SLDS Issue Brief

Data Quality: Striking a Balance

Many states with statewide longitudinal data systems (SLDSs) wrestle with the question of whether to aim for ensuring their data are accurate or more complete. However, data quality and data quantity are not mutually exclusive; completeness is one of the attributes of overall data quality. Both are important to pursue, but given finite resources and time, states often must decide which aspects of data quality to prioritize. It is not realistic to attempt to achieve perfect data quality. Each state has to assess the current quality of the data in its SLDS and weigh the benefits of using that information for certain purposes with the potential risks of not using it. Determining the appropriate level of effort focused on improving quality is essential to ensure the time and resources spent are worth the return.

Defining Data Quality

The following attributes are common to most definitions of data quality:

- *Accuracy:* Are the data correct?
- *Completeness:* Are the data all there?
- *Timeliness:* Are the data available when needed?
- *Validity:* Do the data accurately reflect or measure what they are supposed to reflect or measure?
- *Consistency:* Are the data stable over time and across sources?

Assessing Completeness

Two primary factors influence the completeness of data in an SLDS:

- *How completely is each individual element submitted to the source system?*
- *What is the match rate across data sources?* This factor is particularly relevant for P-20W SLDSs.

Because SLDSs receive data from multiple source systems, the SLDS data can only be as complete as the data contained in each source system at the time it is pulled into the SLDS. If the data elements go through additional edit checks before they are loaded into the SLDS, the SLDS data may be less complete—though more accurate—than the data in the source system. The match rate of records across data sources also influences the completeness of the dataset. The match rate, in turn, is determined by the quality of the source records and of the matching algorithm.

A Purpose-Driven Approach

Identifying the intended use(s) of the SLDS data is one of the first steps to striking the balance between data quality and completeness, and to determining the appropriate level of effort to put into improving data quality. Consider the following questions when identifying intended data use(s):

- *Who will use the data?*
- *For what decision(s) or purpose(s) will they use the data?*
- *How quickly do they need the data?*
- *How often do they need the data?*
- *What stakes are attached to the use(s) of the data?*
- *If the data are not as complete or accurate as desired, what will happen if they are used anyway?*

This product of the Institute of Education Sciences (IES) Statewide Longitudinal Data Systems (SLDS) Grant Program was derived from an SLDS workshop that took place in May 2013. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program. We thank the following people for their valuable contributions:

Corey Chatis
SLDS Grant Program, State Support Team

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

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

If the data are being used for different purposes, varying levels of data quality may be needed for those purposes. For example, data being used for research must provide a sufficient, representative sample from which to draw conclusions with a reasonable degree of confidence. The research can still be performed if the individual-level data are not all accurate and complete. However, using the same data to calculate a cohort graduation rate would require a far higher level of effort to ensure the data are complete and of very high quality. In this case, individual records make a difference in reporting results that have high stakes attached to their use.

A Proactive Focus on Quality

Since the scope of an SLDS frequently expands over time, there are continuing opportunities to improve the quality and completeness of new data before they are incorporated into the system. Data governance establishes processes for managing data quality, availability, security, and other factors, as well as roles and responsibilities for those involved in data management. Use the data governance process to engage the data steward and the data policy committee representative for the new data to discuss and document

- the timeline for the inclusion of the data;
- how the data have been used in the past;
- the intended use(s) of the data by SLDS users; and
- any data quality concerns or other limitations on the data, such as how many years of historical data can be included and any changes in the definition or code set over time.

Strategies to Improve Data Quality

- ✓ Provide data quality training to data submitters.
- ✓ Publish the process and annual timeline for all collections and/or extracts into the SLDS (e.g., initial due date, validation window, final sign-off deadline, release of reports).
- ✓ Publish user-friendly metadata for all data collected. Include non-technical definitions, code sets, the collection/submission calendar—including the validation window—and uses of the data.
- ✓ Establish a communications plan for all data submissions and extracts.
- ✓ Ensure data submitters have adequate notice regarding when data are due and when data will be used for which purposes.
- ✓ For snapshot data, provide a validation window and user-friendly reports that allow submitters to review counts, percentages, and drill down to individual records—if applicable—before the data are finalized.
- ✓ Put automatic edit checks in place that cross-check the same field from year to year and against other related fields.

Following this initial review, a project plan and communication plan for the inclusion of the new data can be developed. The project plan should include steps to address the data quality concerns identified above. The communications plan will help ensure those who submit the data to the source program or agency are aware that these data will be included in the SLDS, on what timeline, and for what purposes. This plan can also establish a means of communicating questions or concerns from data submitters.

Improving the Quality of SLDS Data

There may be data elements or entire data collections within the SLDS that cannot currently be used in analyses because the data steward has determined that they are of insufficient quality. If the SLDS leadership team or data governance group has prioritized a critical policy or research question that requires the use of these data, consider developing a strategic approach to improving the quality of these data over time. This approach might include the following steps:

- Work with the relevant data steward(s) to determine
  - why the data are currently of low quality (for example, they have never been used by the data providers, the providers are unclear about the definitions of elements, the data are redundant with another collection, there is a problematic timeline for submission, etc.); and
  - the appropriate level of quality for the data to be used for the identified purpose(s), including metrics to assess the quality.
- Identify strategies to address each cause of the poor data quality, including a timeline and responsible parties.
- Execute the strategies.
- Implement the data quality metrics to gauge improvement and to determine when the data have reached adequate quality for use.

Strategies to Improve Data Completeness

- ✓ Contact data submitters who have not provided data—or have provided it incompletely—to find out why. Are there resources or training that could address their barriers to submitting complete data?
- ✓ Put automatic edit checks in place that compare current year counts with previous years to proactively identify data submitters who are not on track to submit complete data.
- ✓ Publish the aggregate data on a public website with a clear indicator for entities that did not submit data.

Additional Resources

Growth Models: Issues and Advice from the States
http://nces.ed.gov/programs/slds/pdf/guide_growth-model.pdf

Techniques for Analyzing Longitudinal Administrative Data
http://nces.ed.gov/programs/slds/pdf/Data-Use-Issue-Brief-4_Analysis-Techniques.pdf

Turning Administrative Data into Research-Ready Longitudinal Datasets
http://nces.ed.gov/programs/slds/pdf/Data-Use-Issue-Brief-3_Research-Ready-Datasets.pdf

Workshop Summary: Developing a Data Use Strategy
http://nces.ed.gov/programs/slds/pdf/data_use_strategy.pdf