Data Use Issue Brief 4

Techniques for Analyzing Longitudinal Administrative Data

As SLDS Data Use Issue Brief 3, "Turning Administrative Data into Research-Ready Longitudinal Datasets," explains, administrative data collected for education-related purposes differ in important ways from research-ready datasets. For this reason, analysis conducted using administrative data also requires a slightly different approach. As with preparing the data, additional work may be required, but this extra work will also allow for powerful and actionable analysis. This brief offers analysts some best practices for effectively using longitudinal administrative data for education research.

Suggestions for Conducting Longitudinal Analysis with Administrative Data

Do not limit analysis to commonly used datasets. Assuming that these readily-available files have the best data can lead to underuse and/or misuse of important data for studying education.

Example: Test records often have many years of complete and consistent demographic variables for grade, school, school lunch eligibility, etc., but important student populations, like those at risk of dropping out, will not be found in classes with state testing, such as Algebra II or Physics. Additionally, datasets of state test results often do not contain important control variables or student outcomes like graduation status, suspensions, or course taking.

Information about the data is invaluable for constructing the best research datasets.

Any empirical analysis begins with cleaning the data by checking for errors or anomalies, but analysis of administrative data requires a bit more preparation. Because the data were collected for other purposes, values in an analysis file may be correct but the file may be incomplete. For instance, a variable of interest may not have been collected for every student in the sample grade or year. Alternatively, the variable may have been collected for every student but it may not be accompanied by information necessary to link to other files or years. However, it may be possible to complete the dataset prior to analysis by using reliable data from related files and years to triangulate missing or conflicting data points (for examples, see Issue Brief 3 on prepping administrative data).

If you use analysis subsamples, report on their distributions relative to the population.

If necessary, useful analysis can still be conducted with subsamples of students (or years, teachers, or schools), but it is important to first qualify the analysis sample relative to all students in the state or district. To do this, first compare the number and type of observations in analysis subsamples to total enrollment. Then compare values of key variables in analysis subsamples to the full sample. It is important to do this by comparing distributions versus simply comparing means, because the students most difficult to link across files and years may be both those with the highest and lowest test scores, etc. Finally, make sure to document these comparisons when writing up the analysis.

Example: You wish to conduct a value-added analysis for a particular teaching credential, but only 60 percent of all 5th grade students in a state can be matched to both their 4th grade scores and their teacher of record. To determine whether the subsample is representative, compare the distributions of key variables (e.g., test scores, socioeconomic status (SES) measures,

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

teacher qualifications, etc.) for the 60 percent subsample with the full student and teacher samples. Depending on the data and audience, the comparison may be done with histograms, boxplots, quintiles, etc., though the most rigorous approach would be to conduct an "attrition bias test."

Controlling for school district variation in the administrative data collection process across an SLDS may improve the reliability of estimates.

Administrative data files are compiled by SEAs, but data collection occurs at the school and district level. Data quality procedures ensure that final values are accurate and consistent for school, district, state, and federal reporting. But the raw data files reflect the variation in data collection processes across districts due to both policy and procedural differences (such as different computer systems).

When conducting SLDS research, therefore, it may be important to control for potential variation across local school systems. The methodology by which a researcher controls for the data collection process may vary by discipline. In the case of regression analysis, this could be accomplished with the inclusion of indicator variables for LEAs (referred to by some as "district fixed effects").

Examples of Techniques Used to Leverage the Power of Existing Administrative Data

Use additional, reliable information about the education context to validate data constructs.

Often, variables important for analysis will need to be constructed (e.g., teacher of record, course classifications, SES, grade retention, dropout status, etc.). With administrative data, it is important to check constructed variables against a reliable data source; this may be a data source outside of the state or local education agency data files.

Examples from a statewide, 3rd-12th grade, student-level dropout analysis (Cratty 2012):

- Compared constructed 9th grade retention rates to SEA or LEA annually-reported rates.
- Compared constructed SES indicators to Census aggregates for state by race, age, etc.
- Compared constructed Advanced Placement variables to College Board state reports.
- Compared cohort attrition with Census and National Center for Education Statistics figures for student migration out of state or into private school.

Even a few recent years of high quality, detailed data can be very useful.

As states and districts continue to improve and expand their longitudinal data systems, valuable data (such as transcripts and student-teacher links) become available, but with limited history. Synthetic cohorts[1] constructed from these files allow researchers to model longer trajectories of the new content. Though, it is important to note that these are only analogs to true cohorts. Also, new data can be used to validate outcome variables or subsamples constructed from earlier, less complete data.

Examples from a statewide, 3rd-12th grade, student-level college readiness analysis (Cratty 2012):

- Synthetic cohorts of detailed math outcomes and opportunities were created by stacking and staggering 2006–2008 course data for grades: 10th–12th, 8th–10th, and 6th–8th, etc., allowing for analysis of student-level math preparation from 3rd grade math scores to high school calculus.
- New teacher-student links available as of 2007 were used to check the accuracy of the method used to link teachers in 1998–2006 data files. The earlier method (of matching test group composition to aggregate course and personnel files) was found to be 95 percent accurate.

Aggregate analysis produces valuable results and demonstrates responsible data use.

Sometimes one important data source for the analysis is only available at a higher level of aggregation than another source. This can occur when linking across agencies with different data access policies. Merging an aggregated version of the more detailed file can still allow for valuable analysis. It also allows researchers to demonstrate to those agencies their ability to use the data to conduct relevant and responsible analysis.

Examples:

- North Carolina workforce data (layoffs) were publicly available at the county level only; Ananat, Gassman-Pines, and Gibson-Davis (2011) aggregated ten years of student test scores to the county level to analyze the effect of large plant closings on student achievement.
- Maine Occupational Safety and Health Administration (OSHA) data were available at industry level only; Cratty and Kaufman (2000) generated "industry claim rates" out of ten years of detailed workers' compensation claims to test whether a statewide reduction in benefit levels resulted in fewer claims per injuries.

Visit http://nces.ed.gov/programs/slds for more on administrative data use.

References

Ananat, E., Gassman-Pines, A. & Gibson-Davis, C. 2011. The Effects of Local Employment Losses on Children's Educational Achievement.

Cratty, D. & Kaufman, R. 2001. The Effects of Workers' Compensation Reform in Maine on Injury and Claim Behavior.

Cratty, D. 2012. Do 3rd Grade Math Scores Determine Students' Futures? A Statewide Student-Level Analysis of College Readiness and the Income-Achievement Gap.

Cratty, D. 2012. Potential for Significant Reductions in Dropout Rates: Analysis of an Entire 3rd Grade State Cohort.

[1] An artificial cohort constructed using data derived from actual cohorts present, at different ages, during a single period in time. For example, a synthetic cohort of grade 1 through grade 12 students could be constructed by assembling data from three school years by using four separate cohorts from those years: one cohort from grades 1-3, and three other cohorts from grades 4-6, 7-9, and 10-12.