference between NSC numbers and the data they supply**. When an institution blocks student enrollment data, the NSC can use the data for counts in their reports, but the same data will not be present in the detailed reports (raw data?) that we receive. This means that we will not be able to account for these students. If any of the students in our data file attended a school with such a block, it will lead to a permanent mismatch in the enrollment totals we produce compared to the NSC totals. For example, Columbia University is currently blocking their enrollment records, so this data is not displayed at the detailed level.
  - Some schools do not participate in all of the NSC services and may not be supplying all of the typical data. For example, The Community College of the Air Force is one of a few schools that are currently participating in the NSC DegreeVerify service only. They have not started submitting enrollment data for the EnrollmentVerify service. Therefore, students attending this university would be counted in as "Not in NSC to Date."

- **College names may vary for the same college code**. This can occur when a university has multiple campuses. Cornell University is an example of an institution which reports enrollment data under different unofficial branches even though they are all tied to the same 6-digit college code.