Data Use Issue Brief 3

Turning Administrative Data into Research-Ready Longitudinal Datasets

As SLDS Data Use Issue Brief 2, "Forming Research Partnerships with State and District Education Agencies," discusses, statewide longitudinal data systems (SLDSs) have provided state education agencies and local education agencies (SEAs and LEAs) with valuable data for education research. However, because administrative data files are not created specifically for representative longitudinal research, there may be multiple steps involved in preparing the data in order to make it "research ready." This issue brief (and the next) share important information about the SLDS data collection process that is relevant for analysis and may be of particular interest to researchers outside of the education agencies.

An SLDS consists of administrative data primarily collected for the following purposes:

- record student enrollment and assessments;
- manage teaching staff and financial resources;
- assign and schedule classes and record grades;
- exchange information with parents; and
- meet federal reporting requirements for ED*Facts*, Civil Rights Data Collection (CRDC), Common Core of Data-Fiscal (CCD-Fiscal), etc.

Because these purposes are separate from research, administrative data may look different from the type of data researchers would want. For example, some administrative data files may be longitudinal but not representative, and others may be representative but not longitudinal. Student assessment files may be both longitudinal *and* representative, but not linked to other student, teacher, course, or household files.

As states continue to build their SLDSs, more of these links between files are becoming available. However, to take advantage of the longitudinal nature of the data, research partners and in-house analysts often must first construct representative research-ready datasets. While data systems are unique to a state or district, a number of best practices have emerged. This brief discusses some key challenges and potential solutions when working with administrative data.

Important Considerations when Selecting Administrative Data Files for Longitudinal Analysis

Make sure you know why and how each record was collected.

Find ways to get as much information about the data collection process as possible without burdening SEA and LEA data personnel. Familiarize yourself with the Common Education Data Standards (CEDS) and federal reporting requirements using the National Center for Education Statistics (NCES) and other U.S. Department of Education websites. Find metadata/paradata (data about the data) on SEA and LEA websites.

Example: Suspension files may include all students—some with one or more suspensions recorded, and others who appear to have not been suspended. However, these suspensions may only be the more serious Legally Reportable Offenses, or those meeting the Alternative Learning Program reporting requirements for special education students. Similarly, it is important to confirm whether English Language Learners variables indicate that testing was warranted, services were needed, or services were provided.

July 2012

This publication is part of a series of Data Use Issue Briefs designed to share best practices on data use by states. Issue Brief 1 provides an overview of the different types of data use. Briefs 2-4 share best practices for conducting education research with longitudinal administrative data. Look for forthcoming issue briefs on instructional data use.

Data Use Issue Brief 1:
The Data Use Landscape

Data Use Issue Brief 2:
Forming Research Partnerships with State and Local Education Agencies

Data Use Issue Brief 3:
Turning Administrative Data into Research-Ready Longitudinal Datasets

Data Use Issue Brief 4:
Techniques for Analyzing Longitudinal Administrative Data

This brief is excerpted from the following working paper:

Cratty, Dorothyjean (2010). "Conducting Responsible Education Research with Longitudinal Administrative Data."

For more information on the IES SLDS Grant Program or for support with system development and use, please visit http://nces.ed.gov/programs/SLDS.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Do not assume linked subsets of data are representative: every additional file or year linked leads to additional non-random attrition, leaving a particularly stable—but non-representative—subset of students and teachers.

Example: The state enrolls 100,000 5th graders each year. An analysis sample for teacher value-added estimates requires 4th and 5th grade math and/or reading scores and teacher record linkages, which may reduce the sample to as low as 60,000 5th graders. Adding teacher certification variables requires two or more years of teacher matching, which could further reduce the sample to less than 50,000 students. The omitted students are those who move, miss a test, take an alternative test one year, and/or are in pull-out classes that are harder to match to teachers (see Issue Brief 4 for techniques regarding the use of analysis subsamples).

Do not assume that data files are the same across years. States often have to change what data they collect and/or how they collect them.

Example: Variable codings that can change from year to year include: 1) parents' education changes from four categories to six; 2) school lunch eligibility changes from yes/no to free/reduced/paid; and 3) a variable for raw test scores changes from recording only standard test scores to recording both standard and alternative test scores, which have different scales. In this last case, the files ideally include a new variable indicating which test a student takes, but that information may not be obvious or included in the documentation.

Potential Challenges to Preparing Administrative Data for Research

Longitudinal analysis requires a representative sample of unique observations with a complete panel of consistently coded variables with few missing values. School administrative records pose many challenges that must be addressed in order to meet these requirements. Ideally, solutions should be tailored to each specific research context, but if single, all-purpose decision rules are made, details of the rationale and any sensitivity tests should be well documented. The following are examples of potential problems and solutions.

Non-unique observations for individual students in a given year

Problem: Multiple, identical annual student course records for elementary grades in a course file set up to accommodate semester and block schedules of middle school and high school courses.

Solution: Confirm and delete duplicate or superfluous records (this is important to do within each individual file in each year—before merging them into a longitudinal analysis file).

Problem: Multiple assessment records for the same student in the same year, but with different test score values.

Solution: Find out why there are two scores: were two students matched to the same ID, or did one student change schools mid-year or take a make-up test? Decide how to aggregate to one score per student per year (i.e., take the mean, highest, or latest score).

Non-unique values for the same variable across student-year records

Problem: In the same year, some students have both an 8th grade math score and a 9th grade Algebra I score.

Solution: Determine whether students are 8th graders taking a nearby high school algebra class, or 9th grade algebra students with 8th grade scores carried over in the records.

Problem: In the same year, some students are recorded as being in different schools.

Solution: Determine whether two students are sharing the same ID, students moved mid-year, or students are dual-enrolled. Decide which school to assign to the student-year observation: the school with the test scores, the school with the most days attended, or other. Alternatively, create two student-year school variables: "main school" and "other."

Missing values within panel

Problem: In some years, important student values are missing. There may have been a problem with the source file in a given year, or the data may not be collected at every grade level. For example, school lunch eligibility and special education status are often more frequently recorded for early grades.

Solution: Rather than dropping all student-year records with incomplete longitudinal values from the analysis, decide on an aggregation method to complete the panel of variables. For example, replace missing student-year values for race and gender with the student's modal value. For variables that were not recorded at every grade, consider recoding designations such as "ever school lunch eligible" or "ever learning disabled," as appropriate.

Visit the SLDS website for additional resources on administrative data use (*http://nces.ed.gov/programs/slds*).