North Carolina Education Research Data Center
Technical Report #1: Linking Teachers in the ABC Data to Teachers in the School Activity Report.
April 28, 2003

Data collected by schools do not provide a direct link between teacher and student; however, such a link will permit innovative analyses of the relationships between teachers and students. This report documents the methods the North Carolina Education Research Data Center (NCERDC) employed to link the instructors in the ABC test scores to the personnel file in the School Activity Report. Using the method described below, the NCERDC obtained (approximately) a 95% matching rate.

The Data

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. Activities include traditional academic classes as well as non-class events, such as study hall, lunch period, or counseling. This file includes activities that meet all year, such as English 1, and those that meet for only part of the year, such as a month-long drama workshop. The data for the personnel include fields for last name, first name, middle name, and social security number as well as the district and school in which that person worked.[1] The NCERDC has already created a randomized identifier 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. The purpose of the project described here is to assign that identifier to the instructor name in the ABC data so that researchers can link teacher information to student data.

The ABC data are the End of Grade and End of Course tests with a record for each student. The data include a field called "instructor name" which is the name of the person monitoring the exam. ***We do not know whether the exam monitor is the teacher for those students. Although that is probably the case for many monitors, there may be other instances where, for example, all the eighth graders take a test together, or where a fourth grader takes math in a class from someone other than her homeroom teacher. Additional validation needs to be done by the NCERDC and by individual researchers.*** However, matching these files is the only way we have to begin to link teachers and students. This matching is an essential first step in the process.

The Matching Process

On the ABC test forms, instructors did not have guidelines in completing the instructor name field, so the format of the names has a great deal of variation (last name-first name; first name-last name; last name-nickname; last name-middle name and so forth). This dataset has only one field for name, did not distinguish between first and last names, and some scanning errors did occur. Due to these variations, the NCERDC used many different methods for matching these files. The basic approach is as follows: First,

[1] One can reliably link the SAR data to the Licensure data because the social security number in SAR is validated against that in the Licensure data.

the SAR and ABC data are limited to one record per LEA, school code and instructor name. This is the level at which the data are matched. Then, the NCERDC started with the most conservative match (e.g., full name), set those names aside into a matched data set and excluded them from further iterations of matches. With the names that did not match the first time, the NCERDC tried to match using other information (e.g., last name, first initial), and set those matches aside, and so forth. See below for more detail.

After matching these names, the NCERDC reviewed all of the matches to ensure that the same teacher id was not assigned to different people in the ABC data and created a matchtype variable that indicates the method used to identify the match. With this variable, researchers can choose whether to accept all methods of matching names.

Those That Will Not Match

Not all schools in the ABC data are in the SAR. For example, SAR does not include DHHS schools, DOJ schools, and many charter schools. On the ABC form, some people left the *instructor name* field blank or completed it with something other than a teacher's name such as "Visitor" "Grade 8" or "Longview School." Furthermore, because schools complete the SAR in the fall and collect the ABC data in the spring, teachers who transferred schools or changed their names during the year will not match.

Finding Matches

All matches are done within fiscal year, LEA, and school code. The NCERDC standardized all names by capitalizing all letters, removing punctuation, such as commas, apostrophes, and hyphens, and removing suffixes, such as Jr., Sr., II, and III. Because the ABC instructor name is only 15 characters, many first names were truncated, and many instructors reported only the first initial of the first name. Therefore, most matches employ first initial of first name rather than the entire first name.

Matchtype 1: Last name.

Because many instructors only reported their last names, we matched the *instructor name* in ABC to the *last name* in SAR.

Example:[2] "PUCKETT" matches "PUCKETT" "PAUL"

Matchtype 2: Last name, first initial of the first name.

Many instructors reported their last name and first name, while others reported only last name and first initial. Here we separate *instructor name* in ABC into *last name* and *first initial* and match that to *last name* and *first initial* in the SAR. This method identifies the most matches.

Example: "TUPPER R" matches "TUPPER" "ROBERT"
"WILLIAMS HANK" matches "WILLIAMS" "HENRY"

[2] All examples show the original format of the names in ABC and SAR with the ABC *instructor name* listed first and the SAR *last name* and *first name* listed second. The quotation marks designate separate variables in the databases. The ABC *instructor name* is noted as one variable, as it is in the ABC data. *Last name, first name, and middle name* (if applicable) are noted as separate variables, as they are in SAR.

"CALLAGHAN LISA" matches "CALLAGHAN" "LISA"

Matchtype 3: First five characters of last name, first initial of first name.

This type of match is primarily in response to spelling or typographical errors in the name and names that are subsets of each other (i.e., hyphenated last names). Here we limit the *last name*s in both data sets to the first five characters (e.g., JOHNSON becomes JOHNS) and match the *five-character last name* and *first initial of the first name* in SAR to those variables in the ABC.

Example: "JOHNSON TRACI" matches "JOHNSTON" "TRACI"
"ONEIL M" matches "ONEILL" "MARY"

Matchtype 4: Last name, first initial of middle name.

Some instructors use their middle name rather than their first name.

Example "BROSNAN JOHN" matches "BROSNAN" "EDWARD" "JOHN"

Matchtype 5: First five characters of last name, first initial of middle name.

As in the third match, this match is for people who use their middle names and who have had an error in typing their last name.

Example: "GIFFORDBAKER B" matches "GIFFORD BAKER" "SUE"
"BEVERLY"

Matchtype 6: Last name and first initial of nickname.

Some instructors reported nicknames. Assigned initials for nicknames as follows:
Elizabeth, Elisabeth, Robert, and William became "B" as the first initial.
Virginia and Eugene received "G" as the first initial
Patricia and Anthony received "T" as the first initial
Margaret received "P" as the first initial

Example: "CLINTON WILLIAM" matches "CLINTON" "BILL"

Matchtype 7: First five letters of last name, first initial of nickname.

This match follows the logic of matches 3 and 5.

Example: "BROWNE JIM" matches "BROWN" "JAMES"

Matchtype 8: Last name, first initial of nickname after recoding Elizabeth.

Many nicknames are derived from Elizabeth. Match 6 picked up those who use Beth or Betty. This match recodes the first initial of Elizabeth to "L" to pick up Liza and Libby.

Example: "GILLESPIE LIBBY" matches "GILLESPIE" "ELIZABETH"

Matchtype 9: Concatenating last name and first name in ABC and matching that to the last name in SAR.

There is variation in the way that names' internal punctuation (hyphens, apostrophes) are reported. For example, O'Brien can become O Brien, and Howrey-Peek can become Howrey Peek. In these examples, the computer reads O and Howrey as the *last names*, and Brien and Peek as the *first names*. To pick up these cases, the NCERDC concatenated *last name* and *first name* in the ABC file and matched those to the *last name* in the SAR file.

Example: "O BRIEN" matches "OBRIEN" "JOANNE"
"HOWREY PEEK" matches "HOWREY-PEEK" "KRISTEN"

In the next three matches, the NCERDC switched the order of the first and last names in the ABC data because some instructors reversed the order in which they reported their names. (While most reported last name and then first name, some reported first name and then last name.)

Matchtype 10: Switching first and last names in the ABC data and match that to last and first name in SAR.

Example "ELLEN CASE" matches "CASE ELLEN."

Matchtype 11: Switching first and last names in ABC data and matching that to last name and first initial in SAR.

Some instructors reported their names by first initial and then last name.

Example: "C THEEMAN" matches "THEEMAN" "CHRISTINA"

Matchtype 12: Switching first and last names in ABC data and matching that to first five characters of last name and first initial in SAR.

Because instructor name was only 15 characters, the last name was often truncated if someone completed the record as first name then last name.

Example: "CHRISTINA THEEM" matches "THEEMAN" "CHRISTINA"

Double-checking

After setting all of these matches together, the NCERDC checked for matching errors, such as cases in which an identifier was assigned to more than one instructor, or where the matching method incorrectly assigned an identifier to an instructor.

Because *first name* in the ABC data was often limited to the first initial or otherwise truncated, many matches were identified using first initial rather than first name. In some cases, the ABC *instructor name* did include more information about the first name. For those instances, if the ABC *first name* matched the SAR *first name* the match was kept. If not, names were reviewed by hand. If the matches met the above criteria (e.g., the ABC *first name* was a nickname, or it had a scanning or typographical

error), the matches were kept. As with the other hand matches, if more than two characters of the first names differed, matches were excluded.

Example:
The match “MARTIN DOTTIE” and “MARTIN” “DOROTHY” was **kept**
The match “RODGERS MJ” and “ROGERS” “MARY” “JANE” was **kept**
The match “CLARK SALLY” and “CLARK” “SARAH” was **excluded**.

Second, the NCERDC excluded any duplicate matches where an id was assigned to multiple teachers within the same school. In these cases, there is no way to distinguish between these two teachers.

Example:
“JONES M” could match “JONES” “MARY” and
“JONES” “MICHAEL”
Both names are **excluded** from the match file.