

ADMINISTRATIVE DATA COLLECTION IN EDUCATION: WHAT WE KNOW AND HOW WE KNOW IT

Mark Glander

Richard Reeves

Robert Stillwell

National Center for Education Statistics

Presented at

Federal Committee on Statistical Methodology

Postsecondary Data Collections Using Administrative Records

What is IPEDS?

Integrated Postsecondary Education Data System

- Annual data collection conducted by the U.S. Department of Education's National Center for Education Statistics (NCES).
- Required of every college, university, and technical and vocational institution that participates in federal student financial aid (Title IV) programs.
- A system of 12 interrelated surveys that collects institution-level data three times a year.
- Data collected through IPEDS are made publicly available.

Outcome Measures (OM)

- Applicable to degree-granting institutions
- Four degree/certificate-seeking undergraduate student cohorts:
 - Full-time, first-time students
 - Part-time, first-time students
 - Full-time, non-first-time entering students
 - Part-time, non-first-time entering students

Rendition of Survey

Adjusted 2008 Cohort	Awarded by your institution	Students who did not receive an award from your institution through August 31, 2016	
		Number still enrolled at your institution	Number who subsequently enrolled at another institution

- Full-time, first-time students
- Part-time, first-time students
- Full-time, non-first-time entering students
- Part-time, non-first-time entering students

GR survey common elements

- IPEDS has a graduation rate survey that measures graduation rates for First Time Full Time Degree Seeking Students.
 - 100% of degree time
 - 150% of degree time
 - 200% of degree time

Remember OM is a survey that measures awards 8 years after entry.

Errors in submission

- Cohorts didn't match GR to OM
- Graduation Rates were higher than OM Completion Rates.
- Failure to report transfer out rates.

Risks of erroneous data

- These data will be used by the public to rank and compare institutions.
- These data may be used in the Department's Scorecard tool to compare institutions.

Cleaning up the Data

- Reaching out to institutions to correct errors in subsequent year
- More integrated data checking between GR and OM Surveys has been released

Data Quality & NCES' Common Core of Data How Usage Drives Data Quality

Federal Committee on Statistical Methodology
December 6, 2016

Data Quality & NCES' Common Core of Data

The Common Core of Data (CCD):

- An annual survey of the public elementary/secondary school system in the U.S.
- Data provided by state education agencies (SEAs)
- Three levels of data: state, district & school

Data Quality & NCES' Common Core of Data

Various sources of CCD data:

- Most data come through ED Facts, which collects data for many offices within ED (need to balance our needs with other offices)
- Mix of descriptive data (e.g., enrollment) and program data (e.g. lunch counts) (Program data primarily serves other purposes; NCES not the steward).
- Geography coding derived by another branch within NCES as a separate product. (Some quality control outside of CCD.)

Data Quality & NCES' Common Core of Data

Various uses of CCD data:

- Sampling frame with NCES (NAEP, NTPS)
- NCES publications – Digest and Condition of Education
- Program Administration (Rural Education Achievement Program)
- Best High School Rankings – U.S. News & World Report
- Validation of schools' status for grant applications

Data Quality & NCES' Common Core of Data

CCD as administrative data vs. statistical data:

- Increasing statistical accuracy makes individual records less accurate.
- Data of record for program purposes.
- Legal requirements may dictate how and when data are reported.
 - National: new requirements under ESSA
 - State: may not accept corrections to data once submitted & certified
- Outliers are real – unlike sample surveys, records represent a single entity, not a subpopulation
- Cyclical collection – continuous improvement

Data Quality & NCES' Common Core of Data

CCD Data Quality Measures:

- Prescriptive (vs. reactive) business rules
- Next-day turn-around of data quality reviews
- Minimize false positives; prioritize issues for review
- Process transparency
 - Show data providers how data were processed
 - Show data users how data were collected, derived
- Impute for missing at state level; suppress anomalous data at district & school level

EDFacts

What is *EDFacts*

EDFacts is a U. S. Department of Education initiative to put performance data at the center of policy, management and budget decisions for all K-12 educational programs. *EDFacts* centralizes performance data supplied by K-12 state education agencies (SEAs) with other data assets, such as financial grant information, within the Department to enable better analysis and use in policy development, planning and management.

Respondents

- State Education Agencies (SEAs) report to *EDFacts*
 - 50 states and DC, 2 Federal Jurisdictions, and 5 Outlying Areas
- Reporting levels
 - 58 SEAs
 - >18,000 Local Education Agencies (LEAs)
 - >100,000 schools
- Data aggregations represent over
 - 50 million students
 - 6.3 million staff

Items

- Data are grouped into data groups.

An EDFacts data group is a specific aggregation (i.e., group) of related data that are stored in EDFacts to satisfy the specific information need of one or more ED program offices. Thus, an EDFacts data group does not represent a single data entry but rather a set of related data entries. Each EDFacts data group is intended to be discrete, concise, universally understood, and non-redundant.

- Over 180 data groups are included in *EDFacts*
- Data groups can be reported at:
 - Multiple levels (SEA, LEA, School)
 - Multiple disaggregations

Data Group 052: Membership

Aggregation	Table Name	Grade Level (Membership)	Racial Ethnic	Sex (Membership)	Total Indicator	Comments
Category Set A	MEMBER	X	X	X	N	Enrollment by Grade Level (Membership), Racial Ethnic and Sex (Membership)
Subtotal 1	MEMBER	X	X		Y	Subtotal by Grade Level (Membership) and Racial Ethnic
Subtotal 2	MEMBER	X		X	Y	Subtotal by Grade Level (Membership) and Sex (Membership)
Subtotal 3	MEMBER		X	X	Y	Subtotal by Racial Ethnic and Sex (Membership)
Subtotal 4	MEMBER	X			Y	Subtotal by Grade Level (Membership)
Total of the Education Unit	MEMBER				Y	Total of the Education Unit

Primary Stakeholders

- *EDFacts* operates as a service provider.
Collection is on behalf of ED program offices
- 15 different program offices
 - Office of Elementary and Secondary Education
 - **NCES Common Core of Data (CCD)**
 - Office of Special Education Programs
 - Office of Civil Rights

NCES Common Core of Data (CCD) Non-Fiscal

- Information that describes schools, school districts, LEAs, and SEAs
 - Name, address, and phone number;
 - Demographic information about students (grade, race/ethnicity, sex)
 - Staff (17 categories)
 - Special Populations (English Language Learners, Free & Reduced Price Lunch, Special Education)
 - Graduation and Dropout data

What does this have to do with Data Quality

- SEAs have small staff and are responsible for a huge amount of data
- Distance between respondents and data
- Data systems are not integrated
 - LEAs vs. their SEAs
 - SEAs vs. other SEAs
 - SEAs vs. *EDFacts*

Improving Data Quality

- Focus on the source
 - ED is funding initiatives through grants to improve data quality at it's source (statewide longitudinal data systems)
- *EDFacts* Submission System (ESS)
 - Format and Validation Errors
 - Submission Errors
- EMAPS Metadata Collection
- CCD Data Management System (CCD-DMS)
 - Data Review Warnings and Errors

Prescriptive Editing

- CCD has been around since the 1970s
- *EDFacts* was established in 2003
- Both have a history of reactive editing
 - Often overly specific
 - Hundreds of edits employed
 - Difficult to ascertain what is being covered
 - High rate of false positives
- Moving toward predictive editing model
 - help understand what is being checked and how
 - Focus on major issues with lower false positive rates

Business Rule Types

Business rules are categorized into four business rule types. Each business rule type defines the general purpose of the rule.

- 1. Prior Year Comparisons** – Was the change from the prior year more than expected?
 - Prior year vs. current Year
 - 5 year consistency checks (CCD-DMS only)
- 2. Consistent Reporting** – Are the sums of the counts for the same population equal?
- 3. Reporting Scope**
 - Unreported – Are data that should have been reported missing?
 - Misreported – Are data reported that should not have been?
- 4. Expected relationship** – Do the counts or statuses make sense in relationship to one another?

1. Prior Year Comparisons

- Prior year vs. current year
 - Prior year comparison business rules identify unexpected count, percentage and/or status changes between the current and prior year. Possible comparisons included in this type could include:
 - Both count and percent change over threshold
 - Only percent change over threshold
- 5-year consistency checks (CCD-DMS only)
 - Review of consistency across 4 prior years checked against the current year
 - Checks counts and associated ratios (e.g., student count and pupil/teacher ratio)
 - Fewer false positives

2. Consistent Reporting

Consistent reporting business rules identify two or more aggregated category set counts / subtotals / education unit totals (EUTs) that should be equal but are not. Possible comparisons included in this type could include:

- Category set count aggregation not equal to a different category set count aggregation for same permitted value combination
- Category set count aggregation not equal to subtotal for same permitted value combination
- Category set count aggregation not equal to EUT
- Subtotal not equal to subtotal
- Subtotal not equal to EUT
- EUT not equal to EUT

These include comparisons within the same data group or between different data groups.

3. Reporting Scope

Reporting scope business rules identify unreported data (when no or zero count was reported when a count could reasonably be expected based on the values of other data submitted) or misreported data (when a count was reported when a count was not expected based on the values of other data submitted). Possible comparisons included in this type could include:

- State reported count for permitted value indicated as not collected in *EMAPS* metadata
- State did not reported count for permitted value indicated as collected in *EMAPS* metadata
- State reported membership for non-operational school
- State did not report membership for operational school
- LEA reported grades offered but total student count is zero

4. Expected Relationships

Expected relationship business rules identify counts that violate other relationships. Possible comparisons included in this type could include:

- Reported count exceeds a maximum threshold
- Counts reported at LEA level exceed count reported at state level
- For an education level, the count for a subgroup (e.g., Children with Disabilities, students eligible for free and reduced price lunch) exceeds threshold percentage of total student count or total student count

Error Resolution

- ANOMALIES ARE REAL
- States request ESS business rules turned off for a certain time period to allow for anomalies
- CCD-DMS facilitates data quality review communication
 - Errors presented in a web tool
 - Commenting function (messaging between SEA staff and NCES representatives attached to each error)
 - Error resolution codes
- Long-Term Transparency Goals
 - Publish error reports
 - Publish comments and resolution codes

How are these Data Used

- Internal:
 - Program evaluation
 - Grant monitoring
 - Compliance enforcement
 - Funding allocations
 - Survey Frame
- External
 - Researchers
 - Survey Frame
 - Parents and community members
 - School rankings

Editing

- ED*Facts* does not impute or adjust reported data
- CCD only imputes at the SEA level
 - Non-response
 - Disaggregation
 - Minimal data adjustments at any level
- Research into imputation at lower levels
 - Data adjustments and Imputation at lower levels can have real-world impacts
 - Trade-off between statistical accuracy and the accuracy of individual records (i.e., raking adjustments)

Statistical vs. Administrative Data

- Statistical accuracy can lead to micro-level inaccuracy
- Anomalies are real
 - Charter schools create churn across new and existing schools
 - Virtual schools and special education schools have very different student/teacher ratios
 - Some highly selective schools have high populations and very few dropouts
 - A 7-year old really did bring a gun to school
- Schools, LEAs, and states are individually disclosed
 - Changing data can lead to distrust as locals know their data
- Real world implications to data imputation and adjustments
 - Grant monitoring and legal compliance
 - School rankings (property values?)
 - Allocations