Evaluation of a New Edit Methodology for the Common Core of Data Nonfiscal Surveys

Elizabeth Goldberg, U.S. Census Bureau
Robert Stillwell, National Center for Education Statistics
Jeffrey Little, U.S. Census Bureau

Contact info: elizabeth.j.goldberg@census.gov

Presented at Joint Statistical Meetings, Montreal, Canada, August 7, 2013

Abstract:

The Common Core of Data (CCD) nonfiscal surveys consist of data submitted annually to the National Center for Education Statistics (NCES) by state education agencies (SEAs) in the 50 states, the District of Columbia, territories, and other agencies. CCD survey staff edit the data to produce a clean data file, which NCES uses to construct general-purpose publications and specialized reports. The principal users of CCD nonfiscal data are the federal government, the education research community, state and local government officials, including school boards and LEA (local education agency) administrators; and the general public. With the 2010-11 survey year, public concerns were raised about the potential for extreme erroneous results not being identified for review during the editing process. As a result, NCES adopted a new methodology of editing based on comparing multiple years of data for an edited data element. Previously, NCES only edited current year data against the prior year. This paper discusses the results of this new edit methodology on the data quality for the 2011-12 CCD survey year, as well as the potential for this method to be applied in other ways moving forward.

Keywords: data editing, ratio edits, historical edits, multivariate edits

1. Introduction

The National Center for Education Statistics (NCES) conducts several major surveys/enumerations of school systems annually. The set of surveys known as the Common Core of Data (CCD) consists of data submitted annually to NCES by state education agencies (SEAs) in the 50 states, the District of Columbia, Puerto Rico, the four U.S. Island Areas (American Samoa, Guam, the Commonwealth of the Northern Mariana Islands, and the U.S. Virgin Islands), the Department of Defense Education Agency (DoDEA) dependents schools (overseas and domestic), and the Bureau of Indian Education (BIE). These surveys include:

- Universe surveys of staff counts, student enrollment, etc. at the state, school district (LEA), and school level,
- Universe counts of dropouts and completers at the state and school district level, and
- Universe enumeration of finance data for each state and school district.

The first two sets of data are known as the nonfiscal surveys. The third data set listed is commonly known as the Annual Survey of School System Finances (F-33). In order to provide data comparable across states to the maximum extent feasible, NCES and SEAs have worked since the 1950s to develop and accept common data items and definitions. The remainder of this paper deals strictly with the CCD nonfiscal surveys.

The CCD nonfiscal surveys have been conducted since the 1950s and in their current form since 1986. SEAs report nonfiscal school, agency, and state level education data through the U.S. Department of Education's EDFacts data collection system. For many years, the data processing and editing have been conducted by the U.S. Census Bureau on behalf of NCES. Many of the CCD nonfiscal users are outside NCES. In addition, the Department of Education uses these data directly in calculating allocations for certain formula grant programs, including Title I, Part A of the Elementary and Secondary Education Act, Impact Aid, and Indian Education programs.

The CCD nonfiscal surveys provide the most current and comprehensive directories of U.S. public elementary and secondary schools and local education agencies (school districts). As the only annually updated, comprehensive listing of public school districts and schools, the CCD nonfiscal data serve as the sampling frame for many federal and non-federal programs and surveys.

Table 1 below lists the frequencies for universe counts for the 2010-11 school year.

Table 1: CCD Universe Frequencies for 2010-11 School Year

States	School Districts (LEAs)	Schools
58 states and territories	18,478 LEAs	103, 813 schools

Source: National Center for Education Statistics, 2010-11 School Year Common Core of Data State, Local Education Agency, and School Surveys, http://nces.ed.gov/ccd/ccddata.asp

The main types of information contained in the CCD nonfiscal data are State, LEA, and school characteristics, student membership, staff count data, and high school dropout and completer data.

Student membership data include enrollment counts at the state, agency and school level and, depending on the file, enrollment counts by grade, by gender, by race/ethnicity, and also various combinations of race/grade/gender. In addition, information about the number of special education students, the number of students eligible for free and reduced lunch, the number of students identified as English Language Learners, etc. is included as part of the CCD nonfiscal dataset. This type of information is typically of great interest to many researchers.

Staff count data include counts of classroom teachers, instructional aides, administration, and support staff such as guidance counselors and library specialists. Dropout data include the number of students that drop out of grades 7-12 by race/ethnicity and gender. Completer data include the number of students that successfully complete high school (by race/ethnicity and gender).

2. Editing Process and Methodology

2.1 Historical Process

In a typical survey year cycle, state data coordinators submit CCD nonfiscal data to NCES using the EDFacts data collection system. The data collection cycle for a school year opens in January of that school year. States can submit data to EDFacts at any time after the collection cycle begins. In any given survey year, some states submit a complete set of data all at once, while other states submit data over several months. States can also update data at any time until the data collection officially closes. Currently, a data collection cycle is open for three years. Census Bureau staff edit the data on a flow basis as it is submitted by states.

The CCD contains a large number of numeric data items as shown in Table 2

Table 2: Number of CCD Numeric Data Items for 2010-11 School Year

States	School Districts (LEAs)	Schools
240	244	253

Source: National Center for Education Statistics, 2010-11 School Year Common Core of Data State, Local Education Agency, and School Surveys, http://nces.ed.gov/ccd/ccddata.asp

The CCD data processing system has a complex set of edits to ensure data quality and consistency. A large set of edits is in place to ensure that the universe of schools and local education agencies for any given school year is complete and accurate. Edits are also in place to ensure that counts for membership and staff are consistent and cross check at the school, agency and state level. Due to the volume of numeric data items, creating an editing process that is thorough and yet manageable to review in a timely manner presents a challenge for analysts.

Most edits for numeric data are edits comparing current year to prior year data. These edits are usually of the form:

$$\boldsymbol{Ratio} = \frac{|current\ year - prior\ year|}{prior\ year}$$

$$\boldsymbol{Difference} = |current\ year - prior\ year|$$

where the data item fails the edit when ratio and/or difference is larger than a predetermined cutoff.

Some data items have an edit limit for both ratio and difference, some are ratio only, and some are difference only. These edit limits were set empirically and updated when it was deemed appropriate. When any given data item fails, survey analysts communicate with the state data coordinator to resolve discrepancies.

As is common in many establishment surveys, responding to false positives in the editing process is a time-consuming process and is burdensome for both the survey analysts and also the state data coordinators who act as respondents for the schools and local education agencies in their state or territory.

As mentioned, CCD nonfiscal data are often used by various data analysts as input into other measurements or as a sampling frame for other statistical surveys. Unlike many surveys where analyses are typically rolled up into higher level estimation strata, the CCD nonfiscal data publish detailed information at the state, agency, and school level. These data often receive intense scrutiny at the individual data item level as data users have been known to question individual data items for an agency or school. This creates a dichotomy for survey analysts who need to balance producing timely data while realizing that individual data items may be highly scrutinized by the data user community.

2.2 Revised Process

In the 2010-11 survey year, a concern was raised that the number of false positives was increasing respondent burden without a corresponding increase in data quality, and that potentially erroneous individual data items were being missed as a result. In response, CCD staff developed a new method to identify current year results when they were strikingly different from data in the previous few years.

Historically, reported data flagged by staff for review that were not corrected or verified by the state, were ultimately published as reported. Using the new method, data items flagged as anomalous and not corrected or explained by the states are suppressed. In addition, related data items are also suppressed. For example, if total student membership was identified by an edit and not explained or corrected, then in addition to the total student membership number, all lower level membership data (totals by individual race/ethnicity, grade, gender and various subtotals) were also suppressed. NCES applied this new method on a limited basis retrospectively to 2009-10 and 2010-11. Moving forward, the Census Bureau extended the method in the 2011-12 school year survey to an expanded set of data items.

Given the large number of numeric data items, it was determined that initially the focus of the edit re-design would be on a few key items of most interest to the data user community. Another new aspect of these edits, is that similar data items are combined for comparison purpose vs. examining each data item in isolation. Each edited data item was paired with a second data item that was similar to or contained the first data item. In this way, many false positives were removed from the pool of edit failures (see Tables 5 and 6), reducing both analyst and respondent burden.

Listed below is a description of the variables for which multiyear edits were developed for the school and local education agency data. Note that six edits were developed for the school level data for 2011-12, but only three of the six variables are discussed in this paper.

Table 3: Agency Variables

Edit Type	Edited Items (Dataset Variable Names)
Agency Membership	Total student membership (MEMBER) Pupil Teacher Ratio = MEMBER / TOTTCH
Total Teachers	Total classroom teachers (TOTTCH) Pupil Teacher Ratio = MEMBER / TOTTCH

Table 4: School Variables

Edit Type	Edited Items Description (Dataset Variable Names)
School Membership	Total student membership (MEMBER) Pupil Teacher Ratio = MEMBER / FTE
Total Teachers	Total classroom teachers (FTE) Pupil Teacher Ratio = MEMBER / FTE
Free/Reduced Lunch Membership	Total free and reduced lunch membership (TOTFRL) TOTFRL as % of membership =TOTFRL/MEMBER*100

2.3 New Edit Methodology

Below is a description of the edits applied to the 2011-12 CCD nonfiscal school and agency data files. These edits were based on methodology and SAS code created by NCES for the 2009-10 and 2010-11 CCD nonfiscal data and modified and extended by the Census Bureau for the 2011-12 survey year school and agency data. These limits use four prior years of data to compare to the current year of data. These new edits are collectively known and referred to as the "multiyear edits".

For each variable in Tables 3 and 4, the following criteria are calculated:

$$\textbf{VARIABLE Y1} = \text{mean}\left(all\ |VARIABLE\ Y_i - VARIABLE\ Y_j|\right)\ for\ all\ years\ i\ and\ j \neq to\ current\ year$$

Note: When using 4 prior years of data, Variable Y1 is the mean of 6 differences.

$$\textbf{VARIABLE Y2} = \text{mean}\left(|VARIABLE\ Y_{year} - VARIABLE\ Y_j\right)\ for\ all\ years\ j \neq to\ current\ year$$

Note: YEAR = current year. When comparing current year to 4 other years, Variable Y2 is the mean of 4 differences.

A school or agency is flagged as an edit failure when both variables in each edit "pair" (as defined in Tables 3 and 4 above) fail both 1 and 2 below:

1. **VARIABLE Y2** $\geq$ MINDIF

2. **VARIABLE Y2** $\geq$ **VARIABLE Y1** $*$ LIMIT

Where MINDIF and LIMIT are determined and set by Census Bureau staff for each variable of interest.

Variable Y2 is a measure of how different the current year's data value is compared to the prior four years for that value. Variable Y1 measures the variability in the same measurement in the prior four years and evaluates if the current year's set of absolute differences is unusually large compared to the set of differences observed in the prior four years.

3. Results

Implementing the new multiyear edit methodology produced vastly different rates of school and agency failures when compared to the former edit criteria as displayed in Tables 5 and 6 below. All data below are calculated from the 2011-12 School Year Common Core of Data Local Education Agency and School Surveys. At this time, NCES has not yet published the final 2011-12 school and agency data files. When the data are published it will be available on their website: http://nces.ed.gov/ccd/ccddata.asp.

Table 5: Comparison of School Failures by Method

	Frequency of Failures			Percentage Failed	
	Old Criteria	New Criteria		Old Criteria	New Criteria
School Membership	4,585	216		4.6%	0.1%
Teachers	14,571	836		14.5%	0.2%
Free/Reduced Lunch	3,977	3,571		4.0%	3.6%

Table 6: Comparison of Agency Failures by Method

	Frequency of Failures			Percentage Failed	
	Old Criteria	New Criteria		Old Criteria	New Criteria
Agency Membership	980	75		5.4%	0.5%
Teachers	477	102		2.6%	0.6%

The new multiyear criteria greatly reduce the number of identified schools and agencies for analyst review. With the new edit criteria, each case is given more rigorous review and state data coordinators are not as overwhelmed with potential cases for review. The review is limited to those cases that are highly different from the set of values for that same data item for the prior four years.

After review by analysts, the states are asked to explain the identified edit failures or to provide updated data. CCD analysts will review new data submissions or comments provided by the SEAs and determine if the data for any given school or agency should be suppressed and set the appropriate flag. NCES will publish these flags and the corresponding suppressions with the CCD nonfiscal school and local education agency datasets beginning with the 2011-12 survey year. Listed below are the results of the suppression criteria:

Table 7: Agency Suppression Results

	Identified	Suppressed	% Suppressed
Membership	75	38	50.7%
Teachers	102	34	33.3%

Table 8: School Suppression Results

	Identified	Suppressed	% Suppressed
Membership	216	93	43.1%
Teachers	836	212	25.4%
Free/Reduced Lunch	3571	3527	98.8%

Figures 1 and 2 below show the values of the edit criteria variables for the membership edit for agency and school. The extreme values are easy to identify and represent cases where the current year differs significantly from the prior four years of data.

As a result of this new criteria, data determined to be unreliable / potentially erroneous were suppressed on the data file. Moving forward, if states are able to provide updated data or explanations, and the data collection for the survey year is still open, the data will be updated on a periodic basis until the data collection period closes.

For the processing of 2012-13 CCD nonfiscal data, the plan is to use the edit criteria developed for 2011-12 with minor limit adjustments, if needed. In addition to the above edits for CCD school and agency, multiyear edits have been developed for the not yet publicly released Local Education Agency (School District) Universe Survey Dropout and Completion Data beginning with the 2010-11 School Year. Note that the Dropout and Completion data lag a year behind the universe survey data.

Figure 1: AGENCY Membership Edit Values

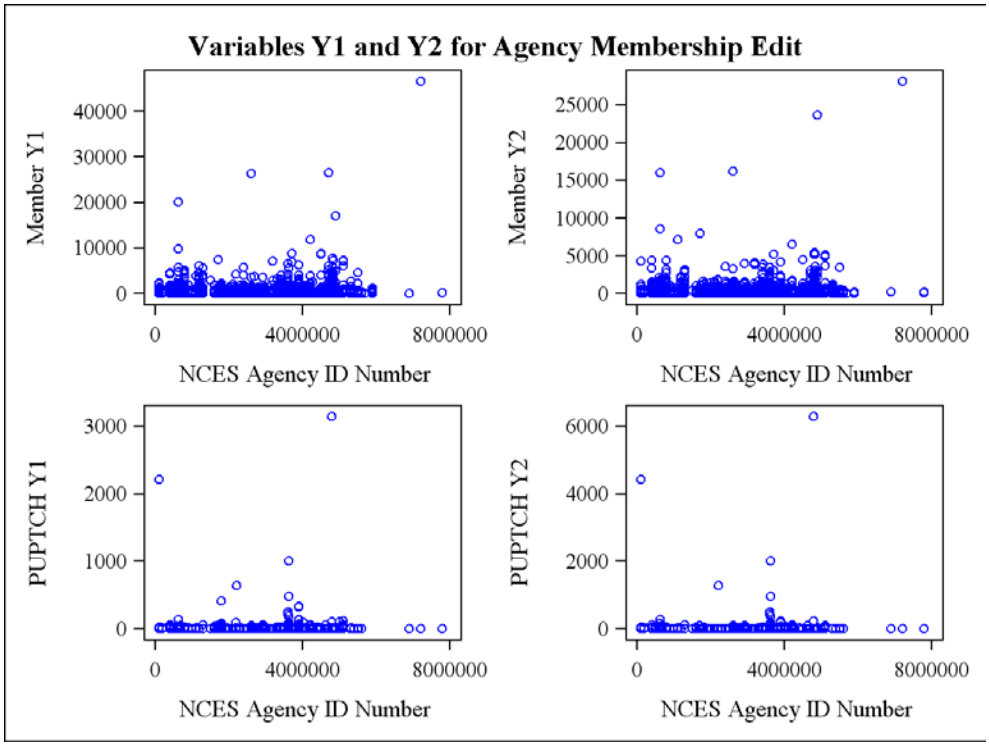
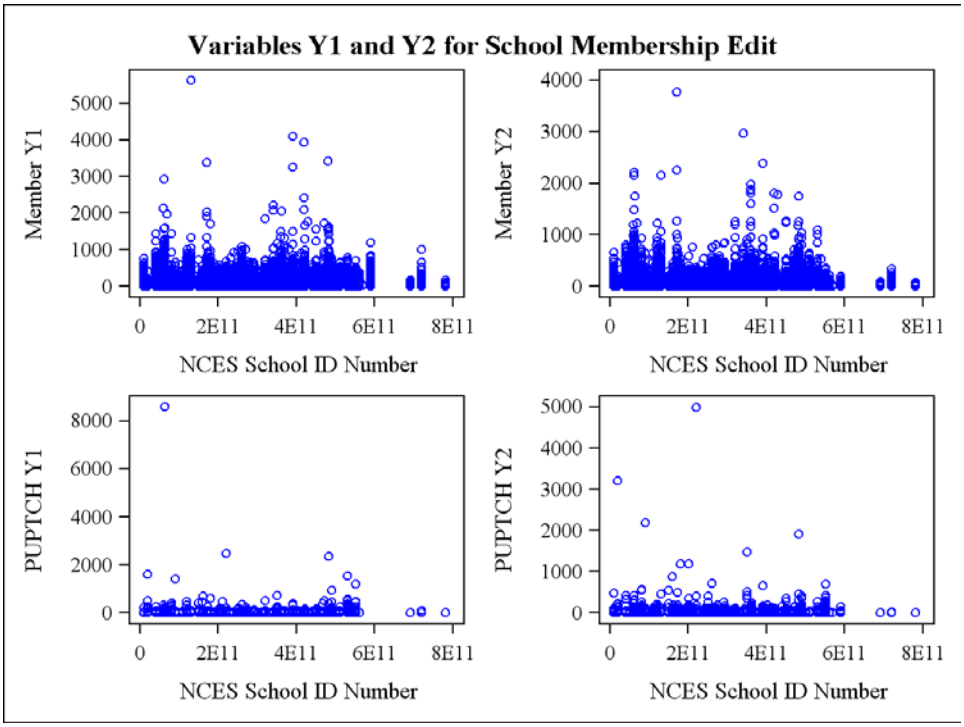


Figure 2: SCHOOL Membership Edit Values



Note: In the above figures, PUPTCH = Pupil Teacher Ratio and MEMBER = membership as defined in Tables 3 and 4 above.

4. Conclusions

The new multiyear edit methodology has been considered a success by the CCD nonfiscal analysts and respondents. It greatly reduces both analyst and respondent burden by reducing the number of data items requiring state data coordinator interaction. It streamlines the editing process by flagging only those items that are outside the typical response range for a given unit. This method takes into account historical trends for a given response school or agency and uses that information to gain insight into the current year's data.

The downside to the new method is that it is more complex to program and requires a larger historical database to be maintained in order to calculate the edit parameters. Also, in the cases where there is no or limited prior year data, the older edit methodology still needs to be used. For schools and agencies that did not have four prior years of data, these units were edited using the new method if they had at least two years of data available in the four prior years. Units without at least two prior years were edited using the old method.

Moving forward, staff are actively working to refine the new method. The use of imputation versus suppression is being considered. In addition, there is interest in expanding the new method to more data groups (administration staff counts, support staff counts, etc.) and also evaluating additional or different criteria.

5. References

Keaton, Patrick (National Center for Education Statistics), June 2012, "Documentation to the Common Core of Data State Nonfiscal Survey of Public Elementary/Secondary Education: School Year 2010–11, Final Version 1a", http://nces.ed.gov/ccd/pdf/STnonfis101agen.pdf

Keaton, Patrick (National Center for Education Statistics), September 2012, "Documentation to the NCES Common Core of Data Local Education Agency Universe Survey: School Year 2010–11, Version Provisional 2a," http://nces.ed.gov/ccd/pdf/pau102agen.pdf

Keaton, Patrick (National Center for Education Statistics), June 2012, "Documentation to the NCES Common Core of Data Public Elementary/ Secondary School Universe Survey: School Year 2010–11, Version Provisional 2a", http://nces.ed.gov/ccd/pdf/psu102agen.pdf

Methodology and Statistics Council (2012). U.S. Census Bureau Statistical Standards. Washington, DC: U.S. Census Bureau.

NCES Common Core of Data (2011) website: http://nces.ed.gov/ccd/ccddata.asp