Creating Linked CCD Data to Improve the Quality of Elementary/Secondary Sample Surveys

Stephen R. Wenck, Albert C.E. Parker, Sameena M. Salvucci, Synectics for Management Decisions,
Carl Schmitt, National Center for Education Statistics
Stephen R. Wenck, Synectics, Suite 900, 1901 N. Moore St., Arlington, VA 22209

Introduction

The National Center for Education Statistics (NCES) has been collecting data about the approximately 15,000 public school agencies (school districts) and the approximately 85,000 public schools as part of the Common Core of Data (CCD) for the past 15 years. These data form the sample frame for all of NCES' elementary/secondary school sample surveys. In order to improve the coverage and efficiency of their elementary/secondary school sample surveys, the quality of the sample frame—the CCD—must be improved and made more consistent. NCES also wishes to make CCD school data for the 1986–87 to 1997–98 academic years available to researchers, public policy analysts, public officials, and the general public in an on-line electronic format in which comparisons across schools, districts, or years can be made easily.

At the school level, the data include:

- Number of teachers (full time equivalent);
- Number of students by grade level;
- Grade levels served by the school;
- Number of students by racial or ethnic (Hispanic) category;
- Number of students eligible for Federally subsidized free lunches.

To be of maximum usefulness, the data should be presented in a consistent manner and should be as correct as possible, without obvious errors arising from misreporting or incorrect data entry. Data for a large set of school districts (about 90 percent of all districts) have been extensively edited. Imputation has been done for districts where data are missing or appear unreasonable. At the school level, however, no editing or imputation has so far been done. In an earlier investigation of the quality and consistency of the data on numbers of students and teachers. The preparatory work was carried out on a data set limited to four types of school districts within the 50 states and the District of Columbia:

1. Local school districts that are not components of supervisory unions;
2. Local school district components of supervisory unions sharing a superintendent and administrative services with other local school districts;
3. Supervisory union administrative centers or county superintendents serving the same purposes as supervisory unions; and
4. Regional education service agencies or county superintendents serving the same purposes as supervisory unions.

This remainder of the paper will discuss the steps we have taken, and those we are still in the process of completing, in developing a multi-year linked CCD data set.

Step 1. Preparation of Data for Linked Datasets

This activity included acquiring all the necessary data and converting the data to a common format. These activities were completed before any changing of the data by editing or imputation.

TECHNICAL APPROACH

Review what schools to include
Our next step was to determine the schools to include in the time series data. We agreed that these data would cover the 1986 to 1997 school years. Over time, schools do occasionally change school districts. Therefore it is possible that a school in a district in 1986 would not still be in a district in 1995.

In addition to reviewing the inclusion and exclusion of school districts with all of their schools, we added and deleted individual schools as necessary. In our preparatory work, we determined that 662 schools were candidates for deletion. These schools appeared only once in the ten-year time span. There was an interruption of reporting of one year or more at another 80 schools. These 80 schools were candidates for addition in the skipped years. In all 742 cases, we worked with NCES to determine which schools to add or delete.

Add master school ID field to all school records
We then created a consistent ID number for linking across all 10 years. This ID is a composite ID (state FIPS number, state-assigned district ID, and state-assigned school ID) for the school in the year that the school was first reported. We maintain this ID for all subsequent years. However, each annual record also includes the ID assigned to the school for that year. Our

preparatory research found numerous instances of changes in district and school ID. The current annual identifying information is maintained for matching to outside data that use it as an identifier.

Make variable names consistent
Next, the variable names were made consistent across all years. Since we are creating a rectangular data set, to compare school characteristics across time, the variable names must be made the same. Currently there is a year identifier on each variable (e.g., "member89" for school enrollment in 1989).

Make special codes consistent
Related to consistent variable names are consistent "special" codes. These codes designate data that are missing, not reported, or not applicable. It is beneficial for the numeric codes for these situations to be consistent across data fields, to facilitate identification of codes that must be treated specially during tabulation or other statistical manipulation of the data. We developed a consistent scheme for these codes and converted all existing codes to the scheme. The scheme was based on existing codes to minimize the amount of conversion required.

Determine which schools have ID number shifts and settle on a consistent ID to use throughout
Once the schools to be included were determined, we also reviewed the consistency of their ID numbers. During the preceding task we found cases that appeared to be the same school (same name, city, and state) with different ID numbers. We identified all such cases and made sure that they have the same ID number in all years. The initial ID number was carried across all years even if the state-assigned ID number changed during the coverage period.

Add fields to all school records for imputation flags
Once the master file was created, we added fields to all school records in which to place flags indicating that data have been imputed. The flags differentiate between data that were imputed because they were missing or not reported and data that were imputed to replace numbers that were implausible or unreasonable. All flags were set to "not imputed." Flags for cases that were imputed would be changed in later steps.

Step 2. Editing and Imputing of School Data

At this point, we had a consistent set of schools, with consistently labeled data, across a period of 11 years. This dataset would be sufficient for a data warehouse or any time series analysis, but we knew from our earlier investigation that there were anomalies in the data that suggested reporting or calculation errors.

TECHNICAL APPROACH

Find student total anomalies, FTE teacher anomalies, and review
We first examined the number of full-time equivalent (FTE) teachers. We looked at total enrollment and FTE teachers together because large accompanying changes are more plausible than large changes in the number of teachers only. Preparatory work included development of a mathematical definition of longitudinal anomalies—what we called a difference measure.

$$D_n = (S_{n+1}/T_{n+1} - S_n/T_n)/S_n/T_n)$$

where D_n = the difference measure for two years, n and $n + 1$, S = the number of students in year n or $n + 1$, and T = the number of teachers in year n or $n + 1$. We investigated the distribution of the difference measure for some of the years and found that about one percent of all the schools in the 1986–97 files showed anomalies over three-year periods that warranted further investigation.

Where the number of students or teachers was inconsistent with other data or with adjacent years, we calculated district-wide totals to determine the likelihood of a data entry error as the cause. For example, suppose that:

1. A school is reported as having 25, 62, and 27 teachers in three successive years;
2. Adding the number of teachers in all schools across the district produces a total that exceeds the total in the district record by 36; and
3. Reducing the middle year total for the school in question by 36 would produce an FTE teacher series of 25, 26, 27.

In this case we assumed that "62" was a transposition in data entry for "26". For situations where a data point appears, we will impute a replacement number.

We discussed marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts.

Impute teacher anomalies due to data entry error, reporting errors, missing/implausible data
We imputed a replacement number using PROC IMPUTE. Our regression equation used the adjacent years of FTE data (year prior and year after target year), three years of enrollment data (prior, target, and

subsequent), school locale, and school size as the predictor variables for the FTE value to be imputed.

Find and review grade total anomalies
We are now in the process of reviewing individual grade enrollments. We will perform a series of tests on the data. After each test a flag will be assigned as to whether the school (or grade) passed that test. After all tests have been performed we will look at the distribution of these flags and from that distribution decide which combination of passes and fails we will chose to impute.

The first test is the "District Imputation Test." This test determines if a school is contained within a district where the district enrollment was imputed. This test is helpful as it indicates a data quality problem at the district level, which may mean something is peculiar at the school, and possibly grade level.

The second test is the "District Sum Test." This test determines if the sum of the grade enrollments for all schools in a district equal the reported enrollment (unimputed) on the district file. If this test passes (sum of school data equals district data), then it is a pretty strong indicator that the data at the school level are okay. Cases which pass the District Sum Test will be set aside; no further testing will be done on grades within these schools, and they will not be considered for imputation.

The next two tests are related. They are the "Cohort Test" and the "Grade Test." In the cohort test, we compare the enrollment in each numbered grade within each school (excluding UG, PK, and KG) with the enrollment of the next lower grade in the preceding year and the next higher grade in the following year. If the grade is the highest in the school, we compare it to the next lower grade in the preceding year and the next lower grade than that in the year before that. If the grade is the lowest in the school, we compare it to the next higher grade in the following year and the next higher grade than that in the year after that. Examples for a school with a grade span of grade 1 to grade 6:

Target		Comparison 1		Comparison 2	
Grade	Year	Grade	Year	Grade	Year
4	1988	3	1987	5	1989
1	1988	2	1989	3	1990
6	1988	5	1987	4	1986

The grade fails if all or one of the following sets of conditions (for a V or a Λ) is true:

General Cohort Test for V:

(1) the previous year's enrollment in the next lower grade is at least twice the current year's enrollment in the grade under examination;
(2) the next year's enrollment in the next higher grade is at least twice the current year's enrollment;
(3) the difference between the previous year's enrollment in the next lower grade and the current year's enrollment is ≥ 15;
(4) the difference between the next year's enrollment in the next higher grade and the current year's enrollment is ≥ 15.

General Cohort Test for Λ:

(1) the previous year's enrollment in the next lower grade is no more than half the current year's enrollment in the grade under examination;
(2) the next year's enrollment in the next higher grade is no more than half the current year's enrollment;
(3) the difference between the previous year's enrollment in the next lower grade and the current year's enrollment is ≥ 15;
(4) the difference between the next year's enrollment in the next higher grade and the current year's enrollment is ≥ 15.

The ratios, 2 and .5, are greater than the 1st and 99th percentiles for within-grade cross-year ratios for 1994. Percentiles vary from grade to grade and year to year. A consistent pair of ratios is easier to understand, explain, and apply than the annual grade-specific ratios that we used in the trial run and that are comparable to the annual percentiles that we used for the earlier school-level task.

V/Λ Cohort Test for Lowest Grades. If the grade is the lowest grade in the school, it fails if it includes ≥ 15 students and the current year's enrollment is at least twice or less than half of the enrollment for both of the successively higher grades in the next two years. For example, the 7th grade enrollment in a school with a grade span of 7-12 would have to be half of the 8th grade enrollment in the next year and half of the 9th grade enrollment in the year after that or half or half of the 8th and 9th grade enrollments in the same years.

V/Λ Cohort Test for Highest Grades. If the grade is the highest grade in the school, it fails if it includes ≥ 15 students and the current year's enrollment is at least twice or less than half of the enrollment for both of the successively lower grades in the preceding two years. For example, the 6th grade enrollment in a school with a grade span of kindergarten-6 would have to be half of

or twice the 5th grade enrollment in the previous year and half of or twice the 4th grade enrollment in the year before that.

V/Λ Cohort Test for End Years. The Cohort Tests cannot be performed as described above for the first and last academic years covered by this project (1986–87 and 1997–98, respectively). Where the grade range is sufficient, the Cohort Test can be performed instead on the next two years for 1986 and the previous two years for 1997. In these cases, the target year enrollment should be significantly different from successively lower grades in both of the preceding two years, or significantly different from successively higher grades in both of the following two years. In these tests, the enrollment if graphed would describe not a V or a Λ but a sharp increase or decrease (doubling or halving) followed or preceded by a plateau (the asterisk indicates the target enrollment):

1986: */‾‾ *__
1997: __/* ‾‾*

The 1986 tests are not applicable to the two highest grades in a school; that is, if (HIGR – g) < 2, where HIGR is the highest grade with enrollment in the school in 1986 and g is the target grade. The 1997 tests are not applicable to the two lowest grades in a school; that is, if (g – LOGR) < 2 where LOGR is the lowest grade with enrollment in the school in 1997.

Some children do not attend public pre-kindergarten classes, some public schools do not offer kindergarten, and ungraded students by definition can not be expected to progress annually from one "grade" to another. Therefore, the V/Λ Cohort Test is not applicable to the three enrollment categories designated as "UG," "PK," and "KG." Only the V/Λ Grade Test applies to enrollments in these categories.

V/Λ Grade Test. The Grade Test compares the current year's enrollment with enrollment in the previous and subsequent years.

V/Λ Grade Test for End Years. The Grade Tests cannot be performed as described above for the first and last academic years covered by this project (1986–87 and 1997–98, respectively). The Grade Test can be performed instead on the next two years for 1986 and the previous two years for 1997. In these cases, the target year enrollment should be significantly different from enrollment in the same grade in both of the preceding or following two years. Thus, the enrollment if graphed would describe not a V or a Λ but a sharp increase or decrease (doubling or halving) followed by a plateau (the asterisk indicates the target enrollment):

1986: */‾‾ ‾‾*
1997: __/* *__

Impute remaining missing/implausible grade totals
We will review the combinations of passes and fails for the above tests, and decide to impute cases after reviewing that distribution.

We will impute a replacement number using PROC IMPUTE. Our regression equation uses enrollment in the same grade for the prior and subsequent years; enrollment in the next lower grade, if available, in the previous year; enrollment in the next higher grade, if available, in the following year; and school size as the predictor variables for the grade enrollment to be imputed. Note that we have to split the dataset into a series of grade datasets (e.g., all schools offering grade 1, all schools offering grade 2, etc.). This is necessary as a school not offering a particular grade will have a missing value for that grade. PROC IMPUTE uses missing values as the values to be imputed. Therefore a school not offering a grade, but included in that grades imputation dataset would be assigned a value for that grade when in fact it never could have had students in that grade.

Compare grade totals by school to grade ranges and identify discrepancies, review discrepancies between grade totals and grade ranges
As we edit individual grade enrollment, we will be able to edit the grade span variable. A difficulty with "grade span" is that a small school may be organized and intended to serve a span of grades but some grades might not have any students in a particular year. For each school, we will compare reported grade span to the grades for which there is positive enrollment. If this grade span differs from the reported grade span, we will retain the reported grade span if the missing grades had enrollment within the previous two years and the subsequent two years and the coefficient of variation of annual grade-specific enrollment changes within the school is close to 1.0. We will have to investigate some actual cases to determine a rule that can be applied to large numbers of schools. When a grade disappears from a school, we will also look at the next lower grade in the previous year and the next higher grade in the subsequent year to see whether the absence of the grade is explainable by an anomaly in the distribution of students in the community across grades.

If a grade at the top or bottom of the grade range disappears for only one year, we will look at the enrollments of other schools in the district to try to confirm a temporary reorganization of the district. If the "missing" grades can be accounted for by increases in enrollments in the corresponding grades in other

schools in the district, we will assume that there was a reorganization of the grade structure of the district schools that was reversed after only one year. If the grade disappears from more than one school in the district, we will assume that it was the result of a deliberate local decision about the grade structures of the various schools. If there are several schools to which the "missing" grade could have been assigned, so that the students could have been dispersed without causing anomalies in the grade structures of other schools, we will also assume that there was a deliberate decision to change the service grade span of the school in question.

We will compare the sum of the individual grade enrollments to the total enrollment for a school. This type of comparison allows for the data to be consistent across years as well as within year. We will complete our data "cleaning" process by conducting a similar comparison on a larger scale. We will sum the enrollment and number of teacher fields for all the schools in a district and compare those against the data in the time series district dataset. This will ensure that school data within a district are consistent between the two time series datasets.

Correct grade range discrepancies
Some grade-range anomalies may be obvious enough for us to decide to replace the reported data without further consideration. However, we will discuss marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts. After we have discussed these cases with him, we will decide what corrections to make and replace what we have good reason to believe are erroneous data. PROC IMPUTE does not work for grade service ranges as the grade range variable is a character variable and PROC IMPUTE is only capable of imputing numeric data.

RACE AND ETHNICITY

We will look for year-to-year anomalies as we did for students, teachers, and grade-specific enrollments, and we will compare anomalous data to data from the Office of Civil Rights. Because of possible district policies regarding the racial composition of schools, such as the institution of busing or the abandonment of busing plans effected before the coverage period, and the creation of magnet schools, we can expect to find some sudden but legitimate shifts in the number and proportion of students from a particular race in a particular school. Therefore, it will be necessary to accept abrupt changes in race data as long as the students can be accounted for elsewhere in the same school district.

Find race/ethnicity data discrepancies
We should be able to detect anomalies within each racial and ethnic group by using the same method as for total students and teachers, that is, by comparing each year's data to the preceding year (except for the first year) and the following year (except for the last year). The difference measure formula should be applied to each race within each school. It is not clear whether it should be applied to the *number* of students of each race within the school or to the *percentage* of students of each race within the school. The number could change for all races if the school is expanded or part of it is closed, changing the total enrollment dramatically. The percentage should not change dramatically unless there is a change in attendance zone boundaries or busing policy or the school becomes a magnet school. We should calculate a difference measure for both number of students and percentage of students for one race for one year and compare the results.

Review race/ethnicity data discrepancies
Discrepancies will have to be reviewed carefully to determine whether they are plausibly related to changing ethnic patterns within a district rather than to a reporting or data entry error. For example, if there is an abrupt change in a school but the same shift occurs within the district, then the shift should be accepted as plausible and not changed on the data set.

Correct race/ethnicity discrepancies
Some cases may be obvious enough for us to decide to replace the reported data without further consideration. However, we will discuss marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts. After we have discussed these cases with him, we will decide what corrections to make and replace data due to data entry and reporting errors for which we can identify an obvious correction. We will impute data for missing cases or other cases for which a single replacement number is not apparent.

Step 3. Documenting Edits and Imputations

Documentation of the editing and imputation will be important for future users to understand the strengths and limitations of the data when they are made available for public use. We will prepare documentation of all the steps we took in creating the time series datasets.

Prepare draft documentation

The documentation should include the following materials:

1. A description of the criteria used to include and exclude schools from the datasets;
2. A table showing the number of schools, and the number of districts they represent, in the database by year;
3. A description of the methodology for linking records across all years and of the policy followed in assigning and maintaining school ID numbers;
4. A codebook listing all variables names, imputation flags, and status codes, and the meaning of all codes that are not representation of numeric values of continuous variables;
5. The record layout used for all years;
6. A description of the criteria used to identify anomalous data that was replaced by imputation; and
7. A description of the imputation procedures used.