North Carolina Education Research Data Center
Technical Report # 4: Creating Public Use Student Data Files
June 27, 2005

Overview

The student test data collected by North Carolina schools contain a variety of identifying fields that should allow researchers to follow students within a year as well as across years. However, variations in these identifying fields make it difficult to do this. This report documents the methods the North Carolina Education Research Data Center (NCERDC) employed to link records for students across all ABC test score data beginning with data from the 2003-2004 academic year. The matching procedure for earlier years is documented in Technical Report #2, which also provides more detailed information on cohort definition and match rates.

All matching described below – within year student matching, year-to-year student matching, and ABC-SAR matching – begin with the most conservative matches and progress to the least conservative matches. Additionally, only records with unique values of matching variables can be successfully matched.[1] For example, if the criteria for a given match attempt included only name and date of birth, and there were two John Smiths with the same birth date, then both students would be excluded from this potential match.

The Data

The ABC data are the End of Grade (EOG) and End of Course (EOC) tests with a record for each student and test. The data contain the following identifying fields for each student: social security number (SSN), Last Name, First Name, Birth Date, Grade, School and LEA. Because of test failures, misadministration, and absences, a child can have multiple records for the same test and year. Also, a child can take multiple EOC tests in the same year. In addition to possibly having multiple records each year, children will have many test records over time as they progress through school. The identifying fields associated with each test record for a student can vary in numerous ways, including:

- Different spelling of names
- Nick names
- School systems that assign their own SSN
- Children who do not have a SSN or whose parents do not report the SSN[2]
- Missing fields including SSN, birth date, and first name
- Birth dates with month and day transposed
- Students who skip or repeat a grade
- Students who change schools
- Scanning or data entry errors

[1] In general, if there are any matching errors, the NCERDC prefers a matching protocol that leads to false negatives rather than false positives.

[2] Currently there is no statewide student id, so where schools need to create student ids, they may use ids that are active in other schools and they may assign different ids to the same student from one year to the next.

The School Activity Report (SAR) contains data for all personnel employed by the public school system who have direct student contact in a classroom or non-classroom activity for which a state course code or personnel assignment type exists. The NCERDC has already created a randomized identifier (teacher ID) for each person in this file that allows researchers to follow a teacher over time and across the School Activity Report and Teacher Salary and Licensure data. When possible, this identifier is also linked to instructor name in the ABC data so that researchers can link teacher information to student data.[3]

The purpose of this project is to assign a unique randomized identifier (ID) to all records for the same child within a year ***and*** across all years for which the child has test records; and to assign teacher IDs to instructors in the ABC data.

The Matching Process

Step 1: Within-Year Matching

The first step in the matching process involved identifying student records from the same year. Students had multiple records in the same year under the following circumstances: They took multiple EOC tests, such as English 1 and Geometry; they took EOG and EOC tests the same year, such as an eighth grader who takes both EOG grade 8 math and reading and the EOC Algebra 1 test; or they retook the test due to failure, absence, or test misadministration.

To make these within-year matches, seven identifying fields were used: SSN, Last Name, First Name, Birth Date, Grade, School Code, and LEA. In all cases, LEA had to match, while the other 6 identifiers were allowed to vary. In most cases, only one identifier was allowed to vary, and in some circumstances two identifiers could vary between records. For example, both grade and school could vary simultaneously, whereas social security number and last name could not. For matches between records with variation in name, social security number, or birthdate, the values were required to be near matches.

<u>Birth Date variations</u> – Records were considered matches if their SSN, Last Name, First Name, Grade, School and LEA were the same. Records identified as matches using this method that had missing SSN were further checked to verify that they were the same student.

<u>SSN variations</u> – Records were considered matches if their Last Name, First Name, Birth Date, Grade, School and LEA were the same. Records identified as matches using this method that had missing Birth Date were further checked to verify that they were the same student.
SSN was also allowed to covary with grade and, separately, with school. In these matches, the SSN could only vary to the extent of one different number (i.e., 123456789 and 128456789) or number transposition (i.e., 123456789 and 123546789).

[3] It is important to note that the instructor listed in the End of Grade and End of Course records is the person monitoring the exam, and we do not know whether the exam monitor is the teacher for those students. (For more information, see NCERDC Technical Report #1.)

First Name variations – Records were considered matches if their SSN, Last Name, Birth Date, Grade, School and LEA were the same AND they fulfilled one of the following criteria:

- The first initial of the first name was the same.

  For Example: "**V**ictoria" and "**V**icky"
  "**L**amarr" and "**L**amaar"

- At least 4 letters of the first name in record A was a subset of the first name in record B.

  For Example: "**Beth**" and "Eliza**beth**"
  "**Rook**" and "B**rook**"

- The first initial of the first name in record A was the same as the initial of the middle name in record B.

  For Example: "Roy **J**ason" and "**J**ason"
  "John **A**" and "**A**lex"

- One of the first names was a nickname for the other.

  For Example: "Bill" and "William"

- Based on reviewing names by hand, we determined a scanning error.

  For Example: "Eduardo" and "_du_ardo"

Additionally, first name was allowed to covary with grade. Records were considered matches if:

- Grade varied by only one year

  AND

- Names were determined to be close matches using SAS spedis value comparisons. This procedure assesses the spelling distance between two names.

  For Example: "Asley" and "Ashley"
  "Katherine" and "Katie"

These matches were checked by hand to ensure that false matches were not assigned.

Last Name variations – Records were considered matches if their SSN, First Name, Birth Date, Grade, School and LEA were the same AND they fulfilled one of the following criteria:

- The last initial of the last names was the same.

  For Example: "**R**odriquez" and "**R**odriguez"

- At least 4 letters of one last name was a subset of the other last name.

  For Example: "T**hompson**" and "**Hompson**"

- Based on reviewing by hand, we determined a scanning error.

  For Example: "Neal" and "Meal"
  "Craigwell-Gra" and "Graham"

Additionally, last name was allowed to covary with grade. Records were considered matches if:

- Grade varied by only one year

  AND

- Based on reviewing by hand, we determined a scanning error.

  For Example: "Rivero" and "Rivera"
  "Wilson" and "Wilsqn"

These matches were checked by hand to ensure that false matches were not assigned.

Grade and School[4] variations – Records were considered matches if their SSN, Last Name, First Name, Birth Date, School and LEA were the same.

[4] School can vary within a year due to retests taken at a different school, or when one of the student's EOC tests is given at a different school.

Summary of Within Year Student Matching

Table 1 shows examples of valid matches between records.

Table 1

First	Last	SSN	Birth Date	Grade	School	LEA	ID	Match*
Tim	Smith	123456789	01/01/85	10	370	430	1	
Tim	Smith	123456789	**01/01/95**	10	370	430	1	1
Tim	Smith	**113456789**	01/01/85	10	370	430	1	2
Timothy	Smith	123456789	01/01/85	10	370	430	1	3
T	Smith	123456789	01/01/85	10	370	430	1	3
Jim	Smith	123456789	01/01/85	10	370	430	1	3
Tim	**S**	123456789	01/01/85	10	370	430	1	4
Tim	**Smyth**	123456789	01/01/85	10	370	430	1	4
Tim	Smith	123456789	01/01/85	**11**	370	430	1	5
Tim	Smith	123456789	01/01/85	10	**355**	430	1	5
Tim	Smith	123456789	01/01/85	**11**	**355**	430	1	5
Timothy	Smith	123456789	01/01/85	**11**	370	430	1	6
Tim	**Smyth**	123456789	01/01/85	**11**	370	430	1	7
Tim	Smith	**113456789**	01/01/85	**11**	370	430	1	8
Tim	Smith	**113456789**	01/01/85	11	**355**	430	1	9

* Match flag: 1 – Birth Date variation
2 – SSN variation
3 – First Name variation
4 – Last Name variation
5 – Grade and/or School variation
6 – First Name and Grade variation
7 – Last Name and Grade variation
8 – SSN and Grade variation
9 – SSN and School variation

Priority in the case of multiple variations – In some cases variations suggest that values for multiple variables may be erroneous. In such cases, the NCERDC based matches on the most likely true match. For example, among four records with variation in first name and SSN, ids would be assigned as follows, under the assumption that name is more likely than SSN to be accurate (Table 2).

Table 2

Record #	**First**	**Last**	**SSN**	**Birth Date**	**Grade**	**School**	**LEA**	**ID**
1	Timothy	Smith	123456788	01/01/85	10	370	430	1
2	Timothy	Smith	123456789	01/01/95	10	370	430	1
3	Tracy	Smith	123456788	01/01/85	10	370	430	2
4	Tracy	Smith	123456789	01/01/85	10	370	430	2

Records 1 & 3 and 2 & 4 would be matches because only first name varies and the students have the same first initial. Records 1 & 2 and 3 & 4 would be matches because only SSN varies. This second match is preferred by the NCERDC, therefore records 1 and 2 are assigned the same id, as are records 3 and 4.

Step 2: Year-to-Year Matching

In the next step, student records were matched across years. Students were matched to the three previous years, with more recent matches taking precedence over older ones (i.e., 2004 students were first matched to 2003, then 2002 and 2001). This longitudinal matching allows students who exit the North Carolina Public School System for up to two years prior to returning or who do not take tests in two years to remain in the longitudinal database. Unique IDs assigned in previous years were attached to matching records in the current year. Students who were not matched to previous years' data were assigned a new unique ID.

In all cases, grade was allowed to vary since, in most cases, a student's grade will change from one year to the next. In all cases, the School and LEA were also allowed to vary to accommodate students who changed schools from one year to the next. This accommodation is particularly important for students changing from elementary to middle school, and then from middle to high school and for those who changed residence or changed schools for some other reason.

The remaining 4 identifying fields were used for matching: SSN, Last Name, First Name, and Birth Date. If records matched by all 4 identifying fields, they were considered a match. Additional matches between records required that 3 of the 4 identifiers were the same AND the records fulfilled the following criteria:

- The 3 fields used for matching were required to have non-missing values.
- All grades associated with the two records were required to be the same or consecutive.
- For last name and first name matches, the varying fields were required to be "close" matches. See below for how "close" was determined for each varying field.

<u>SSN variations</u> – Someone without a SSN who changed schools or districts would receive a new SSN that did not correspond at all to the original one. In LEAs that did not collect SSN, only allowing SSN to vary permits tracking students who changed schools within the LEA or who moved into or out of that LEA. In this instance, all other matching variables had to be identical and non-missing, and grades had to be identical or consecutive.[5]

<u>Birth Date variations</u> – Records matching by first name, last name, and social security number were considered matches. All three identifiers were required to be non-missing, and only unique records from previous years were used as potential matches.

[5] Because SSN is allowed to vary if all other fields match, false positives may occur if two different people have the same full name, grade, and birth date. However, such errors should occur at random. If we force SSN to be "close," students without SSNs and those in LEAs that do not collect SSN will systematically be excluded from any longitudinal analyses. In this case, errors will not occur at random. Therefore, SSN can vary in the year-to-year matching.

First and Last Name variations –Names were considered "close" if they were within a SAS spedis matching distance of 50 or less, one name was a substring of the other, or one name was a nickname for the other.

- Spedis matches allow for several different types of minor variation, including:
  - A letter is omitted or an extra letter is added.

    For Example: "Alexander" and "Alexandr"
    "Brown" and "Browwn"
  - One name is a subset of the other

    For Example: "Chris" and "Christopher"
    "Jones" and "CampbellJones"
  - They were different due to transposing.

    For Example: "Johnathan" and "Jonhathan"
    "Smith" and "Smiht"
  - Two letter variation

    For Example: "Krystal" and "Krystle"
    "Wilkinson" and "Wilkerson"
- Names were also considered matches if one was a nickname of the other name.

  For Example: "Billy" and "William"
  "Becky" and "Rebecca"
- Finally, names were counted as matches if the full name was transposed such that last name was entered as first name and first name as last name.

  For Example: "Sara Thompson" and "Thompson Sara"

Summary of Year to Year Student Matching

Table 3 shows an example of how student records with variations in some identifiers were assigned the same ID across all records in all years.

Table 3

Record	Reason	Year	Grade	First	Last	SSN	Birth Date	School	LEA	ID
A		1996	4	Timothy	Smith	123456789	03/15/86	370	430	1
B	1	1997	5	Timothy	Smith	1234567**98**	03/15/86	370	430	1
C	2	1998	6	Timothy	Smith	123456789	03/15/86	**240**	**655**	1
D	3	1998	6	Timothy	Smith	123456789	**03/1/86**	240	655	1
E	4	1999	7	Timothy	**Mith**	123456789	03/15/86	240	655	1
F	5	2000	8	Timothy	Smith	123456789	03/15/86	240	655	1
G	6	2000	8	Timothy	Smith	123456789	03/15/86	240	655	1
H	7	2001	9	**Tim**	Smith	123456789	03/15/86	**610**	**122**	1
I	8	2001	**10**	Tim	Smith	123456789	03/15/86	610	122	1
J	9	2002	10	Tim	Smith	543219876	03/15/86	**420**	**600**	1
K	10	2002	10	Tim	Smith	543219876	03/15/86	**384**	600	1
L	11	2004	12	**Smith**	**Timothy**	543219876	03/15/86	**351**	**920**	1

Reason for matching:

1 – Timothy's SSN was entered incorrectly. As all other components match Record A, this record is considered a match.

2 – Timothy changed from elementary to middle school. School and LEA are allowed to differ from one year to the next, so record C matches record B.

3 – Timothy had to re-take his 6th grade EOG test. Although birth date in Record D does not match that in Record C, all other components match.

4 – In Record E, Last Name and Birth date differ from Record D; however, when comparing E to C, Last Name fulfills the "close" criteria, and all other components match. Therefore, this E matches C and links to D.

5 – Record F is Timothy's 8th grade end-of-grade test record. It matches to his Record E because the Last Name fulfills the "close" criteria. It also matches completely the Record G, his end-of-course test (see #6).

6 – This is Timothy's first end-of-course test record for a course he took in 8th grade.

7 – Timothy changed from middle to high school, and now prefers to be called “Tim.” LEA and school code are allowed to differ, and because “Tim” is a nickname for “Timothy”, the First Name difference fulfills the “close” criteria, so H matches G.

8 – In Record I, Tim is listed as being in 10th grade. This grade level could be a data entry or scanning error, or he may have retaken this test in the summer as a rising 10th grader. Grade is allowed to vary within year, so Record I matches Record H.

9 – In 2001, Tim’s family moved to Charlotte, which does not track SSN and assigns a new SSN to him. As every other component is a perfect match and grades are identical or consecutive, Record J matches Record H.

10 – One of Tim’s end of course tests was administered at a different school. School is allowed to vary within year, so Record K matches Record J.

11 – Timothy did not take an exam in 2003, so record L was compared to record K. First and last name are transposed, and school and grade are allowed to vary, so this is a match.