John J. Heldrich Center for Workforce Development

research report

Building a Statewide Longitudinal Database: Matching Data without a Common Unique Identifier

by William Mabe, Ph.D., David Seith, and Syeda Fatima

Abstract

For years, states have collected a vast array of data in the course of administering their health, education, and social services programs but have stored it in disconnected data silos. Now, to measure the effectiveness of their educational investments, nearly all states are attempting to merge these silos and link secondary school exiters — graduates and dropouts — to their higher education and employment records. Most, however, lack a common unique identifier, such as a social security number, to link data across state agencies. States can overcome this disconnect by working with a fourth agency, such as the treasury or motor vehicle agency, that possesses identifying information on state residents. This report describes how scholars used data from the New Jersey Motor Vehicle Commission to link 82% of secondary school exiters (18 or older) between 2011 and 2015 with the State of New Jersey's higher education enrollment and Unemployment Insurance wage data.

Practitioner Points

1. It is technically and practically feasible for states to build longitudinal data systems to connect multiple data collections that for decades have been stored in disconnected silos.

2. For states that need to match secondary school graduates and dropouts with their higher education and workforce records but lack a common identifier to support a match, state motor vehicle data, which contain detailed identifying information on state residents, are a legally permissible and technically feasible solution.

3. Practitioners can increase the number and representativeness of the matches they achieve by including processes for exact matching on substrings of first and last names.

4. Practitioners can validate the accuracy of their matches in a variety of ways, including leveraging additional state data collections and implementing well-established "fuzzy" matching algorithms to detect name misspellings.

Introduction

For decades, states have collected a vast array of data in the course of administering state and federally funded programs in health, education, and social services. Because these data collections have primarily been built to meet reporting obligations under the laws and regulations that govern these programs, they have typically not been designed with the ability to link various data collections with one another. They have instead been stored in disconnected data silos across multiple state agencies (Hotz, Goerge, Balzekas, & Margolin, 1998; Abowd, Haltiwanger, & Lane, 2004).

RUTGERS
Edward J. Bloustein School
of Planning and Public Policy

Yet, because each agency serves state residents at different times in their lives and provides them with different services, each data silo stores a trove of valuable, complementary information. For example, a state's health agency serves a young girl and records detailed information on her health status, while the state and local education agencies educate her through adolescence and collect information on the courses she takes, after which the state workforce agency assists her in finding a job and records the services she receives and whether she obtains employment. Across their many administrative departments, states possess the critical building blocks for a robust, person-specific longitudinal data system: data on multiple characteristics of multiple individuals receiving multiple services over time (Card, Chetty, Feldstein, & Saez, 2010).

Unfortunately, breaking down the data silos presents significant challenges. Numerous state and federal laws regulate the data elements that state agencies may collect and some prohibit the collection of certain data elements, such as the social security number (SSN). As a result, few states have a unique identifier that is common to most or all state agency data systems. Therefore, although multiple agencies within a state house data on the same individuals over long periods of time, most states have not yet been able to combine these data to construct a longitudinal data system of the individuals they serve.

The New Jersey Department of Education (DOE) houses data on kindergarten through 12th grade (K-12) students in its New Jersey Standards Measurement and Resource for Teaching (NJ SMART) data system. The Office of the Secretary of Higher Education (OSHE) warehouses data on the enrollees and graduates of the state's public and independent institutions of higher education, and the Department of Labor and Workforce Development (LWD) maintains the Unemployment Insurance (UI) wage record data. Whereas OSHE and LWD use the SSN as the unique identifier, DOE does not collect student SSNs.

This report describes how New Jersey's education, higher education, and workforce agencies' data were combined to construct a robust, longitudinal data system, depite the lack of a common, unique identifier. Data from the state's Motor Vehicle Commission (MVC) were used as a bridge to link 82% of secondary school exiters age 18 or older, between 2011 and 2015, with records in the state's higher education enrollment and the UI wage record data.

Theory

Scholars currently have access to multiple, high-quality survey data collections to support research on education and the labor market, but the principal reason to develop a state longitudinal data system is because it can offer value beyond the longitudinal datasets that currently exist. The most prominent longitudinal datasets are collected and distributed through the U.S. Bureau of Labor Statistics' National Longitudinal Survey (NLS) program. The longitudinal nature of the NLS data supports the study of the same individuals over time, so scholars have used data from the National Longitudinal Survey of Youth (NLSY) to publish studies on topics as diverse as educational attainment (1,034 citations), labor market outcomes (252 citations), marital disruption (106 citations), adolescent pregnancy (81 citations), school quality (52 citations), and obesity (365 citations), among many others (National Longitudinal Surveys, 2016).

Although the NLS data have supported many studies, the data have some limitations. While the data allow scholars to examine the effects of state-level policies on education and labor market outcomes (e.g., Kane, 1995), NLSY does not contain data on individual participation in state-specific programs, so evaluating the effectiveness of such programs either presents significant methodological challenges or is not possible at all. Sample size limitations can also make it difficult to study subgroups of the population. Dynarski (2004), for example, explains, "Many of the states…do not have enough observations in the NLSY97 to allow state-specific estimates of the share of students whose GPA qualifies them for their state's merit program."

Among the numerous advantages that a well-constructed, state-level longitudinal data system can offer over current longitudinal datasets (Hotz, Goerge, Balzekas, & Margolin, 1998, 1998; Card, Chetty, Feldstein, & Saez, 2010), two are particularly salient. First, state data systems can facilitate

research on the effectiveness of state-specific policies and programs for state residents. Although academics may be satisfied with the results of a rigorous study at the national level about the effect of, for example, merit-based scholarships on educational attainment and labor market success, state legislators want the results for the specific programs their state funds. Although information is one of many factors that influence how state legislators vote, previous studies have found that research — and specifically research from a credible, nonpartisan organization such as a university — affects their decisions (Dodson et al., 2013; Woolard, 2015; Weiss, White, Stohr, & Willis, 2015; Tabak, Eyler, Dodson, & Brownson, 2015; Anderson & Goldstein, 2015). While legislators also rely on research based on the experiences of other states, research on the exact policies and programs that they fund is more influential. Second, by providing a much larger sample size than can feasibly be collected through survey data collection, state longitudinal data systems allow scholars to answer questions, such as how effective specific educational strategies and student academic pathways are for different subgroups of students, which are difficult to answer using the relatively smaller samples available in survey data. This is especially important in the field of education where the marginal returns to education are heterogeneous and differ not only across subgroups but also within them (Card, 2001).

Building a state longitudinal data system that can compete with existing survey datasets requires the integration of data from multiple state agencies. Because different agencies serve state residents at different times in their lives, the data system can only be longitudinal if it integrates data from multiple agencies. Multi-agency data are also needed to supply the covariates required to minimize selection bias. Even studies of outcomes, such as high school graduation, that would seem to require data from only a single agency (education) can benefit from the inclusion of additional covariates. Many factors — often stored outside of a state educational agency's systems — can significantly influence students' secondary school outcomes. These include whether a student works during high school (Tyler, 2003; Staff, Schulenberg, & Bachman, 2010; Marsh & Kleitman, 2005; Leos-Urbel, 2014; Lee & Staff, 2007) or whether his family receives benefits from social insurance and social services programs, including UI benefits (Kukla-Acevedo & Heflin, 2014), the Women, Infants, and Children program (Jackson, 2015), and the Section 8 housing choice voucher program (Carlson, 2015). If data allow scholars to leverage geospatial information, they can examine neighborhood effects, which Chetty, Hendren, and Katz (2016) found to powerfully affect individuals' long-run outcomes.

Methods

Like many states, New Jersey does not have a shared identifier for individuals in its K-12, postsecondary, and workforce data systems. Although OSHE and LWD use the SSN as the unique identifier, DOE does not collect student SSNs. DOE's NJ SMART system, therefore, uses its own unique identifier, the NJ SMART student identification number (SID). The OSHE data have a field for the NJ SMART SID, but for the most recent semester of OSHE enrollment data (Spring 2015), less than a quarter of OSHE records had valid NJ SMART SID values. Table 1 illustrates the shared identifier problem by listing the identifying information that each of the three systems collects.

Without a shared identifier, the data cannot be linked across all three agencies. The purpose of the matching project was to build a longitudinal data system by identifying the correct SSN for as many high school exiters age 18 or older as possible.[1] To support the construction of a robust longitudinal data system, the matching strategy had to meet three criteria:

1. **Match accuracy: Minimize false positives.** False positive matches, which involve counting as a match two records that actually represent two different people, can significantly degrade the quality of a database. Studies based on these mismatches would compare the characteristics and educational experiences of a student to the postsecondary and labor market outcomes of a completely different person. Mismatches can bias inferences in either direction because they introduce "noisy" outcomes.

TABLE 1. NEW JERSEY'S SHARED IDENTIFIER PROBLEM

Identifying Data Field	State Data System		
	DOE	OSHE	LWD UI
NJ SMART SID	X	*	
Social Security Number		X	X
First Name	X		X
Last Name	X		X
Date of Birth	X		
Year of Birth	X	X	
Sex/Gender	X	X	

** The most recent semester (Spring 2015) of OSHE enrollment data contains 306,932 student enrollment records, but only 74,239 of them (24 percent) contain an NJ SMART SID.*

To maximize the accuracy of the matches, the following principles were applied to the matching process.

- Minimize the number of false positive matches, even at the risk of excluding some true matches.
- Use multiple methods to identify matching pairs of records.
- When possible, use multiple data sources and multiple methods to validate candidate matches.
- When the information from one or more data sources directly contradicts the information in another data source, classify the record as a non-match.

2. **Population coverage: Minimize false negatives.** Failing to spot true matches — classifying two records that actually belong to the same person as a non-match — can undermine statistical power; that is, the ability to confidently detect small outcome differences.

To maximize population coverage, the statewide dataset that was likely to have data on the greatest number of high school exiters (the MVC driver's license and state identification card holder database) was identified, and the fuzzy matching algorithms were applied to match records with misspelled names.

3. **Subgroup coverage: Maximize the number of accurate matches from key subgroups within the population.**

In addition to obtaining identifying information on state identification card holders, who are likely to earn lower incomes than driver's license holders, the authors also obtained access to an LWD database that tracks workforce services that New Jersey provides to typically low-income job seekers and unemployed persons (the America's One-Stop Operating System, or AOSOS). The authors also developed an algorithm for matching on "substrings" of first and last names in order to match more Hispanic names, which often include double last names (e.g., Lopez-Rivera) that are sometimes transposed or truncated in data entry.

To build a data system that met these criteria, the data sources against which NJ SMART data could be matched were identified, matching methods to generate a candidate pool of matches were applied, and validation rules to the candidate match pool to classify a pair of records as a match were implemented.

Data Sources Matched with NJ SMART Data

OSHE Enrollment Data. Although most OSHE enrollment records are missing the NJ SMART SID, the NJ SMART K-12 data were matched with those OSHE enrollment records that have an NJ SMART SID. This match identified the SSNs for about 13% of all secondary school exiters between 2011 and 2015.[2]

MVC Driver's License and State Identification Card Holder Data. High school exiters 18 or older in NJ SMART were matched with the MVC data on first name, last name, date of birth, and sex/gender.

About 67% of all secondary school exiters between 2011 and 2015 were assigned an SSN from the MVC data.[3]

AOSOS Workforce Services Data. To maximize the match rate among low-income individuals, New Jersey's AOSOS data were also obtained, which record the enrollment of customers in New Jersey's public workforce system and their demographic characteristics, and also track the participation of workforce system customers in the three largest welfare-to-work programs that serve working-age adults: Temporary Assistance for Needy Families, Supplemental Nutrition Assistance Program, and the General Assistance program, a state-funded program that serves adults without dependent children. About 2% of NJ SMART students were assigned an SSN via a match with AOSOS.

Unemployment Insurance Wage Data. The UI wage data use the SSN as the unique identifier and contain the first and last names of individuals employed in New Jersey each quarter since 1998. These data were not directly matched with NJ SMART, but rather combined with the MVC data toward the end of the matching process to assess the validity of candidate matches. The combined UI-MVC data accounted for about 1% of the matches.

Table 2 presents the data sources that were accessed and the identifying data fields in each.

Matching Methods

The core objective of the match process involved finding the correct SSN for as many of the half million (550,600) New Jersey high school exiters ages 18 or older as possible, from 2011 through 2015, in one of the four data source files. But how exactly does one decide that two records are a match? How does one decide that two records, which may be very similar, are non-matches? This section describes how this problem was approached and how these questions were answered.

Although each of the four sources above contains the SSN — unique identifier — they contain different data elements that can be linked to the NJ SMART base file — paths to the unique identifier. Some students are found in each of the three data sources, while others are not found in any. Students enter the four data sources, if at all, in different stages of life — as drivers or state identification card holders, as college students, as participants in employment and training services, and/or as employees. In each of the databases, the same student can appear multiple times, sometimes with

TABLE 2. MATCHING VARIABLES IN EACH ADMINISTRATIVE DATA SOURCE

	Matching Data Source			
Identifying Data Field	**OSHE**	**MVC**	**LWD AOSOS**	**LWD UI**
NJ SMART SID	*			
Social Security Number		X	X	X
First Name		X	X	X
Last Name		X	X	X
Date of Birth		X	X	
Year of Birth	X	X	X	
Sex/Gender	X	X	X	

** The most recent semester (Spring 2015) of OSHE enrollment data contain 306,932 student enrollment records, but only 74,239 of them (24%) contain an NJ SMART SID.*

different names, either because of legitimate name changes (e.g., a transition from a child's name, "Johnny" to an adult name, "John"; a transition from a pre-marriage last name to a hyphenated last name; or to a new surname from different combinations of hyphenated parental and personal surnames, such as "Kahlo y Calderón" to "Kahlo de Rivera"); or due to abbreviations, nicknames, and/or misspellings.

In order to create a pool of candidate matches, four sets of methods were applied.

Exact Matching. A record in NJ SMART had an exact match in one of the other data sources if it had the exact same values on each of the relevant matching variables (NJ SMART SID + year of birth + sex for the OSHE data and first name + last name + date of birth + sex for the MVC and AOSOS data). Table 3 displays an example of an exact match.

The exact matches are the highest quality matches, because the identifying information in every key field is identical. There are, however, many pairs of records that the human eye would spot as matches, but that would fail the exact match test. Table 4 illustrates four examples.

- In the first example in Table 4, the information in NJ SMART and MVC is identical, except in the MVC data the last name appears in the first name field and vice versa.
- The second example has a "double" last name, which is transposed in the MVC data, while all other identifying information is the same.
- In rows three and four, the individual's middle name has been inserted into the last name field and then the first name field in NJ SMART, but has been dropped from MVC.
- Finally, the fifth row illustrates a limitation of the MVC data, which is that the first name field has been restricted to a maximum length of nine characters, so "Christopher" in NJ SMART is "Christoph" in the MVC data. (MVC last names are similarly truncated, to 17 characters.)

Substring Exact Matching. To identify the types of matches displayed in Table 4 when a record in NJ SMART did not have an exact match in MVC, a second matching method called substring exact matching was applied. In order to enter the candidate match pool as a substring exact match, a pair of records had to match exactly on date of birth and sex/gender and then also match exactly on a specific portion (a "substring") of the first name and a substring of the last name. SQL (Structured Query Language) code was implemented to classify records such as those in Table 4 as matches.

Triangulation Matching (Cross-Validation). For records that failed both the exact match test and the substring exact match test, triangulation match-

TABLE 3. EXACT MATCH EXAMPLE

First	Last	DOB	Sex	MVC First	MVC Last	MVC DOB	MVC Sex
Jane	Doe	01/01/90	F	Jane	Doe	01/01/90	F

TABLE 4. EXAMPLES OF VALID MATCHES THAT FAIL THE EXACT MATCH TEST

First	Last	DOB	Sex	MVC First	MVC Last	MVC DOB	MVC Sex
Jane	Doe	01/01/90	F	Doe	Jane	01/01/90	F
Janet	Doe-Dough	01/01/91	F	Janet	Dough-Doe	01/01/91	F
John	Jacob Doe	01/02/93	M	John	Doe	01/02/93	M
John Jacob	Doe	01/01/93	M	John	Doe	01/01/93	M
Christopher	Doe	01/01/92	M	Christoph	Doe	01/01/92	M

ing, also called cross-validation, was engaged, which was used to expand the information available for assessing the similarity of NJ SMART records by leveraging additional data sources. Specifically LWD's quarterly UI wage record data was searched. Over time, the same person (identified by the same SSN) is sometimes associated with several different representations of her/his first and last names. In that case, the MVC data and the UI data were matched on SSN, which produced records such as the one that appears in Table 5.

Table 5 illustrates how a person with an identical SSN and identical last name as well as the one verbatim first name across the MVC and UI data files is the same person. After joining the MVC and UI data, a subsequent match was conducted, in which the NJ SMART data were matched with the combined MVC-UI data file. In order to enter the match pool as a result of this process, a pair of records had to have an exact match on date of birth and sex, and either an exact or substring match on the first name and the last name. Table 6 presents an example of a triangulation match.

Although the human eye can see that the match in Table 6 is valid, Jane Doe would not be classified as either an exact or a substring match if the researcher were to match the NJ SMART data with only the MVC data. She is a match, however, when the triangulation process is implemented and the NJ SMART data are matched with the combined MVC-UI data. The reason for this is that sometimes her first name (Linda) was reported in the first name field, while at other times her middle name (Jane) was reported in the first name field.

Fuzzy Matching (String Distance Validation). Finally, fuzzy matching, also called string distance validation, algorithms were used — Jaro-Winkler (JW) (Jaro, 1989; Winkler, 1990) and Jaccard (1912) distance — developed by scholars at the U.S Census Bureau to identify misspellings, keystroke errors, and name transpositions in first and last names during data entry. These algorithms assign scores to two different names to evaluate their character-by-character similarity on a consistent scale, which often ranges from 0 (completely similar) to 1 (entirely dissimilar). They are useful in identifying candidate matches, which can then be examined visually by a researcher. The fuzzy matching algorithms were only applied to records that had the exact same date of birth and the exact same sex in both NJ SMART and the data sources to which it was matched.

JW is sometimes referred to as a "heuristic" measure because it is based on the idea that spelling differences of the same name often occur because of typos in letters that are positionally proximate within the name. JW counts the number of letters that two strings have in common within a search distance of up to half of the letters of the longest string. The Jaccard index was also used, a general set theory metric often employed in string distance and other applications, defined as one minus the intersection, divided by the union of two sets, or in this case, the number of letters in common, divided by the total number of unique letters. Both algorithms were implemented using the R software (R Core Team, 2016) "stringdist" computer package (van der Loo, 2014).

TABLE 5. MATCHED MVC-UI WAGE DATA

SSN	MVC First	MVC Last	MVC DOB	MVC Sex	UI First 1	UI First 2	UI Last 1	UI Last 2
123009099	Jane	Doe	10/11/84	F	Linda	Jane	Doe	Doe

TABLE 6. MATCHED NJ SMART-MVC-UI DATA

First	Last	DOB	Sex	SSN	MVC First	MVC Last	MVC DOB	MVC Sex	UI First 1	UI First 2	UI Last 1	UI Last 2
Jane	Doe	10/11/84	F	123009099	Linda	Doe	10/11/84	F	Linda	Jane	Doe	Doe

Table 7 shows a valid match between two records that all other matching methods would miss. In this example, the first name has been misspelled by a single character, and the JW algorithm is able to catch the similarity of the names across both data systems and gives it a very low score, which is indicative of a valid match. (In the data used for this project, a summed JW score across the first and last names below 0.2 indicates what a human coder would classify as a match.)

Match Validation Methods

After the large pool of candidate matches was created, the validity of the matches was assessed using a combination of the triangulation (cross-validation) and fuzzy (string distance validation) methods described in the previous subsection. It was at this stage that a pair of records were classified as a true match or a non-match. All NJ SMART records that were exact matched with MVC data were automatically classified as true matches because those records had identical first names, last names, dates of birth, and sex/gender values.

Next, a series of validation processes were implemented to differentiate true matches from false matches. Cross-validation took place after the NJ SMART data had been matched with one data source. Cross-validation involved taking the matched NJ SMART-OSHE data and then matching it by SSN with another authoritative data source, such as MVC. Conducting this match allowed the data elements to be added — in this case, date of birth, first name, and last name — from the MVC data that were associated with the SSN contained in the OSHE data. The combined NJ SMART-OSHE-MVC data file was then examined to determine whether the values for the additional fields from MVC were the same as the values of those fields in NJ SMART. Table 8 presents two examples: one where the cross-validation confirms the match and another where it invalidates the match.

In the first row of Table 8(C), the match is confirmed by adding the values for first name, last name, and date of birth from OSHE with data from MVC that are identical to the first name, last name, and date of birth in the NJ SMART file for this individual. On the other hand, in the second row of Table 8(C) the disagreement on first name, last name, and sex/gender indicate that the match is not a correct match and so it is discarded. To minimize false positives, both records of a match that fail the validation test are ineligible for future matching; that is, they still "count" in the denominator of the 550,600 students to be matched, but they are no longer eligible to count in the numerator of the match rate.

Finally, the string distance, or fuzzy name validation methods previously described, were implemented. The algorithms calculate the degree of difference between first and last names in each pair of candidate matches generated by the non-exact matching methods (i.e., substring exact, triangulation, and fuzzy). After scoring the matched pairs using the JW and Jaccard metrics, the data were analyzed to identify a threshold above which record pairs would be classified as true matches and below which they would be classified as non-matches. The final selection conditions were set as those matched pairs for which (a) the sum of the JW first name score and the JW last name score was less than or equal to 0.2 **OR** the score for the Jaccard distance of the concatenated first and last name was equal to 0, **AND** (b) the birthdates were identical, **AND** (c) the values for sex/gender were identical. This dual condition finds matching but misspelled first and last names, as well as transposed but identical first and last names.

TABLE 7. FUZZY MATCH EXAMPLE

First	Last	DOB	Sex	MVC First	MVC Last	MVC DOB	MVC Sex	JW_ First	JW_ Last	JW_ sum
Jonathan	Doe	01/01/91	F	Jonthan	Doe	01/01/91	F	0.0625	0	0.0625

TABLE 8. CROSS-VALIDATION EXAMPLES

(A) NJ SMART-OSHE Matched File

SID	Birth Year	Sex	First	Last	DOB	SSN
1234567890	1990	M	John	Doe	01/01/90	123001234
2220000000	1991	F	Jane	Doe	01/01/91	111222111

(B) MVC Sample Record

SSN	MVC First	MVC Last	MVC DOB	MVC Sex
123001234	John	Doe	01/01/90	M
111222111	John	Doe	01/01/91	M

(C) NJ SMART-OSHE-MVC Matched File

SID	Birth Year	Sex	First	Last	DOB	SSN	MVC First	MVC Last	MVC DOB	MVC Sex
1234567890	1990	M	John	Doe	01/01/90	123001234	John	Doe	01/01/90	M
2220000000	1991	F	Jane	Doe	01/01/91	111222111	John	Doe	01/01/91	M

Results

In total, four matching methods were applied — exact, substring exact, triangulation, and fuzzy — to four data sources — OSHE, MVC, AOSOS, and MVC + UI — to match 452,240 out of 550,600 K-12 exiters from NJ SMART, an overall match rate of 82%. These matches can be divided into three categories — platinum, gold, and silver — each of which signifies an assessment of the quality of all records matched in that category.[4] Figure 1 displays the matching results in a graphical format.

Platinum Grade Matches: Exact Matches

The first three sets of matches can be classified as "platinum" grade because a combination of familiarity with the data, manual review, and statistical inference shows that the false positive rate for these matches is extremely close to zero. The platinum matches account for nearly 9 out of every 10 matches.

1. Exact match to OSHE on NJ SMART SID, year of birth, and sex/gender [68,029 records]

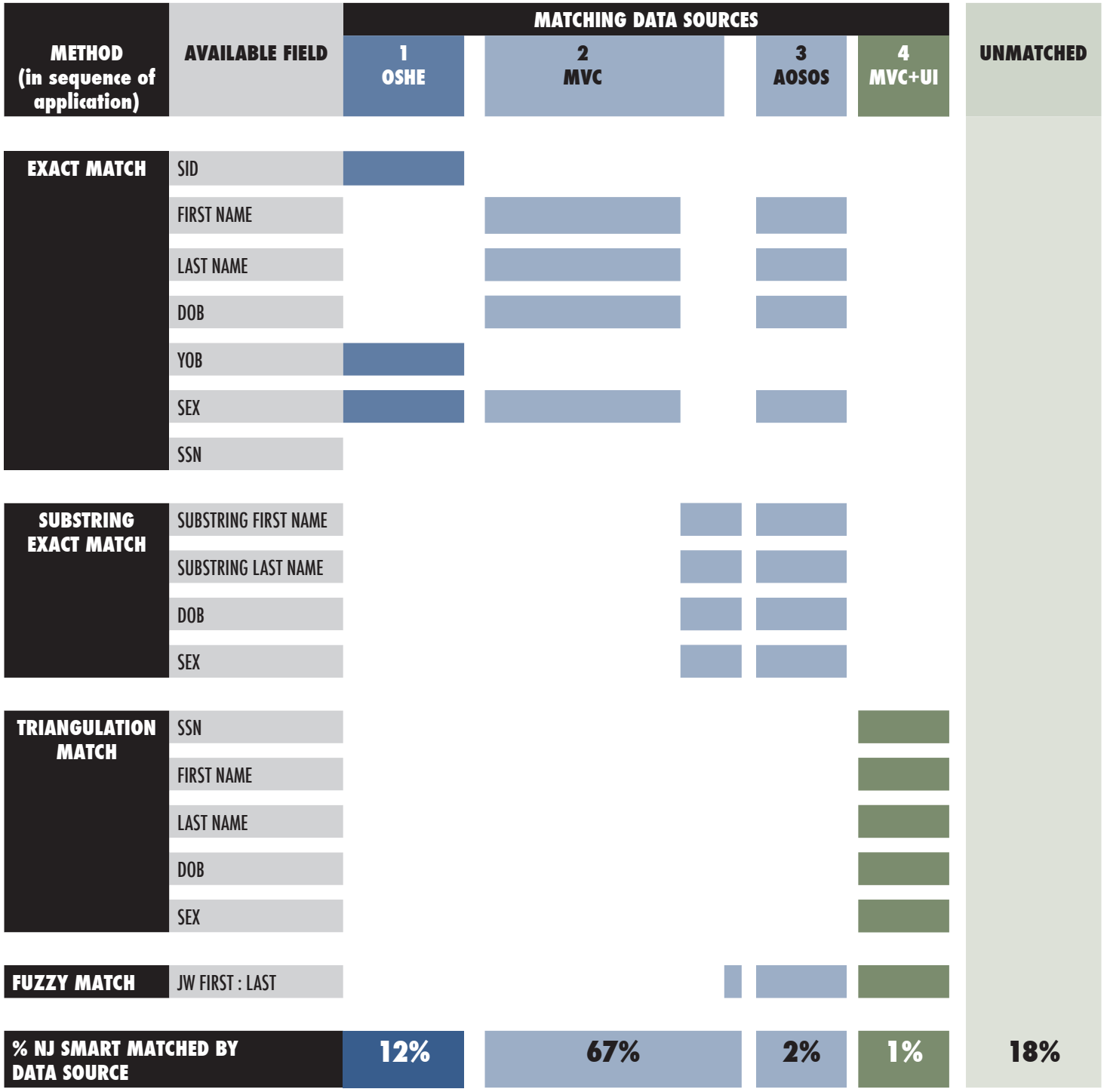
2. Exact match to MVC on first name, last name, birth date, and sex/gender [332,280 records]

3. Exact match to AOSOS on first name, last name, birth date, and sex/gender [1,657 records]

Gold Grade Matches: Substring Exact Matches

The remaining records that matched exactly on birth date and sex/gender were put into a substring of exact name matching NJ SMART exiters first to AOSOS records with first and last names of the same lengths as the first and last names in the NJ SMART data, and then to MVC and AOSOS records with names of different lengths than the NJ SMART names. To minimize the possibility of false positive matches, all names were at least three characters long. In the string-distance validation stage, all name pairs were scored using string-

FIGURE 1. MATCHING RESULTS BY MATCH METHOD AND DATA SOURCE



distance metrics and only accepted those with a summed first and last name JW score less than or equal to 0.20, or concatenated first and last name Jaccard scores of zero.

4. Substring exact match to AOSOS (name of equal lengths) [7,700 records]

5. Substring exact matching to MVC [35,146 records]

6. Substring exact match to AOSOS [815 records]

Gold Grade Matches: Triangulation Matches

After completing the first six groups of matches, a dataset (Cartesian product) of the remaining unmatched NJ SMART records and the remaining unmatched AOSOS and MVC records was created. Separately, they searched through the quarterly UI wage records. Over time, the same person (identified by the same SSN) is sometimes associated with several different representations of her/his first and last names.

As illustrated in Table 8, the "UI multiple names per SSN" file was linked to the "MVC/AOSOS Cartesian Product" file based on SSN. For each pair, the string distance between the MVC/AOSOS first and last name was scored to each of the UI first and last names and then the UI name that best matched the MVC/AOSOS name was accepted. Pairs were accepted as matches when:

- MVC/AOSOS, NJ SMART, and UI birth dates and sex/gender all matched, **OR**

- JW first and last name sum scores equal to or below 0.20 or concatenated first and last name Jaccard scores of zero.

7. Triangulation Matching [5,531 records]

Silver Grade Matches: Fuzzy Matches

The remaining NJ SMART records were exact matched with the remaining MVC and AOSOS records on birth date and sex/gender and accepted as matches to only those records with a summed first and last name JW score less than or equal to 0.20 or concatenated first and last name Jaccard scores of zero.

8. Fuzzy Matching [1,082 records]

Discussion

As explained at the outset of this report, the matching project was not an end in itself, but rather a means to building a longitudinal database that scholars could use to help various state agencies identify the programs and strategies that best meet the needs of students, parents, and job seekers. A key goal of the project was to generate a large enough sample of matched records that, for the study of certain questions specific to New Jersey, could offer significant advantages over available survey data collections. To achieve that goal, the authors sought to generate accurate matches, on as many students as possible, with a high degree of representation across subgroups.

Match Accuracy

Nearly 90% of all of the matched records are exact, deterministic matches, with little opportunity for mismatches. The remainder have been carefully reviewed and scored, and while there will doubtless be some mismatches, the methods employed in this report erred on the side of classifying a pair of records as a non-match as opposed to misclassifying them as a match.

Match Coverage

By using multiple data sources, data for over 80% of all New Jersey students age 18 or older, who exited high school between 2011 and 2015, were matched. The resulting sample size was 452,240.

Subgroup Coverage: Data Representativeness

After completing the matching, the question of how representative the matched data are of all K-12 exiters arose. Due to the limited access to the full complement of variables in NJ SMART, only the information available was used to conduct a preliminary assessment of representativeness. Table 9 compares the match rate by a few important student and school district characteristics. The third column gives an overview of the proportion of sample members with each characteristic who were matched. Match rates were high not only overall, but also within the subgroups for which there were data.

TABLE 9. MATCHING PERCENTAGE BY STUDENT AND SCHOOL DISTRICT CHARACTERISTICS

	Percent of Unmatched Students	Percent of Unmatched Students	Percent Matched	Percent of All Students
Student Characteristics				
Age at High School Exit				
< 17 years old	0.4	0.1	59.5	0.2
17 to < 19 years old	84.8	89.8	83.0	88.9
> = 19 years old	14.8	10.1	75.9	11.0
Gender				
Female	45.7	49.5	83.3	48.9
Male	54.3	50.5	81.0	51.1
Asian Last Name*	8.7	4.2	69.0	5.0
Hispanic Last Name*	18.7	16.6	80.3	17.0
School District Characteristics				
District Factor Group				
A (Lowest SES)	16.6	14.0	79.4	14.4
B	9.3	9.4	82.2	9.4
CD	8.1	8.5	82.8	8.4
DE	11.1	14.2	85.5	13.7
FG	9.3	11.5	85.0	11.1
GH	15.7	17.8	83.9	17.4
I	17.8	14.3	78.6	14.9
J (Highest SES)	5.4	3.2	73.1	3.6
Missing	6.5	7.2	83.6	7.1
NJ SMART Students	98,290	452,310		

** The research team used the R "wru" package (Khanna & Imai, 2016) to identify names that were Asian or Hispanic. The package looks up last names against a master file of last names and race/ethnicity self-descriptions in the U.S. Census. See https://cran.r-project.org/web/packages/wru/wru.pdf.*

Match rates were comparable for women and men, though slightly higher for women, perhaps reflecting their higher college enrollment or labor force participation rates. A key goal of the data match was to maximize the number of matches across and within subgroups. In an effort to assess the rate at which Asian and Hispanic names were matched, the R software package wru — for "Who Are You?" — was used to impute the ethnicity of Asian and Hispanic names (Khanna & Imai, 2016). The name imputations show that although high match rates were achieved for individuals with Asian and Hispanic names, those with identifiably Asian or Hispanic names were somewhat less likely to be matched.

DOE's proxy for the socioeconomic status of each school district — the District Factor Group (DFG) — was used in place of individual student indicators of household income or socioeconomic status. Match rates were high across all DFGs. Students from middle- to upper-middle-income DFGs, however, were more likely to be matched than students from low or high DFGs. For example, the match rate for students from middle DFG factors "DE" and "FG" are at or above 85%, while the match rate for students from the lowest DFG "A" is 79%. The match rate for students from the highest DFG "J" is 73%, perhaps reflecting the fact that middle-income students are more likely to drive to college than lower-income students but less likely to move out of state for college or career than wealthier students.

With respect to subgroup analyses, one advantage that these matched longitudinal data may offer over survey data is that scholars who will use the matched data are likely to know far more about the attributes of the individuals who were not matched than survey data analysts know about the individuals not surveyed. In the case of the NJ SMART data match, scholars have access to detailed information — specifically all of the fields in NJ SMART, including demographics, schools attended, test scores, and courses taken — on the non-respondents, allowing them to weight the data to account for the non-respondents.

Conclusion

The primary conclusion to be drawn from the data matching project is that New Jersey now has a comprehensive statewide longitudinal database that can trace the vast majority of students from high school to college, and into the labor market. This database offers scholars and practitioners the ability to conduct research of the highest quality to inform the state about which educational strategies, programs, and policies most benefit New Jersey residents and to inform the public about which education and training pathways offer them the best opportunity for economic advancement. Several conclusions with respect to data matching, and the state longitudinal data system more broadly, have been made after conducting this data match project.

- Most New Jersey students can be matched one to five years after high school exit, which provides a strong empirical foundation for research on postsecondary education and labor market outcomes in New Jersey. In total, 82.1% of the 550,600 New Jersey high school students who exited the system between 2011 and 2015 were matched. This is both a large number and a high percentage of students. This high rate would support a comprehensive evaluation of the postsecondary and labor market outcomes of New Jersey high school exiters into early adulthood as well as of the programs and strategies used to serve them.

- The majority (89%) of the matches were high-quality, deterministic "platinum" matches. This suggests that empirical postsecondary and labor market outcomes studies will provide unbiased and accurate estimates for future research.

- Using multiple data sources is fundamental to achieving a high match rate, especially among traditionally under-represented groups. Three percent of all of the matches that were identified were individuals who had records in AOSOS data, and these data record information almost exclusively on low-income individuals.

- To boost future match rates, especially among the low-income population, New Jersey should consider including data on the individuals served by the state's largest public income support programs. The use of the AOSOS data showed that using administrative data sources that track the economically disadvantaged can help bolster the match rate among those individuals. Expanding the data sources, such as Temporary Assistance for Needy Families, Supplemental Nutrition Assistance Program, and the General Assistance program, that include information on low-income individuals' data should further enhance the match rate.

- MVC data are an essential data source, key to New Jersey's data linking strategy and to its longitudinal data system. Of the three data sources, the MVC data provide the most matches. More than half (60.3%) of New Jersey students could be matched to MVC data using simply birth date, sex, and verbatim first and last name. In New Jersey, MVC data are also a remarkably representative source, resulting in a balanced representation of students from all but the lowest and highest school DFGs. Over time, it is likely that it will become easier to match New Jersey students to MVC data as MVC implements planned changes to improve data quality.

References

Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. *The American Economic Review*, *94*(2), 224-229.

Anderson, K. S., & Goldstein, M. (2015). *Engaging state legislators: Lessons for the education sector*. Washington, D.C.: Aspen Institute.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127-1160.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). *Expanding access to administrative data for research in the United States*. American Economic Association.

Carlson, D. (2015, November). *The effect of housing voucher receipt on student achievement and attainment: Evidence from Wisconsin*. Paper presented at the APPAM Fall research conference, Miami, FL.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *The American Economic Review*, *106*(4), 855-902.

Dodson, E. A., Stamatakis, K. A., Chalifour, S., Haire-Joshu, D., McBride, T., & Brownson, R. C. (2013). State legislators' work on public health-related issues: What influences priorities? *Journal of Public Health Management and Practice*, *19*(1), 25.

Dynarski, S. (2004). The new merit aid. In C. M. Hoxby (Ed.), *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 63-100). Chicago: University of Chicago Press.

Hotz, V. J., Goerge, R., Balzekas, J., & Margolin, F. (1998). *Administrative data for policy-relevant research: Assessment of current utility and recommendations for development*. Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*, 37-50.

Jackson, M. I. (2015). Early childhood WIC participation, cognitive development and academic achievement. *Social Science & Medicine*, *126*, 145-153.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association*, *84*(406), 414-420.

Kane, T. J. (1995). *Rising public college tuition and college entry: How well do public subsidies promote access to college?* Washington, D.C.: National Bureau of Economic Research.

Khanna, K., & Imai, K. (2016). *Who are you? Bayesian prediction of racial category using surname and geolocation*. Retrieved July 15, 2016, from: https://cran.r-project.org/web/packages/wru/wru.pdf.

Kukla-Acevedo, S., & Heflin, C. M. (2014). Unemployment insurance effects on child academic outcomes: Results from the National Longitudinal Survey of Youth. *Children and Youth Services Review*, *47*, 246-252.

Lee, J. C., & Staff, J. (2007). When work matters: The varying impact of work intensity on high school dropout. *Sociology of Education*, *80*(2), 158-178.

Leos-Urbel, J. (2014). What is a summer job worth? The impact of summer youth employment on academic outcomes. *Journal of Policy Analysis and Management*, *33*(4), 891-911.

Marsh, H. W., & Kleitman, S. (2005). Consequences of employment during high school: Character building, subversion of academic goals, or a threshold? *American Educational Research Journal*, *42*(2), 331-369.

National Longitudinal Surveys. (2016). *Bibliography*. Retrieved July 15, 2016 from: https://nlsinfo.org/bibliography-start.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved July 15, 2016 from: https://www.R-project.org/.

Staff, J., Schulenberg, J. E., & Bachman, J. G. (2010). Adolescent work intensity, school performance, and academic engagement. *Sociology of Education*, *83*(3), 183-200.

Tabak, R. G., Eyler, A. A., Dodson, E. A., & Brownson, R. C. (2015). Accessing evidence to inform public health policy: A study to enhance advocacy. *Public Health*, *129*(6), 698-704.

Tyler, J. H. (2003). Using state child labor laws to identify the effect of school-year work on high school achievement. *Journal of Labor Economics*, *21*(2), 381-408.

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111-122.

Weiss, D., White, J. M. Stohr, R. A., & Willis, M. (2015). Influencing healthcare policy: Implications of state legislator information source preferences for public relations practitioners and public information officers. *Online Journal of Communication and Media Technologies*, *5*(1), 114.

Winkler, W. E. (1990). *String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage*. Proceedings of the Section on Survey Research Methods American Statistical Association (pp. 354-359).

Woolard, N. (2015). Academic capitalism and the impact on state spending for higher education: Perceptions from members of the Oklahoma State Legislature. *Academy of Educational Leadership Journal*, *19*(3), 319.

Endnotes

1. Each high school exiter is then assigned a random digit identifier that is untraceable to the student but that is consistent across all data sources. After the random digit identifier has been assigned, the SSN is stripped from every student's record. In addition, the data are not stored in a matched state, but can be matched using the random digit identifier, which is shared across all three data systems.

2. All research for this project took place in a data environment with a high degree of electronic and physical security.

3. At the conclusion of the process, all MVC data were deleted from the secure data environment.

4. The matching process was non-greedy, meaning that once a match for an NJ SMART record was found, it was removed from the match pool along with the record with which it matched.

Acknowledgments

The authors of this report were William Mabe, Ph.D., David Seith, and Syeda Fatima. Christine Jenter and Robb C. Sewell provided editorial and graphic design assistance.

About the Heldrich Center

The John J. Heldrich Center for Workforce Development at Rutgers University is a university-based organization devoted to transforming the workforce development system at the local, state, and federal levels. The Center, located within the Edward J. Bloustein School of Planning and Public Policy, provides an independent source of analysis for reform and innovation in policy-making and employs cutting-edge research and evaluation methods to identify best practices in workforce development, education, and employment policy. It is also engaged in significant partnerships with the private sector, workforce organizations, and educational institutions to design effective education and training programs. It is also deeply committed to assisting job seekers and workers attain the information, education, and skills training they need to move up the economic ladder.

As captured in its slogan, "Solutions at Work," the Heldrich Center is guided by a commitment to translate the strongest research and analysis into practices and programs that companies, community-based organizations, philanthropy, and government officials can use to strengthen their workforce and workforce readiness programs, create jobs, and remain competitive. The Center's work strives to build an efficient labor market that matches workers' skills and knowledge with the evolving demands of employers. The Center's projects are grounded in a core set of research priorities:

- Disability Employment
- Education and Training
- Unemployment and Reemployment
- U.S. Labor Market and Industry
- Workforce Policy and Practice
- Work Trends Surveys

Learn more: *www.heldrich.rutgers.edu*