

Taking Your Data to the Laundry



How to improve data accuracy

December 2008 (revised)



Data
Accountability
Center

Introduction

Data are vital to decisionmakers in the fields of education, early intervention, health, public welfare, and other social services. Accurate data are critical for *IDEA* Part B and Part C for demonstrating accountability for public funding and for understanding educational results for children with disabilities. Both Part B and Part C require valid and reliable data to successfully demonstrate program effectiveness to state legislatures and the public.

Some state-reported data are used by the Office of Special Education Programs (OSEP) to meet its requirements under the *Government Performance and Results Act (GPRA)*, respond to the Performance Assessment Rating Tool (PART) administered by the Office of Management and Budget (OMB), and to address ad hoc requests from members of the Executive and Legislative Branches of the Federal government. *GPRA* requires that Federal programs report annually on the agencies' progress in meeting their goals, objectives, and performance indicators. OSEP has established goals for Parts B, C, and D of *IDEA*. Performance on these indicators is used by Congress in determining budget allocations for these programs and thus the amount of money to be allocated to states to serve infants, toddlers, children, and youth with disabilities. PART is used by OMB to evaluate Federal programs during budget formulation. It consists of four assessment areas—purpose and design, strategic planning, management, and results and accountability. PART also requires the Department of Education to establish long-term child outcome goals and data as well as develop strategies to collect performance data.

At the Federal, state, and local levels, evidence-based decisionmaking is now a key element in guiding systemic improvement efforts. Simply put, evidence-based decisionmaking is the use of information to guide decisions. Data used to guide decisions must be accurate, systematically collected and analyzed, and consistent with the experience of individuals familiar with the system.

This document is organized in four parts. First, we describe the purpose of this document. Second, we briefly discuss our approach to the terminology of editing. Third,

we provide Part B and Part C examples of data definition edits, cross-field or relationship edits, and historical edits. Finally, we draw some conclusions and describe our next steps in providing technical assistance on various data-related topics.

Purpose of this Document

This document is intended to help those who collect *IDEA* data to improve data accuracy. One key element in data accuracy is collecting clean data.¹ Cleaning data is more than applying clever edits, comprehensive error checking, or one-time verification routines. It is the consistent application of procedures that encourage policymakers to trust the data collected and use the data to guide and focus decisionmaking for improvement. The editing techniques proposed below will improve the quality of special education and early intervention data, but they are not all inclusive. Over time, state data managers will define their own unique editing techniques that are specific to their unique data, laws, regulations, and policies.

Beyond a few general recommendations, this document does not provide advice on when you should apply an edit or what to do when an edit fails. Those topics are too system specific to cover here. It is our belief that, to the extent practical, you should apply edits as close as possible to the data source. Editing at the data source makes it easier to correct errors as they arise. We also recommend that at a minimum you apply your data definition edits whenever the data are moved from one database into another. For example, if you create a data extract to move information from the local database into the state database, the data should be edited before they enter the state database.

Several Part B and Part C data managers contributed the edits listed in this document (see page 15). The list of edits presented below is certainly not exhaustive, and some edits will NOT be appropriate for your data. Some edits presented below are state specific, for example, the number of days in the school year or the maximum age for Part

¹ This document will not deal with other elements of data accuracy such as data validity. Readers may want to consult the U.S. Department of Education's Data Quality Standards at <http://www.ed.gov/about/reports/annual/2003plan/annualplan2003.pdf>. Technical assistance documents covering the requirements of the data reported under *IDEA* are available at <http://www.IDEAdata.org/TAMaterial.asp>.

B services. However, we hope the suggested edits will help you think about new ways to identify and prevent errors in your own data. The edits are organized according to what they do (e.g., define what data values are acceptable, what relationships are expected, what trends are projected) as well as the level of data they apply to (e.g., individual child/student records, aggregated counts).

Terminology

Terminology related to data editing is used in many different ways by data collectors and managers. Terms commonly used by one group may mean something entirely different to another. For example, validation edits can mean edits on a single record, edits across records, or edits against an alternate data source. To avoid confusion in this document, we avoid technical jargon and define those terms we use. That way, even if you disagree with our choice of language, you will be able to understand our meaning. We welcome your feedback where additional clarification is needed.

What is an edit?

In this document, data edits include any activity that detects or prevents errors. These activities often take the form of rules that define what values or relationships between fields are acceptable.² How and when those rules are implemented varies according to how the data are collected, the technology in use, and other factors such as available programming resources and the ability to correct errors.

Data Definition Edits

Data definition edits are the most basic. These are the rules that delineate what values are acceptable for each field. At their most general level, they involve defining the data type (e.g., text, integer, date) and the range of acceptable values. No information about other data fields is required for these edits. A few examples of data definition edits follow.

² By field, we mean a specific unit of information (e.g., the child's name, the child's address, the child's disability).

Edits on all types of fields

- The field must hold one of a specified set of values or must be blank. No other values are accepted.
 - Children’s race/ethnicity must be one of a predefined set of race/ethnicity codes.
 - Each LEA, school, and Part C locality must be identified by one of the codes on the master list of districts/schools/localities.
 - Each service provided must be identified by a valid service code.
 - Each service provider must be identified by one of the valid service provider codes.
 - Disability or eligibility codes must be one of a predefined list of codes.
- The following fields can never be blank.
 - Child name;
 - Race/ethnicity;
 - Grade-level;
 - Student/child ID; and
 - School/locality ID.
- Some fields must be blank.
 - New fields must be blank for records existing prior to the creation of the new field.
 - The Part B disability code for a student not in special education must be blank.
- Some values are valid only for records created before a specified date (the value was retired). Other values are only valid for records created after a specified date (newly added values).
- Some values must contain specific information.
 - Social Security numbers must contain no fewer than nine characters (excluding hyphens).

- Telephone numbers must contain no fewer than 10 numbers.

Edits on dates

- Dates must be valid or blank. A valid date is a date that exists in the past, present, or future.
 - All months must be between 01 and 12.
 - All days must be between 01 and 31 and must appear in the designated month (February 29 only permitted for leap years, September 31 is never permitted, etc).
 - All years must be within an expected range of values.
- Date of entry into special education/early intervention cannot be blank if the student/child is participating in a special education/early intervention program.
- Dates must be entered in the correct format as read by the data system. For example, in some states, December 6, 2008, can only be entered as 12062008 (month day year). In other states, December 6, 2008, must be entered as 06122008 (day month year).
- Dates must fall within a specified range. For example:
 - For Part B, assuming that all students in special education are between the ages of 3 and 21, all student birth dates must fall between the range today minus 3 years and today minus 22 years. On July 22, 2008, that means that the birth dates must be on or before July 22, 2005, and must be after July 22, 1986.
 - For Part C, assuming that all infants and toddlers in early intervention are between birth and age 2, all birth dates must fall between the range today and today minus 3 years. On July 22, 2008, that means that the birth dates must be on or before July 22, 2008, and must be after July 22, 2005.
 - Data collection period date must be valid for current data collection. For example, for the 2007-08 school year, only dates between September 1, 2007, and June 30, 2008, might be considered valid.

- All exit dates must be after the date the database was established.
- Most dates should not be future dates. However, there may be exceptions for future dates within a specified window. For example, in a special education database, on July 22, 2008, a date such as 03/03/2088 would most likely be unacceptable. However, a date such as 06/19/2018 might be OK (e.g., expected graduation date).
- The date of the most recent IEP or IFSP meeting cannot be in the future, but it should also not be more than 1 year ago. On July 22, 2008, that means that the date of the most recent IFSP meeting must be on or before July 22, 2008, and should not be before July 22, 2007.
- Mother's date of birth cannot be more recent than today minus 12 years. For example, on July 22, 2008, mother's date of birth must be on or before July 22, 1996.

Edits on numeric fields

By definition, numeric fields contain only numbers. No letters or special characters are permitted.

- Part B examples of numeric range edits
 - Students' FTE in special education must be between 0.00 and 1.00.
 - Number of hours with peers without disabilities cannot exceed the total number of hours in the school day (e.g., no more than 6 hours).
 - Sum of the number of hours with peers without disabilities and the number of hours outside regular class cannot exceed the total number of hours in the school day.
 - Number of hours of services cannot exceed the number of hours in the school day.
 - Number of days present for school during the school year cannot exceed the number of days in the school year (e.g., no more than 300 days).
 - The number of days absent in a school year may not exceed 180.
- Part C examples of numeric range edits
 - Birth weight should be between 750g and 6000g.
 - Gestation period should be between 25 weeks and 42 weeks.

- The maximum amount of time for services delivered should not exceed 40 hours per week.

Edits on text fields

Text fields can contain letters, numbers, and special characters. However, sometimes additional edits prevent the use of specific values. Some examples of edits on text fields include:

- Child’s first, middle, last name cannot contain non-alpha characters (except hyphen), for example, no commas, no parentheses.
- Child’s first and last name must contain at least two characters.
- SSN must have nine characters (excluding hyphens).
- The student/child’s ID must have numbers in the first 3 fields, a letter in the 4th and 10th fields, and numbers in fields 5 through 9.

Cross-field or Relationship Edits

Cross-field, or relationship, edits define the logical relationships between two or more fields. These logical relationships are often defined in terms of Boolean operators (and, or, not) and relative size (<, =, >) or temporal order (before, after). Some of these edits define relationships that must exist, and others define relationships that cannot occur. What follows are examples of edits for child-level data followed by a few examples of edits for aggregated (summary) data. For child-level data, the edits all assume the condition “If the child is in special education or early intervention, then the following conditions must hold.” We do not restate this condition for each edit below.

Child-level data

- Edits on dates
 - End/exit dates cannot be prior to a start/entry date.
 - Date of entry into special education/early intervention must be on or after date of birth.

- Date of birth must be on or before date of eligibility determination.
- Current IEP/IFSP date must be before date of next IEP/IFSP review.
- Services initiation date must be on or after eligibility determination date.
- Edits between child characteristics
 - If special education/early intervention eligibility status is eligible for services, then there must be an eligibility determination date.
 - If there is an IEP/IFSP date, then eligibility status must be eligible.
 - If current status is inactive, then there must be an exit reason.
- Part B examples of edits on children’s characteristics
 - If current status is active, then there must be a valid disability code.
 - If student’s age is 9 or older as of December 1 of the current school year, then student disability cannot be developmental delay.
 - If student’s age is 3 (w/ birthday after July 1 of current year), then status must be new referral.
 - If student is flagged as a transfer student, he/she must have an original district ID.
 - If student is flagged as a transfer student, he/she must have a transfer date.
 - If student grade is kindergarten, then expected student age is between 5 and 7.
 - If student age is between 6 and 9, then expected grade is 2 or 3.
- Part C examples of edits on children’s characteristics
 - If current status is active, then there must be an eligibility reason.
 - If eligibility status is ineligible, then there must be an evaluation date.
 - If eligibility reason is a diagnosed condition, then the diagnosis must be one of the physical or mental conditions that, in your state, result in Part C eligibility.
 - If eligibility reason is developmental delay or at-risk, then there must be an evaluation date.
 - If Medicare paid for the service, then the child must have a valid Medicare number.

- If the child has a Medicaid identification number, then the family must meet the eligibility requirements for Medicaid (i.e., family size, annual income).
- If within 90 days of child’s first birthday, then current status should be transition planning.
- Examples of edits on Part B educational environments
 - If age on the state-designated child count date (between October 1 and December 1) was 5 years or less, then educational environment must be one of the following: regular early childhood program, home, residential facility, separate class, separate school, or service provider location.
 - If age on the state-designated child count date (between October 1 and December 1) was 6 years or more, then the educational environment must be one of the following: special education inside regular class less than 40 percent of day, special education inside regular class at least 40 percent of day and no more than 79 percent of day, special education inside regular class at least 80 percent of day, separate school, residential facility, homebound/hospital, correctional facility, or parentally placed in private school.
 - If the student’s location code indicates that he/she is receiving his/her education outside the home district, then the student’s school code must be “89” (student does not attend school in district).
 - If student is age 6 or older and the student’s school ID is for a regular public school, then student’s environment type must be either special education inside regular class less than 40 percent of day, special education inside regular class at least 40 percent of day and no more than 79 percent of day, or special education inside regular class at least 80 percent of day.
- Examples of edits on Part C settings
 - If there is an IFSP date, then there must be a valid primary setting code.
 - If no services are provided in the home, then the primary setting cannot be home.
 - If the service type is respite care or transportation services, then service location cannot be service provider location.
- Examples of general edits on exit codes
 - If there is an exit reason, then there must be a valid exit date.
 - If there is an exit date, then there must be a valid exit reason.

- If there is an exit code, then current status should be inactive.
- Examples of Part B edits on exit codes
 - If reason for exit is reached maximum age for services, then age must be greater than state maximum age for services (e.g., 21).
 - If reason for exit is received regular high school diploma or received a certificate, then age should be greater than 16.
 - If current status is graduate, then graduation date must be less than or equal to today (and not blank).
 - If graduation date is not blank, then student status must equal graduate.
- Examples of Part C edits on exit codes
 - If reason for exit is Part B eligible; not eligible for Part B, exit to other program; or not eligible for Part B, exit with no referral, then the Part B eligibility evaluation date must be less than or equal to the date of exit (and not blank).
 - If reason for exit is completion of IFSP prior to reaching maximum age, moved out of state, withdrawn by a parent, or attempts to contact unsuccessful, then date of exit minus date of birth must be less than 3 years.
 - If reason for exit is Part B eligibility not determined, then age must be at least 3 years.
- Other Part B edits
 - If related services field 1 is blank, then related services field 2 must also be blank.
 - Related services field 1 cannot have the same value as related services field 2.
 - If disability is speech/language impairment, then related services cannot include speech.
 - If disability is a specific learning disability, then alternate assessment flag should not equal yes.
 - School and district IDs must be compatible. That is, the school code must be one of the legitimate values for the district, as indicated by the district code.
- Other Part C edits

- If the service requires medical equipment only available in a physician’s office or hospital, then service setting cannot be the home.
- If service type is not covered by Blue Cross, then Blue Cross cannot be the source of payment for that service.

Aggregate data

Cross-field edits should also be applied to aggregated (summary) data. The data tables states submit to OSEP are an example of aggregate data that require this type of editing. DAC applies many logical edits to these data. Complete lists of the Part B and Part C edits are available at <https://www.ideadata.org/618DataCollection.asp>. Some examples of edits on aggregated state data include the following:

- General examples
 - The sum of the number reported by race/ethnicity must equal the total number reported.
 - The sum of the number reported by educational environment/setting must equal the total number reported.
 - The number reported by age must equal the number reported by race/ethnicity.
 - The number reported for the child count must equal the number reported by educational environment/setting for the same year.
- Part B examples
 - The sum of the number reported by disability must equal the total number reported.
 - The number reported by disability must equal the number reported by race/ethnicity.
 - The number reported by gender must equal the number reported by disability.
 - The number reported by LEP must equal the number reported by disability.
 - The number of students with specific learning disabilities reported on the child count must equal the number of students with specific learning disabilities reported on the educational environments table.

- Part C examples
 - The total number of Part C infants and toddlers reported on the child count must be greater than the number of at-risk infants and toddlers reported on that form.
 - For the settings data collection, the number of children reported for each setting category cannot be greater than the number of children reported on the child count for the same year.
 - The total number reported by age group must equal the number reported by gender
 - The total number reported by settings must equal the number reported by race/ethnicity.
 - The total number reported by settings must equal the number reported by age group.

It also makes sense to apply aggregate data edits to district/locality and school/program data. This will help identify where data problems are introduced. In addition to the edits DAC applies to the data, aggregated data edits may reference other data sources. For example:

- The number of special education students reported must be less than the total number of students enrolled.
- The number of infants and toddlers referred for services should be greater than or equal to the number of eligible infants and toddlers.
- The number of special education students should not be more than 13 percent of the total student enrollment.
- The number of infants and toddlers in Part C should not be more than 10 percent of the birth through 2 population.
- If the school ID is for a regular public school, then the number of students reported in the separate school, residential facility, and home/hospital environment categories must be zero (note this edit does not apply at the district level).

Historical Edits

Historical edits compare the current data with data for a previous reporting period. While a degree of variability is expected for some fields, other fields should remain constant. Historical edits should look at both kinds of fields. For the state-reported data, the only fields OSEP expects to remain constant are state name and the reporting periods for child count, educational environments, exiting, and Part B discipline. For all other fields, some change is expected. In fact, a lack of change is suspicious. That is, if all the data on a table remain constant from the previous year, then we assume it is a mistake.

- For fields that are expected to change, historical edits are meant to identify changes outside the parameters of normal change. OSEP developed a set of historical edits to apply to the state-reported data. These edits are based on changes observed in the data over a period of several years and include both the numeric size of a change and that change as a percentage of the previous year's value. These edits are now contained in the year-to-year change reports. They consist of flagging any change that is more than 10 cases and more than 10%.

Historical edits also make sense for data containing individual student records. Some examples provided by Part B and C data managers include:

- All children reported by the school/district/Part C locality last year (e.g., present on their data snapshot) must be accounted for (present or valid exit code).
- All students not reported by the school/district last year, but reported this year, must have a valid entry code.
- If a student's status at school A is transfer to school B, then school B must have that student identified as a new transfer.

- No infant or toddler identified as moved out of state should be active in any Part C locality within the state.

Conclusion

Cleaning data is a time-consuming and resource-intensive process. A good rule of thumb is to take as much time cleaning the data you collect as you do in analyzing your data. Clean data are needed for evidence-based decisionmaking. Decisions based on dirty data will be poor decisions. Poor decisions are not likely to lead to improved results.

As recently as 10 years ago, the technological infrastructure was not available to create the comprehensive information systems or longitudinal data systems states need to demonstrate accountability for state (and local) special education and early intervention programs. Today, data managers have the tools and technology to assist with the data cleaning process. Most localities and schools are connected to the Internet; personal computers are inexpensive and prolific; the Web is used to access and explore data; and child-level data systems are widespread, and more states are developing them. The tools now available allow us to collect cleaner data and use these data to demonstrate improved results for children with disabilities.

We hope that you found this document useful. If you have any additional edits to share, please send them to us. We will periodically revise this document to include any additional edits you share with us. DAC will be developing additional materials to support the collection of high quality data. Please send your suggestions for additional materials to IDEAData_PartB@westat.com or IDEAData_PartC@westat.com.

Acknowledgment of contributors to the original document (April 2006):

Authors:

Marsha Brauen, Westat

Roberta Garrison-Mogren, Westat

Shafali Srivastava, Westat

Lucian Parshall, Consultant

A special note of appreciation to the following colleagues for documentation and feedback:

Bruce Bull, Oregon Department of Education

Karen Denbroeder, Florida Department of Education

Wayland Gillingham, Pennsylvania Office of Mental Retardation

John Hunter, Intelligence Group

Mary Keenan, Connecticut Department of Education

Cheryl Pulley, Wyoming Department of Education

Paul Raskopf, Virginia Department of Education

Alice Ridgway, Connecticut Birth Through Three System

Peggy Staples, Arizona Department of Education

Mason Vosburgh, Kansas Department of Education

Tong Zhang, Texas Department of Education