North Carolina Education Research Data Center
Technical Report # 2: Creating a Longitudinal Student Database
May 20, 2003

Overview

A longitudinal student database permits innovative analyses of issues children face as they progress through school. Such a database allows defining cohorts of students and studying their educational careers. With this approach, one can evaluate the impact of educational transitions, such as changing schools, and that of new educational policies on student achievement.

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.

Cohort Definitions and Matching Rates

Due to attrition, we do not expect to match all students from one year to the next. However, with these longitudinal matches, one can define cohorts in various ways. These are examples of some cohorts and matching rates.

- The first year of available End of Grade data file is 1995. We can match 60% of the 3rd graders in 1995 to their 8th grade test scores in 2000.
- The ABC of Accountability System began in elementary and middle schools in 1997. We can match 68% of the 4th graders in 1997 to their 9th grade English 1 test data in 2002, thus capturing their transitions to middle and high school.
- The ABC of Accountability System included End of Course subject tests in 1998. In 1998, we can match 20% of the 8th graders to their Algebra 1 End of Course test data that year and 34% of them to Algebra 1 End of Course test data in the following year, thus defining tracks of students by the timing of their math courses.
- Defining cohorts of high school students presents additional challenges because students vary in whether and when they take different subjects. For example, most students never take physics, and those who do may take it in 10th, 11th, or 12th grade. If we cannot find a student in 11th or 12th grade that may mean the student did not take any classes with End of Course tests, left the North Carolina public school system but is enrolled elsewhere, or dropped out of school. Of the 9th graders in 1999, we have the following match rates for subsequent grades and years: 84% of them as 10th graders in 2000, 71% as 11th graders in 2001, and 33% of them as 12th graders in 2002.
- The student identifiers permit tracking students who were retained. For example, of the 1995 3rd graders, 1.5% were retained; of the 2001 3rd graders, 3% were retained, and this difference may have resulted from policies ending "social" promotion for 3rd graders.

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[1]
- 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

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.

The Matching Process

Step 1: Within-Year Matching

The first step in the matching process involved identifying student records from the same year and type of test (EOG or EOC). Students had multiple records in the same year under the following circumstances: They took multiple EOC tests, such as English 1, Geometry, and US History; 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. Matches between records required that 6 of the 7 identifiers were the same. In all cases, LEA had to match, while the other 6 identifiers were allowed to vary – one at a time. 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	

[1] 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.

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
Tim	**mith**	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	**314**	430	1	6

* Match flag: 1 – Birth Date variation
2 – SSN variation
3 – First Name variation
4 – Last Name variation
5 – Grade variation
6 – School variation

Birth Date variations – 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.

SSN variations – 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.

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"

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 the last name in record A was a subset of the last name in record B.

  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"

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

School variations[2] – Records were considered matches if their SSN, Last Name, First Name, Birth Date, Grade and LEA were the same.

Step 2: Year-to-Year Matching

In the next step, student records in consecutive years (*e.g.* 1998 and 1999) were matched. 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:

[2] 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.

- 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 birth date, 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.

Birth Date variations – Birth Dates were considered "close" if two of the components (month, day, or year) were the same, OR if the month and day components were transposed. Table 2 shows examples of Birth Dates that were considered "close."

Table 2

Birth Date 1	Birth Date 2	Match?	Reason
06/12/1985	09/12/1985	Yes	Day and Year the same
04/23/1988	04/03/1998	Yes	Month and Year the same
07/25/1985	07/25/1986	Yes	Month and Day the same
10/04/1992	04/10/1992	Yes	Month and Day transposed
11/13/1993	12/31/1993	No	Only Year the same

First Name variations – First Names were considered "close" if:

- They were different by only one letter, whether inserted, deleted, or replaced.

  For Example: "Cassandr**a**" and "Cassandr**ia**"
  "Mat**t**hew" and "Mathew"
  "Lat**a**sha" and "Lat**i**sha"

- One of the names was a complete subset of the other name.

  For Example: "**Chris**" and "**Chris**topher"

- They were different due to transposing.

  For Example: "Do**mn*oe***a" and "Do**nm*eo***a"

- One of the names was a nickname of the other name.

  For Example: "Billy" and "William"
  "Becky" and "Rebecca"

- One of the names was a name suffix (*e.g.* "Jr", "III", or "Sr"), which indicated a scanning error.

Last Name variations – Last Names were considered "close" if:

- They were different by only one letter, whether inserted, deleted, or replaced.

  For Example: "Robert**s**" and "Robert**ss**"

- One of the names was a complete subset of the other name.

  For Example: "**Mc**" and "**Mc**Call"
  "**Lark**" and "C**lark**"

- They were different due to transposing.

  For Example: "Sant**ia**go" and "Sant**ai**go"

SSN variations – 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.[3]

Step 3: Assignment of ID

In the final stage of the matching process, all of the within-year matches and the longitudinal matches from year to year (*e.g.* 1998 to 1999) were combined across all years. Since variations in the identifying fields were accommodated in the previous steps, no additional variations were allowed in this final step. The EOC/EOG records within a year (step one) were combined with the longitudinal matches between consecutive years (step 2). Then, consecutive longitudinal matches were combined using the identifiers from the shared year. The 1995/1996 longitudinal matches were combined with the 1996/1997 using the 1996 identifiers. Then, the 1997/1998 longitudinal matches were added using the 1997 identifiers. And so forth until all combinations of all identifiers across all years were added to a master file of identifiers.

A unique ID was then attached to each set of identifiers. 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.

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

Table 3

Record	Reason	Year	Grade	First	Last	SSN	Birth Date	School	LEA	ID
A		1995	4	Timothy	Smith	123456789	03/15/86	370	430	1
B	1	1996	5	Timothy	Smith	123456**98**	03/15/86	370	430	1
C	2	1997	6	Timothy	Smith	123456789	03/15/86	**240**	**655**	1
D	3	1997	6	Timothy	Smith	123456789	**03/1/86**	240	655	1
E	4	1998	7	Timothy	**Mith**	123456789	03/15/86	240	655	1
F	5	1999	8	Timothy	Smith	123456789	03/15/86	240	655	1
G	6	1999	8	Timothy	Smith	123456789	03/15/86	240	655	1
H	7	2000	9	**Tim**	Smith	123456789	03/15/86	**610**	**122**	1
I	8	2000	**10**	Tim	Smith	123456789	03/15/86	610	122	1
J	9	2001	10	Tim	Smith	543219876	03/15/86	**420**	**600**	1
K	10	2001	10	Tim	Smith	543219876	03/15/86	**384**	600	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.