TECHNICAL MEMORANDUM FOR

The Invisible Achievement Gap, Part One

Linkage Between Data

From the California Department of Education and

From the California Department of Social Services

THE CENTER FOR THE FUTURE OF Teaching & Learning at WestEd

Technical Memorandum for *The Invisible Achievement Gap, Part One*: Linkage Between Data From the California Department of Education and From the California Department of Social Services

Vanessa X. Barrat
BethAnn Berliner

THE CENTER FOR THE FUTURE OF Teaching & Learning at WestEd

This technical memorandum is available online at
http://cftl.org/documents/2013/IAG/IAG_TM.pdf

Suggested citation:
Barrat, V. X., & Berliner, B. (2013). *Technical Memorandum for* The Invisible Achievement Gap, Part 1: *Linkage Between Data From the California Department of Education and From the California Department of Social Services.* San Francisco: WestEd.

Contents

List of Figures

List of Tables

There has been limited information about the education experiences, needs, and academic outcomes of California students who are in foster care under the state's child welfare system.[1] The purpose of this study was to produce a baseline portrait of the education status of students in foster care in California, identifying areas in which research, policy, and practice might be strengthened to better meet the needs of these vulnerable students.

Because there are no identifiers that directly link students between the two data systems, this study required a process for matching individual K–12 student records from the California Longitudinal Pupil Achievement Data System (CALPADS), which falls under the authority of the California Department of Education (CDE), with individual client records for clients in foster care from the Child Welfare Services Case Management System (CWS/CMS), which falls under the authority of the California Department of Social Services (CDSS). Both departments had agreed to share data with WestEd to compare, for the first time, the education experiences and academic outcomes of K–12 students in foster care with the general student population and with other student subgroups identified as being at risk academically. Personally identifiable information from the records in each data set was used to match individuals across the two data systems. Once the match was complete, the data were de-identified and used to compare the characteristics and education environment, as well as the academic and graduation outcomes, of K–12 students in foster care to those of the state's K–12 population as a whole and of other at-risk subgroups with documented achievement gaps, specifically low-socioeconomic-status (low-SES) students, English learners, and students with disabilities.

This memorandum is presented in three sections. First, we define the populations of students to be matched and describe how the data were prepared for the matching process. Second, we explain in detail the six-step matching process and matching rates. Third, we specify how the quality of the match was evaluated.

---

[1] California Education Collaborative for Children in Foster Care. (2008). *Ready to succeed: Changing systems to give California's foster children the opportunities they deserve to be ready for and succeed in school. Recommendations and implementation strategies from the California Education Collaborative for Children in Foster Care.* Santa Cruz, CA: Center for the Future of Reaching and Learning. Retrieved from http://www.eric.ed.gov/PDFS/ED501333.pdf

I. Before the Match: Definition of the Populations and Data Preparation

First, the populations of analysis to be linked were defined as school-age students from the CDE dataset and school-age clients from the CDSS dataset. Then, to prepare for the linkage, matching variables were thoroughly examined to evaluate their discriminating power and the presence of compound/hyphenated names. Additional variables available in both datasets were also merged in the analysis dataset and were used to sort out duplicate matches.

1-Populations of analysis

CDE population of analysis

Student-level education data for all students ages 5–17 enrolled in a California public school during school year 2009/2010 were obtained from the CALPADS Statewide Student Identifier (SSID) extract. These data contained information on student demographic, enrollment, and school characteristics. The final population of analysis consisted of 5,969,112 students and was defined as follows:

- Students enrolled in a California school as of October 7, 2009 (census date) in the CALPADS SSID 2009/10 file extract;
- Students ages 5 years old or older at the beginning of the school year (born before August 1, 2004);
- Students younger than 18 years old as of the 2009/10 census date (born after October 8, 1991);
- Students enrolled in adult education regardless of age were excluded;
- Students with no name information registered in the CALPADS extract were excluded; and
- Student records marked as duplicate entries in CALPADS were excluded.

CDSS population of analysis

Individual client records for clients in foster care were obtained from the CWS/CMS. The population of clients in child welfare in California to be matched to the CDE population of analysis was defined as all CDSS clients ages 5–17 with an open placement episode during the 2009/10 school year. The final population of clients in child welfare consisted of 62,274 clients and was defined as follows:

- Clients from the CLIENT_T table extract with an open placement episode between August 1, 2009 and June 1, 2010;

- Clients ages 5 years old or older at the beginning of the school year (born before August 1, 2004);
- Clients younger than 18 years old as of the 2009 census date (born after October 8, 1991);
- Client records showing the agency responsible for placement to be "private adoption," "mental health," or "Kinship Guardianship Assistance Payment Program (Kin-GAP)" were excluded; client records showing the agency responsible for placement to be "child welfare," "probation," and "other agencies" were included for the match (but only "child welfare" placements were later kept for the analysis).
- Clients had to have some name information registered in the CWS/CMS extract (around 300 records with last name= 'CONFIDENTIAL' and/or first name 'ADOPTED' were deleted); and
- A few clients (139 records for 69 clients) had duplicate unique identification numbers for last names, first names, and dates of birth. Those records were further checked by WestEd and CDSS staff and seemed to correspond mostly to real duplicate entries in the CWS/CMS data system. The following decision rule was followed to select only one CWS/CMS unique identifier for those clients:
    1. Select the most recent placement episode start date;
    2. If the records have the same placement episode start dates, then select the record that has ended; and
    3. If the records have the same placement episode start dates and both records have ended, select the one with the latest end date.

2-Characteristics of the matching and control variables

This study used a deterministic and "fuzzy" sequential matching process. Given the absence of a student and client identifier that is common to both data systems, a set of personal descriptors—first name, last name, and date of birth—was used to link across the two data systems, and a match was made if they all agreed.

Matching variables

The characteristics of the matching variables—first name, last name, and date of birth—were examined closely in both datasets to prepare for the match.

- **Discriminating power of the matching fields:** Since CDE data represents the population of students to which we were matching, we examined the specificity of our planned matching variables on this dataset: out of 5,969,112 records in the CDE population, 14,781 combinations of first names, last names, and dates of birth appeared more than one time, representing a percentage of duplicate values on our matching variables of less than a quarter of a percent (0.25 percent). When adding middle name, city of residence, or city of school to sort out the duplicates, we were able to unduplicate virtually all records that had this information available. The rare cases

that could not be sorted out corresponded to real duplicates in the CDE data (the exact same first name, last name, date of birth, city of residence, and city of school).

- **Compound/hyphenated names:** The name fields were evaluated for the presence of compound/hyphenated names (names with two or more words separated by a blank or a special character in the same data field), or for the use of abbreviations, since the presence of several names in a field can pose problems for the match. Examples of matching issues due to compound/hyphenated names are provided below:

Dataset 1

First name	Middle name	Last name
Vanessa	Ximenes	Barrat
Beth-Ann		Berliner
John Jr		Smith

Dataset 2

First name	Middle name	Last name
Vanessa		Ximenes Barrat
Beth	Ann	Berliner
John		Smith

In the examples above, a matching process that uses the first name and last name fields as provided in the dataset would not result in a match for these names. The percentage of compound/hyphenated names in both datasets is presented in table 1.

Table 1. Percentage of compound/hyphenated first and last names in the California Department of Education and California Department of Social Services datasets

	CDE students	CDSS clients
Total number	5,969,112	62,274
With compound/hyphenated first names (e.g., Beth-Ann, Mary Jane, Carl JR, William III)	271,329 (5%)	665 (1%)
With compound/hyphenated last names (e.g., Ximenes-Barrat, Bonham Carter, Pollack-Pelzner)[2]	492,601 (8%)	3,480 (6%)

Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.
Note. CDE = California Department of Education. CDSS = California Department of Social Services.

In cases of compound/hyphenated names, three versions of each name were kept in three separate fields: one corresponding to the name as it was provided with no blank or separator, one storing only the first part (as defined by the presence of a blank or special character) of the compound/hyphenated name, and a third one storing the second part. All fields were used sequentially in the matching process. Table 2 presents examples illustrating the decomposition of compound/hyphenated names.

Table 2. Examples of decomposition of compound/hyphenated first and last names

First Name—original value	First Name	First Name—1	First Name—2
BETH-ANN	BETHANN	BETH	ANN
BETH ANN	BETHANN	BETH	ANN
BETHANN	BETHANN	BETHANN	BETHANN
Last Name—original value	**Last Name**	**Last Name—1**	**Last Name—2**
XIMENES BARRAT	XIMENESBARRAT	XIMENES	BARRAT
XIMENES/BARRAT	XIMENESBARRAT	XIMENES	BARRAT
BARRAT-XIMENES	BARRATXIMENES	BARRAT	XIMENES

[2] Examples given are not real client names from the CDE or CDSS datasets.

Control variables for duplicate matches

In cases where a CDSS client matched to more than one CDE student, the middle name, city of residence, and city of school, when available, were used to unduplicate the data. The additional information—the middle name, city of residence, and city of school—was merged into the analysis datasets as follows:

- A County-District-School (CDS) code uniquely identifying the school in which a student is enrolled is available for all records of the CALPADS SSID file. There are four types of enrollment in CALPADS: primary enrollment, secondary enrollment, short-term enrollment, and receiving specialized services only enrollment. In cases of several active records as of October 7, 2009 in the data extract, the following decision rule was applied to define the main school of enrollment as of the census date:
  - In cases of concurrent records with different types of enrollment, the record used to document the school of enrollment as of the census date (for matching) is selected in the following order of priority: Primary Enrollment, Short-Term Enrollment, and Secondary Enrollment. Students whose sole enrollment as of the census date was receiving specialized services only are not kept in the sample; and
  - In cases of concurrent records with the same type of enrollment, the record used to document the school of enrollment as of the census date (for matching) is selected in the following order of priority:
    - If one or both records was an open enrollment, then the record with the most recent enrollment date is used; and
    - If one of the records had a more recent withdrawal date, then the most recent withdrawal date is used.
- The CDE website lists all schools in the state by CDS code, along with main characteristics of the schools (e.g., address, school size, grade configuration, and school type). This information was merged by CDS code into the CDE population to define the city of the school.
- Residential addresses were merged to the CDE population file to identify, when available, a city of residence. In cases of multiple records for residential or school addresses, the record active as of October 7, 2009 was given priority, followed by the record with the start date closest to the census date.
- In parallel, the CDSS school information table and the CDSS residential addresses table were merged to the CDSS client file to identify, when available, a city of residence and a city of school for the CDSS population file. In cases of multiple records for residential or school addresses, the record active as of October 7, 2009 was given priority, followed by the record with the start date closest to the census date.

The availability of the control variables for the matching process is summarized in table 3.

Table 3. California Department of Education students and California Department of Social Services clients with information on city of residence, city of school, and middle name

	CDE students	CDSS clients
Total number	5,969,112	62,274
With a city of residence as of October 7, 2009	5,456,984 (91%)	59,291 (95%)
With a city of school as of October 7, 2009	5,936,142 (99%)	47,518 (76%)
With a middle name	4,095,049 (69%)	36,212 (58%)

Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.
Note. CDE = California Department of Education. CDSS = California Department of Social Services.

3-Data preparation

All text fields used for the match (names and cities of residence and school) were cleaned before matching to ensure a better match:

- Names fields were transformed into all capital letters.
- City names were corrected for misspellings. Frequencies on the cities of residence and school were examined for CDSS and CDE data and obviously misspelled city names (e.g., LLOS ANGELES or LOS ANGELLES) were corrected. Some common abbreviations (e.g., LA, SF, HGTS for HEIGHTS, VLY for VALLEY, BCH for BEACH) were recoded as well.

II. The Matching Process

The six-step matching process was written in SAS software* and entailed a sequence of deterministic and "fuzzy" matches using SAS SOUNDEX and SAS SPEDIS functions.

1-Description of SAS SOUNDEX and SAS SPEDIS

SAS SOUNDEX

Soundex is an algorithm that codes a name as a short sequence of characters and numerals based on the way a name sounds rather than the way it is spelled. It was originally developed by Robert C. Russell and Margaret K. Odell in 1918. An updated version, the American Soundex, was used in the 1930s for a retrospective analysis of U.S. censuses from 1890 through 1920. The National Archives and Records Administration (NARA) maintains the current set of rules that defines the algorithm for the official implementation of Soundex as used by the U.S. Government. The SAS built-in function SOUNDEX is based on the American Soundex algorithm without the restriction to four characters (see appendices A and B for details about the SOUNDEX and SPEDIS SAS function, respectively).

SAS SPEDIS

SPEDIS is a built-in SAS function that computes the spelling distance between two words as a measure of how close two text strings really are to one another. It is more specific and restrictive than the SOUNDEX operator and was used to restrict the pool of possible matches identified by SOUNDEX to the final set of matches. SPEDIS accepts two arguments, a query and a keyword. It returns the 'spelling distance' required to convert the keyword into the query as the normalized cost of operations used in the conversion process, such as character insertion, deletion, and replacement. SPEDIS evaluates all of the possible ways to translate the keyword into the query and then returns the smallest possible value, rounded down to a whole number. The SPEDIS distance was computed for all potential matches obtained using SOUNDEX. Based on a close examination of those results, a conservative cutoff score of 34 was selected for this study (see operations and "cost" of the SPEDIS function in appendix B).

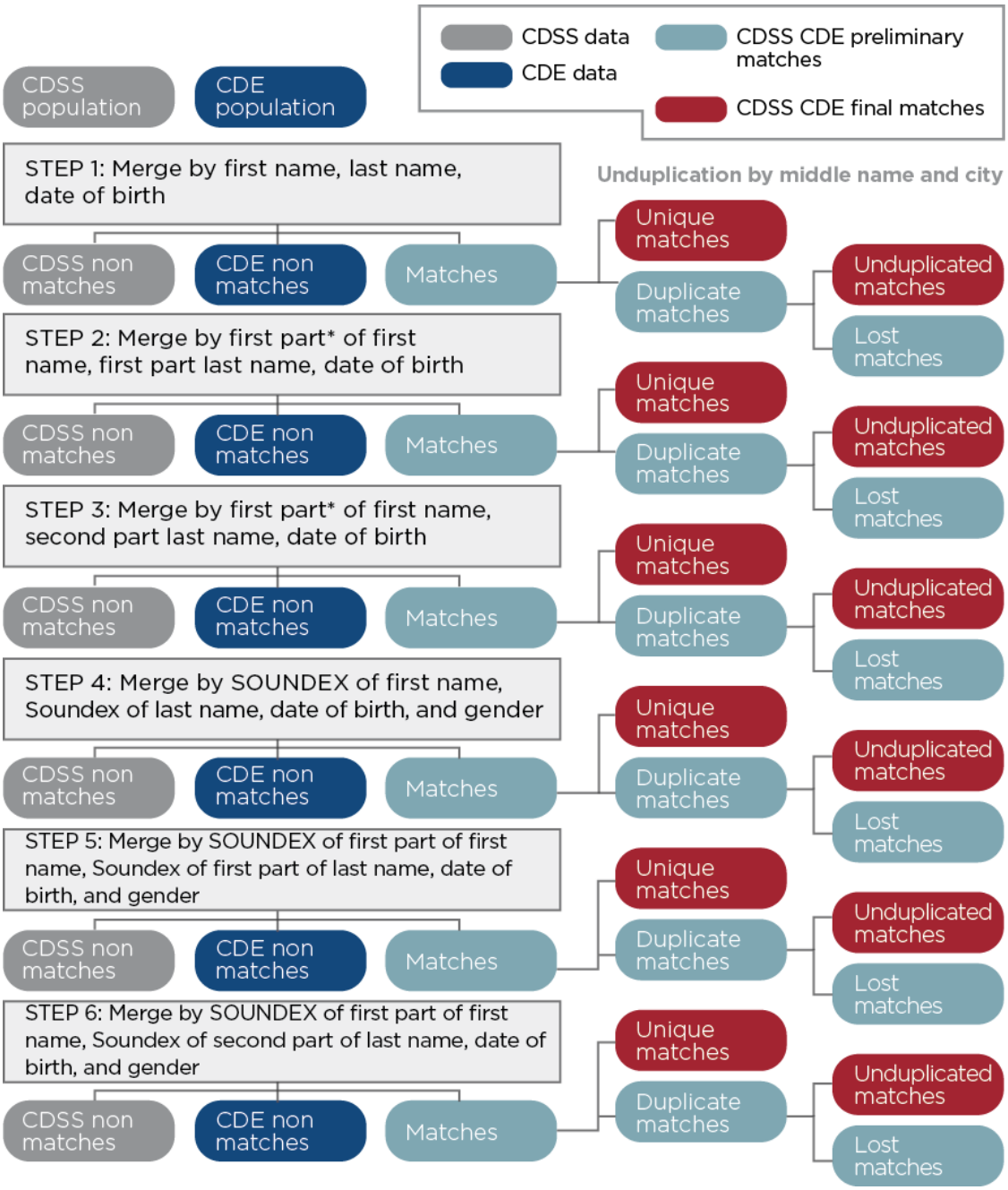
---

* Version 9.3 of the SAS System for Windows. Copyright © 2002–2003 SAS Institute Inc.

2-The Six-Step Matching process

The matching process for this study was conducted by following a sequence of six steps. Figure 1 summarizes the matching process. It is followed by a detailed step-by-step description of the process and the corresponding matching rates.

Figure 1. Overview of the matching process



Note. CDE = California Department of Education. CDSS = California Department of Social Services. * First part differs from the original name only for compound /hyphenated names; in those cases it refers to the first name in the field

Step 1 of the matching process used the exact text strings recorded for first names, last names, and dates of birth to match the two datasets.

Because of the prevalence of compound/hyphenated names (8 percent of the last names in CDE and 6 percent of the last names in CDSS), steps 2 and 3 were structured to capture different combinations for entering compound/hyphenated last names, along with the date of birth. Step 2 of the match used only the first word (as separated by a blank or special character) from the first name and the first word in the last name. Step 3 used the first word in the first name field and the second word in the last name field.

Steps 4, 5, and 6 repeated the sequence of steps 1, 2, and 3, but instead of relying on the spelling of names, steps 4, 5, and 6 used a SOUNDEX transformation on the first and last name fields. The pool of potential matches obtained at each of these steps was further limited by imposing a restriction on the spelling distance between the two names being matched, as calculated by the SPEDIS function.

From one step to the next, only the residual records—those not matched in a previous step—were kept in the pool to be matched in a subsequent step. At each step, the set of CDSS clients who matched exactly to only one student in the CDE dataset were kept as final matches, while the set of CDSS clients for whom there were duplicate matches in the CDE dataset were further studied to be unduplicated. When a CDSS client matched to more than one CDE student, we looked at the city of school, the city of residence, and the middle name to pick the right match. If a one-to-one match could be achieved using the additional information, the record was identified as a final match. If confirming data (i.e., the city of school, the city of residence, and the middle name) were not available for any of the duplicate records, or if the data were available but the information was the same for all duplicates (e.g., same middle name), then we did not unduplicate the data and the CDSS client did not get matched.

The six steps of the matching process are described in detail below:

- *Step 1: Matching by full first name, last name, and date of birth.* The first step of the match used the exact text strings recorded for the first and last name fields after transforming the names in capital letters and deleting all blanks and special characters (so compound/hyphenated names appeared as one long single word; e.g., XIMENESBARRAT for Ximenes Barrat).

- *Step 2: Matching by the first part in the first name, the first part in the last name, and the date of birth.* The second step of the match used only the first word (as separated by a blank or special character) from the first name and the first word in the last name (so only the first name in a compound/hyphenated name was kept; e.g., XIMENES for Ximenes Barrat).

- *Step 3: Matching by the first part in the first name, the second part in the last name, and the date of birth.* The third step is symmetrical to step 2 but used the second word in the last name field (so only the second name in compound/hyphenated name was kept; e.g., BARRAT for Ximenes Barrat).
- *Step 4: Matching by SOUNDEX transformation of the first name, SOUNDEX transformation of the last name, the date of birth, and the gender.* Step 4 matched on a SOUNDEX transformation of first and last names, date of birth, and gender. Gender was added as a matching field because the SOUNDEX function tends to erase the gender specificity of some first names (e.g., Alexandro and Alexandra are coded the same). Only matches with a spelling distance less than a score of 34 as computed with SPEDIS were kept as final matches.
- *Step 5: Matching by SOUNDEX transformation of the first word in the first name, SOUNDEX transformation of the first word in the last name, the date of birth, and the gender*. Step 5 is similar to step 2 but used the SOUNDEX of the first word in the first name and the SOUNDEX of the first word in the last name. Only matches with a spelling distance less than a score of 34 as computed with SPEDIS were kept as final matches.
- *Step 6: Matching by SOUNDEX transformation of the first word in the first name, SOUNDEX transformation of the second word in the last name, the date of birth, and the gender.* Step 6 is similar to step 3 but used the SOUNDEX of the first word of the first name and the SOUNDEX of the second word of the last name. Only matches with a spelling distance less than a score of 34 as computed with SPEDIS were kept as final matches.

3-Matching Rates

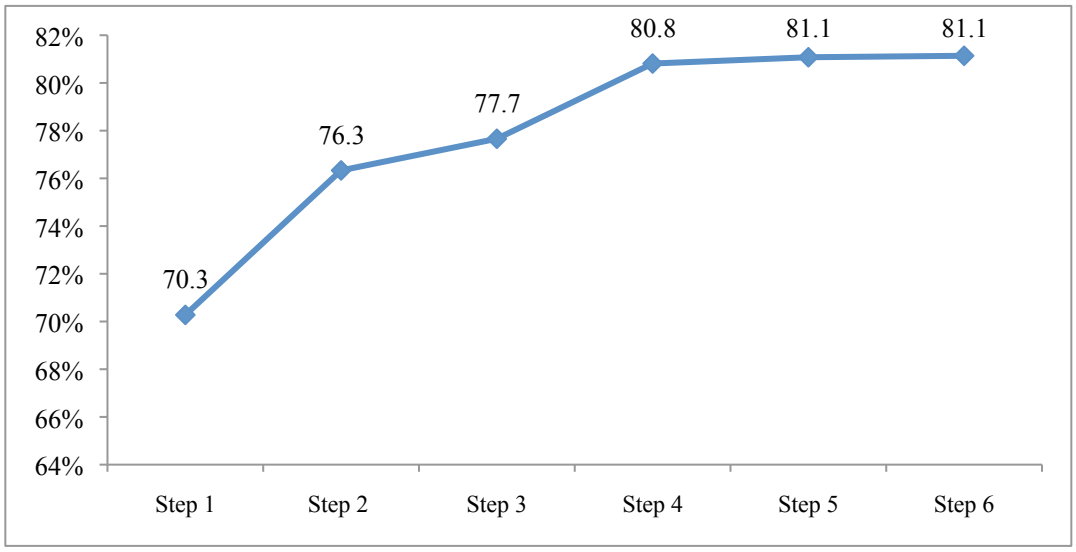
The sample size and match rate at each step of the matching process are presented in table 4.

At step 1, using the exact text strings (after cleaning) recorded for first names, last names, and dates of birth, 43,764 CDSS clients (70.3 percent of the CDSS population) were matched. Decomposing compound/hyphenated names (step 2) allowed the match of an additional pool of 3,772 CDSS clients (6.1 percent of the CDSS population), and the fuzzy match using the first name and the last name as they were provided in the dataset (step 4) resulted in an additional 1,964 CDSS clients (3.2 percent of the CDSS population) being matched. The other steps each provided 1 percent or less of matches.

The final total number of matches was 50,528 out of 62,274 CDSS clients, representing an 81.1 percent match rate.

The cumulative match rate obtained through the matching process is presented in figure 2.

Figure 2. Match rate at each of the steps of the matching process



Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.

Table 4. Number and percentage of matches at each step

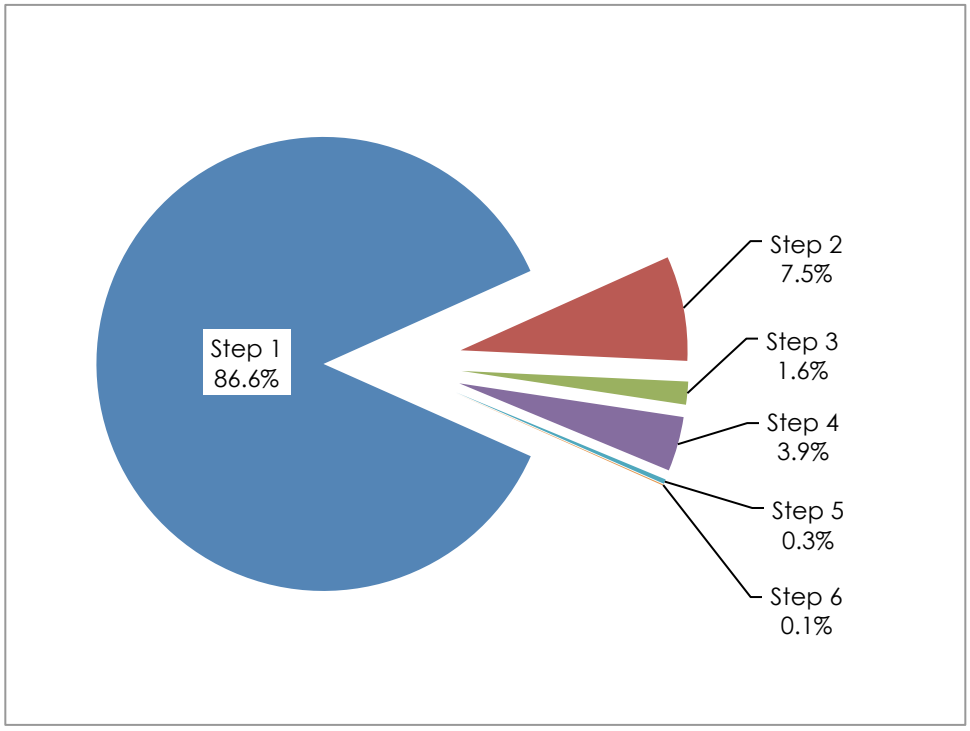
	Number of matches	Cumulative number of matches	Match rate (percentage)	Cumulative match rate (percentage)	Percentage of all matches
Step 1–First name, last name, and date of birth	43,764	43,764	70.3	70.3	86.6
Step 2–First part in the first name, the first part in the last name, and date of birth	3,772	47,536	6.1	76.3	7.5
Step 3–First part in the first name, the second part in the last name, and date of birth	826	48,362	1.3	77.7	1.6
Step 4–SOUNDEX transformation of the first name, SOUNDEX transformation of the last name, date of birth, and gender	1,964	50,326	3.2	80.8	3.9
Step 5–SOUNDEX transformation of the first word in the first name, SOUNDEX transformation of the first word in the last name, date of birth, and gender	164	50,490	0.3	81.1	0.3
Step 6–SOUNDEX transformation of the first word in the first name, SOUNDEX transformation of the second word in the last name, date of birth, and gender	38	50,528	0.1	81.1	0.1
Total	50,528	n/a	81.1	n/a	100

Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.

The number of unique and duplicate matches and the results from the unduplication process are presented in table 5. For example, step 1 provided 44,123 matching records, of which 43,562 corresponded to one-to-one matches and 561 corresponded to one client in the CDSS matching to two or more students in the CDE. For 202 CDSS clients, the unduplication process selected one record since the additional information available led to the selection of only one of the CDE records, but for 149 duplicate records, no additional information matched (61 records) or additional matching information was available for several records (88 records), making the identification of a unique match impossible.

Figure 3 shows that step 1 provided the great majority (86.6 percent) of the matches. The following steps combined, dealing with spelling errors and compound/hyphenated names, added less than 14 percent of all matches. Relaxing some of the strict conditions in the fuzzy match in steps 4–6 might have led to a higher match rate; however, since these matches would be compared to all K–12 students in California, a much larger population, the goal for this first linkage between the two data systems was to limit false-positive errors (i.e., when a match is made between two records that, in fact, do not belong to the same child). When trying to limit this type of error, one might increase the false-negative errors (i.e., when a match is not made between two records that, in fact, do belong to the same child).

Figure 3. Match rate at each step of the matching process



Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.

The matching process developed for this study resulted in a final matching rate of 81.1 percent, which is relatively high and comparable to the match rates of other studies in the field using similar methods. Specifically, in a study of the education experiences of the children in out-of-home care, University of Chicago researchers reported a match rate of 81 percent for students in Chicago.[3] Furthermore, the match rate obtained for this study likely underestimates the real match rate since CDSS clients with delayed entry in school (i.e., starting kindergarten after age 5), dropping out of school, or not attending a public school would not be expected to appear in the CDE data system.

[3] Smithgall, C., Gladden, R. M., Howard, E., Goerge, R., & Courtney, M. E. (2004). *Educational experiences of children in out-of-home care*. Chapin Hall Center for Children at the University of Chicago. Retrieved from Chapin Hall Center for Children website http://www.chapinhall.org/research/report/educational-experiences-children-out-home-care

Table 5. Number of unique and duplicate matches at each step

	Matching records	Unique matches	Number of records corresponding to duplicate matches	Only one record had matching additional information	Several records had matching additional information (lost match)	No record had matching additional information (lost match)	Number of students from unduplicated matches	Number of matched clients
Step 1	44,123	43,562	561	412	88	61	202	43,764
Step 2	3,789	3,767	22	10	8	4	5	3,772
Step 3	829	823	6	6	0	0	3	826
Step 4	1,975	1,957	18	14	2	2	7	1,964
Step 5	165	163	2	2	0	0	1	164
Step 6	38	38	0	n/a	n/a	n/a	n/a	38

Source. Authors' analysis of linked California Department of Education and California Department of Social Services administrative data, 2009/10.

III. Evaluating the Quality of the Match

The quality of the match was evaluated in three steps. Each step is described below and was run iteratively during the matching process, and the results were used to improve the process until we reached the final algorithm for making the match.

- The middle name, the city of school, and the city of residence variables were also examined as quality control for the matches:
  - When full middle names were available in both CDE and CDSS datasets (17,458 of the final matches), they matched at 86 percent (93 percent if looking at SOUNDEX of the middle names); and
  - When only an initial was available as a middle name in either of the CDE or CDSS datasets (6,221 of the final matches), this initial matched at 96 percent with the initial or first letter of the middle name in the other dataset (when available).

  When the cities of school or residence were available, they matched at 80 percent. However, since school information collected through the CDSS system was not systematically collected as of October 7, 2009, cities of school and residence in the CDE and the CDSS could legitimately differ for the same child.[4]
- Additionally, a random sample of 20 matches for each of the six steps (direct one-to-one matches and matches resulting from the unduplication process, totaling 240 records) was evaluated manually to verify the quality of the matches obtained at each step. All matches were legitimate with the final matching algorithm.
- Finally, the characteristics of the matched sample were compared to the characteristics of the unmatched sample and of the whole CDSS population, and subgroups of children under- or overrepresented in the matched sample were investigated. Underrepresented subgroups included clients who were 5 years old as of October 7, 2009; clients who were 17 years old as of October 7, 2009; and clients who were marked as enrolled in non-public schools in the CDSS data. See tables in appendix D for a comparison of the matched sample, the unmatched sample, and the CDSS population by agency responsible of placement, age, types of school, gender, and ethnicity.

[4] One possible strategy to reconcile these differences could be to use a geo-coding tool to evaluate proximity of locations (e.g., Riverside and Moreno Valley).

Appendix A. SOUNDEX Algorithm and SAS SOUNDEX Function

American Soundex Coding Rule

The American Soundex code consists of a letter and three numbers. The letter is the first letter of the name. The numbers encode the remaining consonants with similar sounding consonants sharing the same digit as shown in table A.

Table A. Soundex coding

Number	Represents the Letter(s)
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Letters A, E, I, O, U, H, W, and Y are not coded. Zeros are added at the end if necessary to produce a four-character code. Additional letters are disregarded. If the surname has any double letters, they should be treated as one letter. If the surname has different letters side-by-side that have the same number in the Soundex coding guide, they should be treated as one letter. If a vowel (A, E, I, O, U) separates two consonants that have the same Soundex code, the consonant to the right of the vowel is coded.

Examples:

- **Washington** is coded W252 (W, 2 for the S, 5 for the N, 2 for the G, remaining letters disregarded).
- **Lee** is coded L-000 (L, 000 added).

SOUNDEX function and Sounds-like operator in SAS

The SOUNDEX built-in SAS function is different from the standard American Soundex code. The differences are as follows:

1. SAS SOUNDEX function generates all the possible codes for the string. For example, Washington has a code of W25235 instead of W252. Even if there are spaces in the string, SOUNDEX function will generate the SOUNDEX code for the entire string.
2. No additional zeros will be added. Jackson is coded as J25 instead of J250. There are no supplemental zeros being added to make it a four-character code.

Appendix B. SAS SPEDIS Function

The SPEDIS Spelling Distance can be defined as the "cost" of operations used in the conversion process. These operations include character insertion, deletion, and replacement.

Table B1. SPEDIS operations and costs

Operation	Cost	Explanation
Match	0	No change
Singlet	25	Delete one of a double letter
Doublet	50	Double a letter
Swap	50	Reverse the order of two consecutive letters
Truncate	50	Delete a letter from the end
Append	35	Add a letter to the end
Delete	50	Delete a letter from the middle
Insert	100	Insert a letter in the middle
Replace	100	Replace a letter in the middle
FirstDel	100	Delete the first letter
FirstIns	200	Insert a letter at the beginning
FirstRep	200	Replace the first letter

The costs of the operations are summed and then divided by the length of the query to represent the spelling distance. SPEDIS evaluates all of the possible ways to translate the keyword into the query and then returns the smallest possible value (always rounded down to a whole number).

Table B2. Examples of spelling distances

	First	Last	First_B	Last_B	SPEDIS (First, First_B)	SPEDIS (Last, Last_B)
1	George	Washington	Goerge	Washnigton	8	5
2	George	Washington	greg	Wa.sh	75	32
3	George	Washington	Thomnas	Wasshington	108	2
4	Thomas	Jefferson	Thmas	Jefrson	16	16
5	Thomas	Jefferson	Thomnas	Wasshington	8	100
6	Thomas	Jefferson	TANK	JEEPERS	89	92

Source. Roesch, A. (2011). Matching data using sounds-like operators and SAS® Compare Functions. Educational Testing Service, Princeton, NJ. Retrieved from http://www.nesug.org/Proceedings/nesug11/ap/ap07.pdf

Table B3. Examples of popular first name spelling distances

	First	First_B	SPEDIS (First, First_B)
1	Marc	Mark	25
2	Sophia	Sofia	33
3	Elaina	Elena	33
4	Antonio	Anthony	35
5	Pedro	Peter	50
6	Julio	Joel	75

Appendix C. Matching rates by California Department of Social Services client characteristics

Table C1. Matching rate by agency responsible for placement, 2009/10

Agency	Total	Matched	Match Rate (percentage)
Probation	7,324	5,566	76.0
Child Welfare	54,906	44,928	81.8
Other (Indian Child Welfare, State Adoptions District Office)	44	34	77.3
All	62,274	50,528	81.1

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table C2. Matching rate by student's age as of October 7, 2009

Age (As of 10/7/09)	Total	Matched	Match Rate (percentage)
5	3,237	2,454	75.8
6	3,877	3,151	81.3
7	3,621	2,953	81.6
8	3,570	2,976	83.4
9	3,568	2,950	82.7
10	3,590	3,010	83.8
11	3,664	3,083	84.1
12	3,948	3,347	84.8
13	4,806	4,052	84.3
14	5,821	4,854	83.4
15	7,040	5,757	81.8
16	8,016	6,380	79.6
17	7,516	5,561	74.0
All	62,274	50,528	81.1

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table C3. Matching rate by school category, 2009/10

	Total	Matched	Match Rate (percentage)
Public School	44,740	38,039	85.0
Missing School Information	14,739	10,832	73.5
Private School	2,024	1,148	56.7
Home School	615	416	67.6
Independent School	156	93	59.6
All	62,274	50,528	81.1

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table C4. Matching rate by gender, 2009/10

Gender	All	Matched	Match Rate (percentage)
Female	29,576	23,912	80.8
Male	32,680	26,602	81.4
Missing	18	14	77.8
Total	62,274	50,528	81.1

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table C5. Matching rate by race/ethnicity, 2009/10

Ethnicity	Total	Matched	Match Rate (percentage)
Hispanic	28,830	23,598	81.9
Black	16,010	13,001	81.2
White	14,779	11,836	80.1
Asian	1,634	1,317	80.6
Native American	756	598	79.1
Missing	265	178	67.2
Total	62,274	50,528	81.1

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Appendix D. Comparison of the characteristics of the population of students in foster care, the matched sample, and the unmatched sample

Table D1. Distribution of students by agency responsible for placement for all students in foster care, the matched sample, and the unmatched sample, 2009/10

Agency Responsible for Placement	Percentage for all students in foster care (N=62,274)	Percentage for Matched Sample (N=50,528)	Percentage for Unmatched Sample (N=11,746)
Probation	11.8	11.0	15.0
Child Welfare	88.2	88.9	85.0
Other (Indian Child Welfare, State Adoptions District Office)	0.1	0.1	0.1
Total	100.0	100.0	100.0

Source. Authors' analysis California Department of Social Services administrative data, 2009/10.

Table D2. Distribution of students by age as of October 7, 2009, for all students in foster care, the matched sample, and the unmatched sample, 2009/10

Age (As of 10/7/09)	Percentage for all students in foster care (N=62,274)	Percentage for Matched Sample (N=50,528)	Percentage for Unmatched Sample (N=11,746)
5	5.2	4.9	6.7
6	6.2	6.2	6.2
7	5.8	5.8	5.7
8	5.7	5.9	5.1
9	5.7	5.8	5.3
10	5.8	6.0	4.9
11	5.9	6.1	5.0
12	6.3	6.6	5.1
13	7.7	8.0	6.4
14	9.4	9.6	8.2
15	11.3	11.4	10.9
16	12.9	12.6	13.9
17	12.1	11.0	16.6
Total	100.0	100.0	100.0

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table D3. Distribution of students by school category for all students in foster care, the matched sample, and the unmatched sample, 2009/10

School category	Percentage for all students in foster care (N=62,274)	Percentage for Matched Sample (N=50,528)	Percentage for Unmatched Sample (N=11,746)
Public School	71.8	75.3	57.1
Private School	3.3	2.3	7.5
Home School	1.0	0.8	1.7
Independent School	0.3	0.2	0.5
Missing	23.7	21.4	33.3
Total	100.0	100.0	100.0

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.
Note. School information collected through the CDSS system is not systematically collected as of October 7, 2009.

Table D4. Distribution of students by gender for all students in foster care, the matched sample, and the unmatched sample, 2009/10

Gender	Percentage for all students in foster care (percentage) (N=62,274)	Percentage for Matched Sample (percentage) (N=50,528)	Percentage for Unmatched Sample (percentage) (N=11,746)
Female	47.5	47.3	48.2
Male	52.5	52.7	51.8
Unknown/Missing	0.0	0.0	0.0
Total	100.0	100.0	100.0

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

Table D5. Distribution of students by ethnicity for all students in foster care, the matched sample, and the unmatched sample, 2009/10

Ethnicity	Percentage for all students in foster care (N=62,274)	Percentage for Matched Sample (N=50,528)	Percentage for Unmatched Sample (N=11,746)
Hispanic	46.3	46.7	44.5
Black	25.7	25.7	25.6
White	23.7	23.4	25.1
Asian	2.6	2.6	2.7
Native American	1.2	1.2	1.4
Missing	0.4	0.4	0.7
Total	100.0	100.0	100.0

Source. Authors' analysis of California Department of Social Services administrative data, 2009/10.

References

California Education Collaborative for Children in Foster Care. (2008). *Ready to succeed: Changing systems to give California's foster children the opportunities they deserve to be ready for and succeed in school. Recommendations and implementation strategies from the California Education Collaborative for Children in Foster Care*. Santa Cruz, CA: Center for the Future of Reaching and Learning. Retrieved from http://www.eric.ed.gov/PDFS/ED501333.pdf

Fan, Z. (2004). *Matching character variables by sound: A closer look at SOUNDEX function and Sounds-Like Operator (=).* Retrieved from Proceedings of the Twenty-Ninth Annual SAS® Users Group International Conference website http://www2.sas.com/proceedings/sugi29/072-29.pdf

Goerge, R. M., & Lee, B. J. (2002). Matching and Cleaning Administrative Data. In M. Ver Ploeg, R. Moffitt, and C. Citro (Eds.), *Studies of welfare populations: Data collection and research issues* (pp. 197–219). Retrieved from the U.S. Department of Health and Human Services website http://aspe.hhs.gov/hsp/welf-res-data-issues02/pdf/07.pdf

Roesch, A. (2011). *Matching data using Sounds-Like Operators and SAS® Compare Functions*. Princeton, NJ: Educational Testing Service. Retrieved from the SAS Users Group International 11 proceedings website http://www.nesug.org/Proceedings/nesug11/ap/ap07.pdf

Smithgall, C., Gladden, R. M., Howard, E., Goerge, R., & Courtney, M. E. (2004). *Educational experiences of children in out-of-home care.* Chapin Hall Center for Children at the University of Chicago. Retrieved from Chapin Hall Center for Children website http://www.chapinhall.org/research/report/educational-experiences-children-out-home-care

SOUNDEX function. (n.d.). Retrieved from SAS(R) 9.2 Language Reference: Dictionary, Fourth Edition http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000245948.htm

SPEDIS function. (n.d.). Retrieved from SAS(R) 9.2 Language Reference: Dictionary, Fourth Edition http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000245949.htm

The Soundex indexing system. (2007). Retrieved from the National Archives and Records Administration (NARA) website http://www.archives.gov/research/census/soundex.html

THE CENTER FOR THE FUTURE OF
Teaching & Learning
at WestEd