Paper 886-2017

Data Validation Using the Basic SAS® Sort Procedure and Merge Statement

SAS Global Forum 2017

Barry Frye, Appalachia Intermediate Unit 8

ABSTRACT

Data validation plays a key role as an organization engages in a data governance initiative. Better data lead to better decisions. This applies to public schools as well as business entities. Each Local Educational Agency (LEA) in Pennsylvania reports children with disabilities to the Pennsylvania Department of Education (PDE) in compliance with IDEA (Individuals with Disabilities Education Act). PDE provides a Comparison Spreadsheet to each LEA to assist in their data validation process. This Comparison Spreadsheet provides counts of various categories for the current and previous year. LEAs use the Comparison Spreadsheet to validate data submitted to PDE. This paper discusses how the Base SAS® Sort procedure and Merge statement extract hidden information behind the counts to assist LEAs in their data validation process.

INTRODUCTION

Each year, PDE provides a Comparison Spreadsheet to every LEA within Pennsylvania. This spreadsheet contains counts for various measures for the current and previous year and is based on what an LEA has submitted to PDE. The spreadsheet contains counts for disability, educational environment, ethnicity, limited English proficiency, and gender. An LEA is required to review the counts between the two years. Further investigation would be warranted should there appear to be a larger than expected variation between the two years. A large variation may indicate a change is a reporting process or an error. Typically, though, since there are usually small normal variations from year to year, additional inquiry may not occur in these instances. This paper provides an example of how the Sort procedure and Merge statement are used to drill down behind counts to uncover supporting information that might be overlooked because it may fall into the 'small normal variation' category. Providing this additional information to an LEA will enhance their ability to perform data validation. For the sake of simplicity, this paper will deal only with disability categories. (Figure 1)

There are 29 Intermediate Units[1] (IUs) in Pennsylvania. IUs provide a broad range of educational services to their member LEAs. One service includes assisting LEAs with their PDE reporting requirements with respect to data associated with services provided for special education children. Appalachia IU8 developed the SAS program discussed in this paper that provided additional information to a sampling of its 37-member LEAs. This information was in the form of a Proposed Comparison Spreadsheet containing five additional measures or counts as well as reports containing student information behind those counts. PDE is constantly striving to assist LEAs with improving the quality and accuracy of their data used in decision-making. Based on the usefulness of these additional special education data, PDE is planning to provide this expanded information to all of its LEAs.

NOTE: **All output information in this paper is for demonstration purposes only.**

[1] http://www.dot7.state.pa.us/BPR_PDF_FILES/MAPS/Education/Statewide_IU_and_Districts_web_map.pdf

CURRENT COMPARISON SPREADSHEET

The Individuals with Disabilities Education Act (IDEA) is a law ensuring services to children with disabilities throughout the nation. IDEA governs how states and public agencies provide early intervention, special education and related services to more than 6.5 million eligible infants, toddlers, children and youth with disabilities. IDEA provides funding each year to LEAs to supplement and/or increase the level of special education and related services provided to eligible students with disabilities.

PDE provides approximately $1,100 of IDEA funding to each LEA for each student that qualifies to receive special education services. This funding is determined by a one-day count of qualified students whose biological or foster parents reside within the boundary of an LEA. This one-day snapshot count occurs on December 1 of each year.

Information on each publically educated student is submitted to the Pennsylvania Information Management System (PIMS). PIMS is a statewide longitudinal data system, which is improving data capabilities by enhancing school districts' capacities to provide robust decision support tools and to meet student-level data reporting requirements.

Each year following the December 1 snapshot collection, the Bureau of Special Education (BSE) of PDE provides Comparison Spreadsheets (Figure 2) to each IU for all LEAs within their region. Each IU then distributes a Comparison Spreadsheet to the appropriate LEA. The Comparison Spreadsheet contains counts for various categories for the current and previous year and represents what an LEA has submitted to PDE via PIMS. Included in the spreadsheet are category counts for disability, educational environment, ethnicity, limited English proficiency, and gender. An LEA uses the spreadsheet to verify the data in each category. An unusual variation may indicate a change in a reporting process or it may indicate a potential reporting error.

In addition to reporting a snapshot of special education children in their district on December 1, LEAs also report any student that exited special education. LEAs report all students that received special education services from July 1 through June 30 who exited at any time during the current school year. Some of the reasons for Exiting Special Education are: 'Returning to Regular Education', 'Graduating', or 'Moved Outside of Pennsylvania'.

As mentioned earlier, for the sake of simplicity, this paper deals only with disability types. This paper discusses how the Base SAS Sort procedure and Merge statement were able to extract 'hidden' information from what an LEA has submitted to PIMS. This information is shown later in this paper as a Proposed Comparison Spreadsheet (Figure 7). This expansion includes five additional measures to assist an LEA in data verification and accuracy improvement. Also included are program output reports (Figure 3) that include the five additional counts as well as the student information behind the counts.

Figure 1 shows the Disability Codes and Descriptions used by PDE.[2]

[2] Appendix H from http://www.education.pa.gov/Documents/Teachers-Administrators/Pennsylvania%20Information%20Management%20System%20-%20PIMS/PIMS%20Manuals/PIMS%20Man%202016-2017%20Vol%202%20v1.pdf

Disability Code	Disability Description
2121	Autistic/Autism
2122	Deaf-blindness
2123	Hearing impairment including deafness
2124	Intellectual disability (formerly Mental Retardation)
2125	Multiple disabilities
2126	Orthopedic impairment
2127	Emotional disturbance
2128	Specific learning disability
2129	Speech or language impairment
2130	Traumatic brain injury
2131	Visual impairment including blindness
2132	Other health impairment

Figure 1

Disability Codes (Challenge Types)

The Current Comparison Spreadsheet in Figure 2 shows a partial example of what an LEA might currently receive. The counts would be based on what an LEA had submitted to PIMS. The LEA would use this information to compare the December 1 snapshots of submitted data for the current and previous year.

January 6, 2016

B	C	D	E	F	G	H
Disability Code	DESCRIPTION	2015-2016	2014-2015	Difference	Percent Difference	Justification for Difference
2121	Autism	220	202	18	8.91%	
2122	Deaf-Blindness	0	0	0	0.00%	
2123	Hearing Impairment including Deafness	10	10	0	0	
2124	Intellectual Disability (MR)	175	184	-9	-4.89%	
2125	Multiple Disabilities	9	10	-1	-10.00%	
2126	Orthopedic Impairment	11	10	1	10.00%	
2127	Emotional Disturbance	215	191	24	12.57%	
2128	Specific Learning Disability	455	473	-18	-3.81%	
2129	Speech or Language Impairment	149	151	-2	-1.32%	
2130	Traumatic Brain Injury	4	3	1	33.33%	
2131	Visual Impairment including Blindness	7	7	0	0.00%	
2132	Other Health Impairment	298	282	16	5.67%	

Figure 2

Current Comparison Spreadsheet

Following the review of the information to determine correctness, an LEA is required to submit a comment in the Justification of Difference cell for any non-zero variation between the two years. Looking at the Disability Code 2123, (Hearing Impairment including Deafness), no comment would be required because there is no variation between the two years. An LEA might assume that the counts of 10 represent the same students for both years and may not investigate further. Using the Sort procedure and Merge statement, a number of situations were revealed that were not apparent by only looking at the yearly counts.

As mentioned earlier, an LEA also reports any student who received special education services from July 1 through June 30 and exited special education at any time during that current school year. For example, any student counted on December 1, 2014 and exited special education would be not appear in the December 1, 2015 count. Some of the reasons for Exiting Special Education are: ‘Returning to Regular Education’, ‘Graduating’, or ‘Moved Outside of Pennsylvania’.

UNCOVERING 'HIDDEN' INFORMATION

Finding 'hidden' information behind the counts required looking at individual student records and fields of interest. This investigation looked at:

- Students
- Primary disabilities
- Activities that occurred between the three PIMS submissions

The logic was determined only after number of output iterations. It then became apparent that there were five different situations affecting the counts between the two December 1-snapshot collections.

PROGRAM VARIABLES AND LOGIC

The SAS program used simulated input that an LEA might submit to PIMS. The input was contained in the following three comma separated variable (csv) files:

- Special Education Submission Details – Student Template School Year 2014-2015 Snapshot Date December 1, 2014
- Special Education Submission Details – Student Template School Year 2014-2015 Special Education Exits
    - All students who exited special education July 1, 2014 to June 30, 2015
        - Keep only those who exited between December 2, 2014 and June 30, 2015
- Special Education Submission Details – Student Template School Year 2015-2016 Snapshot Date December 1, 2015

Program Variables

Main Variables:

- SY (School Year 2014_2015 and 2015_2016)
- STUDENT_ID
- PRIMARY (Disability Code)

Macro Variables:

- &P resolves to 2014_2015 (Previous School Year from variable SY)
- &C resolves to 2015_2016 (Current School Year from variable SY)

Comparison Variables:

- PRIMARY&P resolves to PRIMARY2014_2015
- PRIM_EXIT&P resolves to PRIM_EXIT2014_2015
- PRIMARY&C resolves to PRIMARY2015_2016

Program Logic

- Sort the three input files by STUDENT_ID
- Merge Previous SY and Exit file to determine:
    - Students active on 2014-12-01 and exited between 2014-12-02 and 2015-06-30
    - Students active on 2014-12-01 but do not appear as EXITED
    - Students that EXITED but do not appear as active on 2014-12-01
- Merge Previous SY and Current SY to determine:
    - Same students that appear in both years
        - Students that have the same disability between the two years
        - Students that changed to a disability between the two years
            - Comparing PRIMARY&P and PRIMARY&C
    - Students that appear only in Current SY
    - Students that appear only in Previous SY.

.

PROGRAM OUTPUT

The Proposed Report, Example Area SD Reports, and the Proposed Comparison Spreadsheet were created from a simulation of data that an LEA might submit to PIMS.

Proposed Report

Figure 3 shows the additional counts with title descriptions

Example Area SD ADDITIONAL COUNTS FOR DISABILITY COMPARISON
SAME STUDENTS=Same students with same disability between the current and previous years

EXITED=Students reported December 1 2014_2015 and exited before current year December 1

CHANGED_TO_THIS=Students that changed to this disability in current year Decenber 1

NEW STUDENTS=New students for this disability reported in current year December 1

MOVED_EXITED=Verify if these students moved to another LEA or should appear on current year exit submission

PRIMARY	SY_2016_2017	SY_2015_2016	SAME_STUDENTS	EXITED	CHANGED_TO_THIS	NEW_STUDENTS	MOVED_OR_EXITED
2123	10	10	7	1	.	3	2
TOTALS	10	10	7	1	0	3	2

Figure 3

Proposed Report

Example Area SD Reports

Figure 4 shows the names and student identification numbers of the 10 students for both years. Looking at just the counts of 10 previously shown in Figure 2, one could make an incorrect assumption that these are the same students.

There are seven of the same students appearing in both years. In addition, note there are some students who appear in one year but do not appear in the other year.

January 6, 2016

Example Area SD STUDENTS ACTIVE ON DEC 1 SY 2014_2015

PRIMARY=2123

SY	STUDENT_ID	LAST_NAME	FIRST_NAME	SECONDARY
2014_2015	!$&(##*<	CASSIDY	BUTCH	2129
2014_2015	!&(!&&%%!&	GRAZIANO	ROCKY	2129
2014_2015	!*!%%#&#?%	BOBBSEY	FREDDIE	
2014_2015	#%(!&?(#$$	BOBBSEY	FLOSSIE	
2014_2015	%($(<&(#	BONNEY	WILLIAM	2128
2014_2015	%(%<?<!#%<	CODY	WILLIAM	
2014_2015	(#?$(&%%$(	JACKSON	COOL HAND LUKE	2128
2014_2015	(#(<!<(*	DUNLOP	REGGIE	2129
2014_2015	?$<<?#<?*	FELSON	FAST EDDIE	2129
2014_2015	?$*!(%&?&	MADDUX	LARRY	2121

N = 10

Example Area SD ALL STUDENTS WITH THIS DISABILITY FOR 2015_2016

PRIMARY=2123

SY	STUDENT_ID	LAST_NAME	FIRST_NAME	SECONDARY
2015_2016	!$&(##*<	CASSIDY	BUTCH	2129
2015_2016	!&(!&&%%!&	GRAZIANO	ROCKY	2129
2015_2016	!((#&<!%<	POLLITT	BRICK	
2015_2016	%#*$#&%#	BANNERMAN	HARRY	
2015_2016	%($(<&(#	BONNEY	WILLIAM	2128
2015_2016	%%$(<(?<	BEAN	JUDGE ROY	2129
2015_2016	(#?$(&%%$(	JACKSON	COOL HAND LUKE	2128
2015_2016	(#(<!<(*	DUNLOP	REGGIE	2129
2015_2016	?$<<?#<?*	FELSON	FAST EDDIE	2129
2015_2016	?$*!(%&?&	MADDUX	LARRY	2121

N = 10

Figure 4

Ten Students from Each Year

Example Area SD Reports (continued)

Additional information is shown in Figure 5 when expanding the counts from the EXITED and NEW_STUDENTS colums from Figure 3.

In Figure 5, there are:

- 3 new students in SY 2015_2016
- 1 student that exited SY 2014_2015 between December 1, 2014 and June 30, 2015
- 2 students indicated by '**?**' that appear in SY 2014_2015 and:
    - Do not appear in SY 2015_2016 and also do not appear as "EXITED FOR SY 2014_2015'
        - These students are worthy of further investigation.

January 6, 2016

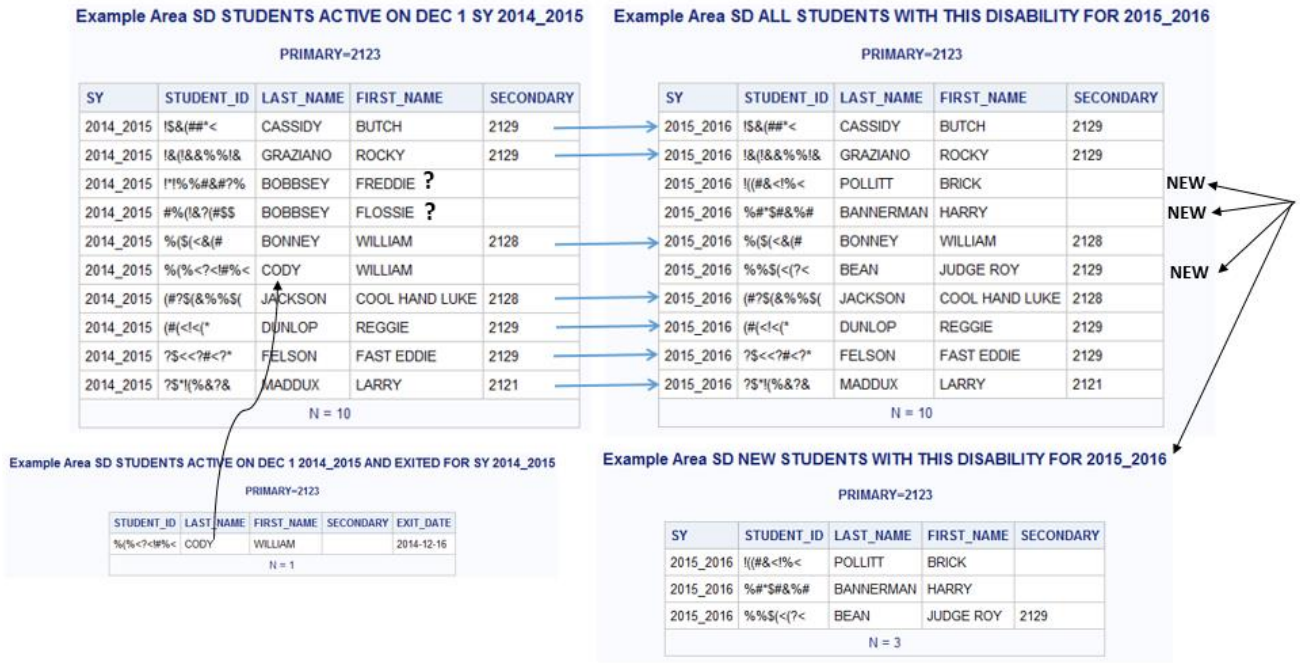

Figure 5

Same, New, Exited, and '?' Students

Example Area SD Reports (continued)

'?' Students appearing in Figure 6 could fall under one of the following scenarios:

- Moved to another LEA within Pennsylvania
- Exited Special Education the following SY between July 1, 2015 and November 30, 2015
- Were incorrectly coded.

The LEA should determine if '?' students have been reported correctly.

January 6, 2016

Example Area SD
VERIFY IF MOVED OR EXITED AFTER JUNE 30 2014_2015
IF 'MOVED TO ANOTHER LEA/SCHOOL DISTRICT' NO ACTION REQ'D
IF 'EXITED SPECIAL EDUCATION', STUDENTS SHOULD APPEAR ON 2015_2016 EXITING SUBMISSION

PRIMARY=2123

SY	STUDENT_ID	LAST_NAME	FIRST_NAME	SECONDARY	
2014_2015	#%(!&?(#$$	BOBBSEY	FLOSSIE		?
2014_2015	!*!%%#&#?%	BOBBSEY	FREDDIE		?
N = 2					

Figure 6

'?' Students

Proposed Comparison Spreadsheet

Expanding the current spreadsheet to include additional counts along with access to detailed student data reports will provide LEAs with supplementary information when performing their data validation process. Below are details on the Proposed Comparison Spreadsheet similar to the Proposed Report shown in Figure 3.

The Proposed Comparison Spreadsheet in Figure 7 shows those 'hidden' situations that were revealed by looking more closely at the data using the Sort procedure and Merge statement.

Additional Columns

1. How many students were reported in both years, same students with the same disability? (**Column H**)
2. How many students who were active on December 1 of the previous year but were reported as "Exiting Special Education between December 2, 2014 and June 30, 2015 of the previous school year? (**Column I**)
    a. These students would not appear in the current December 1 count.
3. How many students from the previous year appear in the current year but changed to this disability? (**Column J**)
    a. This most likely would occur following a re-evaluation of a student.
4. How many students were 'new' in the current year? (**Column K**)
    a. This would occur when:
        i. A new student enters an LEA.
        ii. A current LEA student was identified in the current year as qualifying for special education services.
5. How many students appeared in the previous year but did not appear in the current year and also were not reported as 'Exiting Special Education'? (**Column L**)
    a. This could occur when:
        i. A student transferred to another LEA within Pennsylvania
        ii. A student exited special education between July 1, 2016 and December 1, 2016.
        iii. A student was incorrectly coded.

Figure 7 shows a spreadsheet example of a Proposed Spreadsheet with the five additional columns.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	IU	Disability Code	DESCRIPTION	2015-2016	2014-2015	Difference	Percent Difference	2015-2016 (Same Students With Same Disability From 2014_2015)	Students Active Dec 1 2014_2015 And Exited	Students From 2014_2015 Who Changed To This Disability	New Students In 2015_2016 With This Disability	Verify If Students Are 'Moved-Known To Be Continuing' or 'Exited'	Justification for Difference
2													
3													
4	8	2123	Hearing Impairment including Deafness	10	10	0	0.00%	7	1	0	3	2	
5													
6													
7													
8													
9													
10													
11													
12													
13													

Figure 7

Proposed Comparison Spreadsheet

CONCLUSION

This paper is an example of how 'hidden' information was uncovered by using the Base SAS Sort procedure and Merge statement. The Proposed Report as shown in Figure 3 provides additional counts to an LEA when starting the data verification process. More importantly though is providing LEAs with the ability to run reports that provide student-related information. Counts for the additional categories are helpful, but names behind the counts are most helpful. LEAs know their students and related information. Having additional counts and access to reports behind counts enable LEAs to reduce time for the data verification process. It also improves the accuracy of information used for decision-making as well and information reported to PDE. Better data lead to better decisions. As noted earlier, data validation plays a key role as an organization engages is a data governance initiative.

Appalachia IU8 developed the SAS program discussed in this paper that provided additional information to a sampling of its 37-member LEAs. This information was in the form of a Proposed Comparison Report containing five additional measures or counts as well as reports containing student information behind those counts. PDE is constantly striving to assist LEAs with improving the quality and accuracy of their data used in decision-making. Based on the response from a sampling of IU8 LEAs on the usefulness of these additional special education data, PDE has begun the process of providing this expanded information to all its LEAs.

REFERENCES

Pennsylvania Association of Intermediate Units https://www.paiu.org

Commonwealth of Pennsylvania, Harrisburg, PA 17110 http://www.state.pa.us

Pennsylvania Department of Education, Harrisburg, PA 17126 http://www.education.pa.gov

Base SAS® Procedures Guide

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

<Barry Frye>
<Appalachia Intermediate Unit 8>
<814-940-0223 (ext. 1345)>
<bfrye@iu08.org>
<http://www.iu08.org/>