Assessing Alternative Sources of Data for Education Research

Sallie Keller (Project Lead),
Director Social and Decision Analytics Division
Distinguished Professor in Biocomplexity,
Professor of Public Health Sciences, School of Medicine
https://orcid.org/0000-0001-7303-7267
sak9tr@virginia.edu

Kathryn Schaefer Ziemer,
Kyle Angelotti,
Berk Norman, https://orcid.org/0000-0002-8156-3467
Mark Orr, https://orcid.org/0000-0001-7950-8752
Bianica Pires, https://orcid.org/0000-0002-4710-4849
Aaron Schroeder. https://orcid.org/0000-0003-4372-2241
Stephanie Shipp https://orcid.org/0000-0002-2142-2136

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

December 13, 2016

Funding: This evaluation was partially funded by the U.S. Census Bureau under contract to the MITRE Corporation

Citation: Ziemer, K., Keller, S., Angelotti, K., Norman, B., Orr, M., Pires, B., Schroeder, A., Shipp, S. (2016). Assessing Alternative Sources of Data for Education Research. Proceedings of the Biocomplexity Institute, Technical Report. TR# 2021-061. University of Virginia. https://doi.org/10.18130/bxcv-8h44.

Abstract

A wealth of alternative sources of education data, such as administrative and commercial data, are now available to researchers. These data have certain benefits over traditional survey research, including timeliness, lower cost, larger samples, geographical granularity, and longitudinal tracking. However, these data also present several challenges and require a different approach to data discovery, acquisition, and processing. These challenges include 1) identifying relevant data, 2) determining the usefulness of these data for answering specific research questions, 3) choosing among the data sources, 4) acquiring the data within a reasonable timeframe, and 5) assessing the quality of the data. This paper presents potential solutions and methodologies for addressing each challenge. Potential solutions include developing a data inventory, developing of a metric, and collecting additional detailed information in order to filter the data sources and determine which source(s) to target. In addition, we provide examples from an education project with the U.S. Census Bureau to highlight each challenge and solution. The purpose is to provide researchers with methods that better enable them to use alternative education data in research.

Asessing Alternative Sources of Data for Education Research

Kathryn Schaefer Ziemer, Sallie Keller, Kyle Angelotti,
Berk Norman, Mark Orr, Bianica Pires, Aaron Schroeder, and
Stephanie Shipp

The wealth of data now available has the potential to transform educational research, but also reveals the need for strong data acquisition and processing methodologies. Where the typical education research scenario once relied on the analysis of a standalone survey or data collection, these alternative sources of data, such as administrative and commercial data, are often cheaper, faster, more detailed, and lack many of the limitations of surveys. However, these alternative sources require a different approach to data discovery, acquisition, and processing. This paper explores the challenges and principles of data acquisition and processing in this new data environment.

Alternative data sources offer numerous benefits for educational research, some of which may include larger samples, longitudinal tracking, geographical granularity, timeliness, and lower cost. One example includes the statewide longitudinal data systems (SLDS) which contain administrative information on all students enrolled in public school in the state. These data are not samples, but are a census of the entire population of public school students. Moreover, the SLDS data can be linked across years which provides a way to track students longitudinally as they progress through the education system. Some data sources also identify the school each student attends, providing more detailed geographic granularity than survey data. Another example, the National Student Clearinghouse (NSC), provides student enrollment information for participating institutions in higher education and includes information on financial aid and students' majors. In addition, commercial data sources, such as Location Inc., are updated in real time which allows for more timely data. Finally, some of these data sources are free or available

at low cost. For instance, Washington state allows researchers access to the SLDS data free of charge whereas other states, like Kentucky, charge for the SLDS data. While alternative education data sources offer a number of benefits, these benefits also come with challenges.

In our experience, researchers face five major challenges in using these alternative data sources that have not been the result of research-based designed data collections such as surveys. These challenges include 1) identifying relevant data, 2) determining the usefulness of these data for answering specific research questions, 3) choosing among the data sources, 4) acquiring the data within a reasonable timeframe, and 5) assessing the quality of the data. In the following sections, each challenge is discussed in greater detail along with potential solutions. Figure 1 illustrates the challenges and potential solutions. Each solution for challenges 1 to 4 narrows the options for which data sources to pursue and challenge 5 addresses the quality of the data in the context of the selected use. The examples we present are primarily from a project for the U. S. Census Bureau (Keller et. al., 2016). A case study based on the Census project is provided in the Appendix.

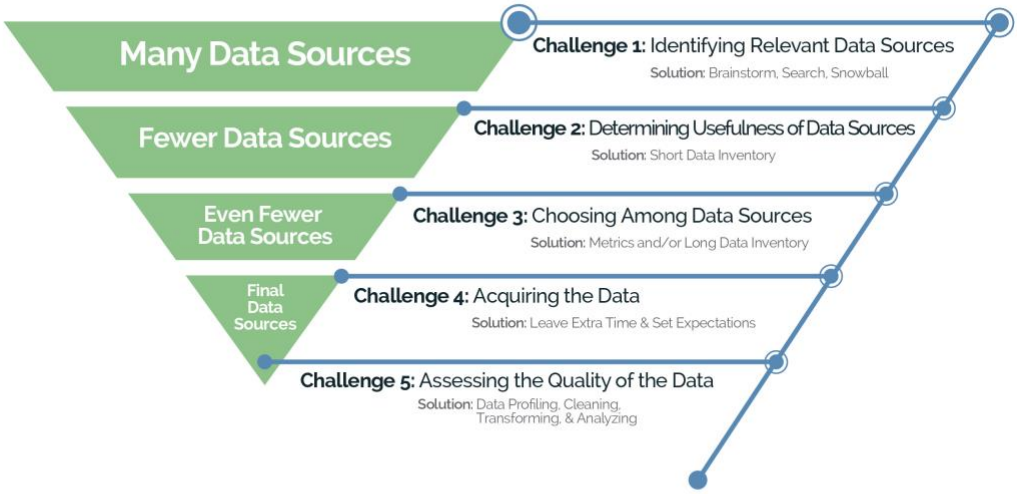


Figure 1. Challenges and potential solutions for choosing and using alternative education data sources in research. Challenges 1 to 4 narrow the choices for potential data sources. Challenge 5 addresses the quality of the data in the context of the selected use.

Challenge 1: Identifying Relevant Data

When identifying relevant alternative data, the researcher must consider a number of different factors, including the potential data source agencies, the appropriateness of their data collection methods and respondents, data access rules, policies and procedures, and the potential for integrating the various data. This process differs from traditional education research that uses a custom designed instrument in which the main concern involves reaching the appropriate respondents. Many potential barriers could prevent the actual acquisition or usefulness of the alternative data. Therefore, the researcher should in the beginning cast the broadest net possible to identify as many potential data sources as possible.

Meeting this first challenge involves generating a list of potential data sources that the researcher may be able to use to answer the research questions. The research questions need to drive this search for data sources. For example, the Census Bureau project research questions focused on comparing the American Community Survey (ACS) with education data sources external to the federal statistical system. The project aimed to find alternative education data sources (commercial, administrative, and international) that had the same or similar education variables as those in the ACS, such as student enrollment and educational attainment.

Several strategies can facilitate the search process. First, online searches using keywords, such as education data, K-12 data, higher education data, and workforce data, can provide a first pass. Second, the first search process can be the foundation for later searches where the data

sources found with the first search lead to other related sources or where the initial data sources are used to exclude or restrict terms found to have ambiguous meaning. Third, researchers may want to think about how other education stakeholders might find information. For example, how would parents find information on the schools their children attend, how would high school students discover information about colleges when applying to schools, and how would job seekers identified jobs and careers? Table 1 displays the list of alternative data sources discovered for the Census project.

Table 1. Alternative Education Data Sources - Commercial, Administrative, and International

Commercial	**Administrative**
College Board Donors Choose eSparks Glassdoor Great School Ratings LinkedIn Location Inc. Maponics Monster Resume Database National Student Clearinghouse School Attendance Boundary Information System School Digger US News and World Report Rankings	Statewide Longitudinal Data Systems (50 states and the District of Columbia)* • K-12 data systems (34) • Higher education data systems (3) • Workforce data systems (2) • Combination systems (21) **International** Organization for Economic Co-operation and Development • Program for International Student Assessment (PISA) • Teaching and Learning International Survey (TALIS)

* Some states have separate SLDS systems for K-12, higher education, and/or workforce, whereas other states have combination systems that link across these areas.

Challenge 2: Determining Usefulness of the Data

Once the researcher gathers potential data sources, the next step involves looking more closely at each source to further assess its usefulness for research purposes. How the researcher defines and assesses usefulness will depend upon the specific research situation, including aspects of the data and aspects of the research application. For instance, aspects of the data may

include the lack of code books or an unstructured format, and aspects of the research application may include restricted access due to proprietary information, all of which affect the usefulness of the data.

The extent to which the data have the potential to provide insight into the specific research questions for the research application should be the primary driver in assessing the usefulness of the data. Creating a short inventory process based the key factors for distinguishing data sufficiency based on information relevant to the particular project helps to focus efforts on this goal. The inventory process needs to be rigorous to provide a sound structure and process, but flexible to not be too time-consuming.

The short inventory process consists of two stages. First is the creation of the inventory. Researchers should determine key factors that will help determine the usefulness of a data source and distinguish one source from another. In this stage, researchers should develop a list of definitions for key factors related to the metadata of each data source. Guidance on the factors and definitions can be obtained from the literature and then refined to be as comprehensive as possible for the intended purposes of the data and research. Table 2 presents the short inventory of important factors, definitions, and rationale used for the Census project. Researchers may wish to modify that table to suit their needs.

Table 2. Short Inventory Factors and their Definitions, and Rationales

Key Factors	Definition	Rationale
Purpose	Purpose of the organization collecting the data and the reason for collecting the data.	Purpose may affect the quality and type of information collected, e.g., advocacy.
Method	Method of data collection and the raw source of the collected data.	Errors or bias in the data and the information reported may differ depending on the method and the raw source.
Description	Variables included in the dataset and the dates for which the data are available.	To determine whether the dataset has the variables they need to answer the research questions.

Key Factors	Definition	Rationale
Selectivity	Unit of the data, universe of the data, sampling technique used, and coverage of the data.	To determine whether the data are representative enough and provides enough granularity to answer the research questions.
Accessibility	Process for acquiring the data, cost, and any variables collected but not included in the data.	To determine whether the data are available for research purposes, and whether the timeline and cost fit with the timeline and budget of the study.

Source: Definitions adapted from AAPOR (2015), Iwig et al. (2013), and UN (2014).

Second, researchers need to gather specific information about the key factors for each data source under consideration. Finding the information to complete the inventory may also prove challenging and may require an iterative process. For the Census project, web searches often provided sufficient information to answer questions related to the purpose and description of the data. However, online information often did not provide detailed information regarding the accessibility of the data and the process for obtaining the data. In these cases, researchers may need to contact the company or organization directly and engage in multiple conversations with different staff members to obtain the necessary information. Since the information gathered may vary depending on the staff member, it is often most useful to talk to the technical staff who worked directly with the data.

Challenge 3: Choosing Among the Data Sources

The short inventory can help narrow the number of data sources that would be useful, however, further filtering may be necessary to determine which sources to pursue. For the Census project, the short inventory helped identify state SLDS systems as good data sources to answer our research questions. However, the project did not have the budget or timeline to pursue all 50 states' SLDS and therefore required a method for determining which state's SLDS to pursue. Two ways to narrow the choices are 1) developing a metric and 2) collecting additional important information on the data sources.

Developing a Metric. Developing a metric involves identifying a question, or set of questions, from the inventory that are most important for the research and then determining the data measures that could potentially be used to answer the question(s). This method is especially useful when the researcher has identified a large number of potential data sources. From the short inventory, selectivity and accessibility were the most important factors for the Census project and hence used as metrics to guide our data source selection process (see Table 2). Selectivity was one metric since the research questions required access to K-12, higher education and workforce data. As a result, data sources were categorized according to the universe of the data represented (i.e., K-12, K-12 linked with higher education, or pre-kindergarten through workforce; see Table 3).

Accessibility was also deemed a metric since the project required student-level data. Therefore, data sources were further categorized based on which sources presented challenges to gain access to student-level information, as represented by a purple diamond in Figure 2. The results indicate a great degree of variability regarding the selectivity and accessibility of states' SLDS. Researchers can use a similar method to categorize and rank potential data sources.

Table 3. Virginia Tech's Categorization of states' SLDS

Categories	Description	Number of States
1	States with P-20W systems (e.g., linkages across K-12, higher education, and workforce)	15 (light green)
2	States with linkage across K-12 to higher education	15*
3	States with linkage across K-12	19
4	States that do not have SLDS or are in the process of creating an SLDS	2 (dark green)

*Includes the District of Columbia
Source: Virginia Tech, Social and Decision Analytics Division; SDAD (2016)

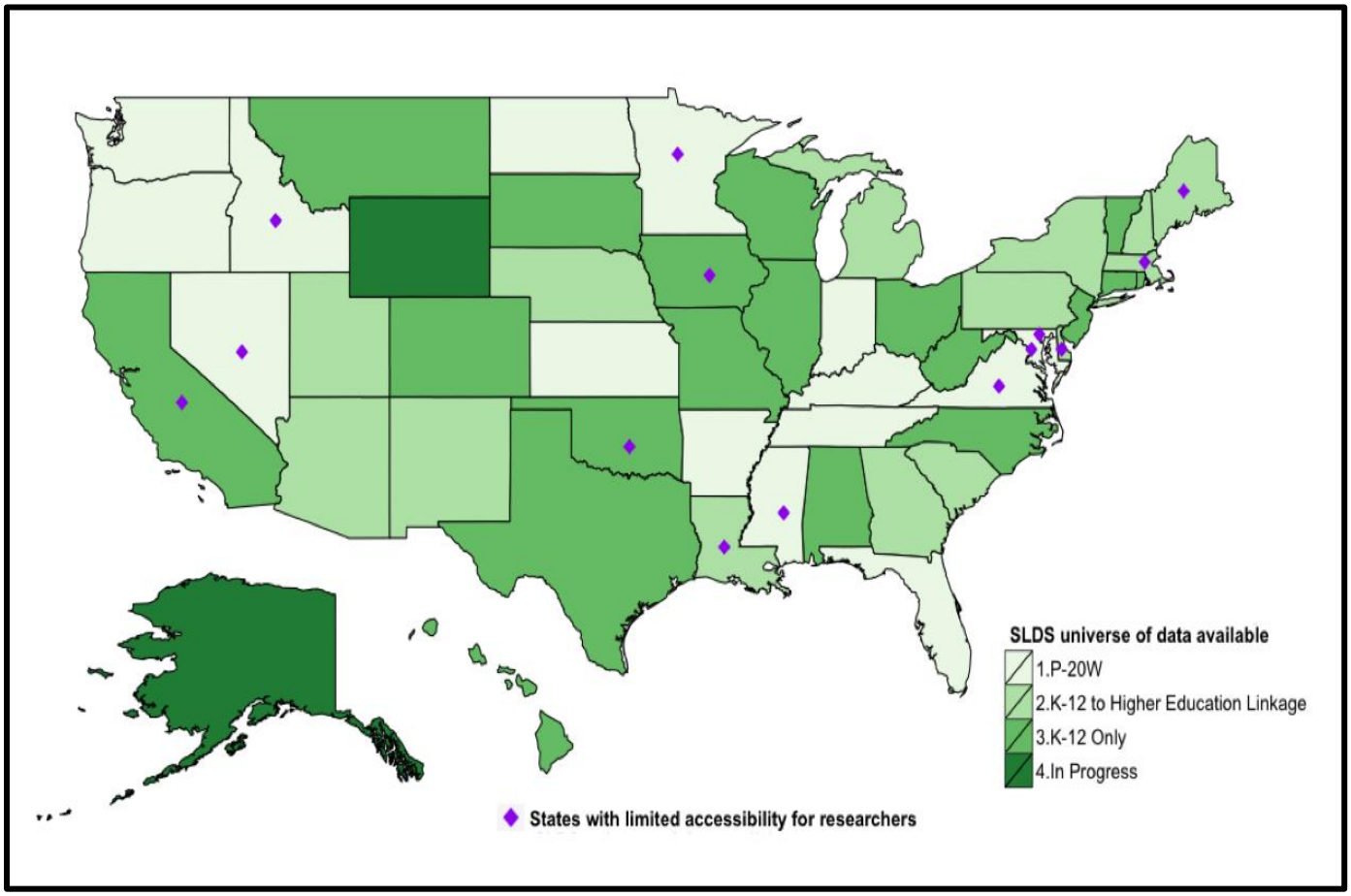


Figure 2. Virginia Tech's metrics of SLDS displayed across the 50 states. Purple diamonds represent the states with limited data accessibility for researchers. Source: Social and Decision Analytics Division (SDAD), 2016.

Collecting Additional Information. In addition to or instead of creating a metric, researchers can collect additional important information on the data sources under consideration. Researchers can collect this information in the same manner as the short inventory, and can be thought of as a "long inventory". This represents a useful method if new barriers or benefits arise when conducting the initial short inventory. For the Census project, we added questions under existing key factors in the initial inventory as well as added new factors. For the existing key factors, we added questions about the cost of the data and the timeliness of receiving the data to the accessibility factor. Answering these additional questions ensured that the project stayed

within budget and finished on time. Table 4 presents the key factors added to the long inventory and the rationale for including each one. The ultimate purpose of collecting this additional information is to identify nuances about the data sources that differentiate sources that should be pursued first from those that should not.

Table 4. Long Inventory Key Factors and their Definitions and Rationales

Key Factors	Definition	Rationale
Metadata	Unique IDs, codebooks or data dictionaries, and any other information that could be used to assess the soundness of the data.	Necessary for cleaning and interpreting the data.
Stability/Coherence	Changes in the universe of data captured, data capture method, or sources of data.	To determine the extent that data can be compared across time spans or to other datasets.
Accuracy	Quality control checks and known sources of error in the data.	To determine whether the data are high enough quality for the research.
Privacy and Security	Legal restrictions, informed consent, and confidentiality policies (e.g., FERPA, MOUs)*.	This can affect the amount and type of data that are made available and the timeline to receive the data.
Research	Research that has been conducted with this data or data source.	Provides information about what has already been done and the kind of studies that are possible with the data.
Gaps and Concerns	Anything that could affect the research that was not already captured.	Can be tailored to fit researchers' specific needs.

*FERPA=Family Educational Rights and Privacy Act; MOU=Memoranda of Understanding

Challenge 4: Acquiring the Data

Once the researcher chooses which data source(s) to pursue, the process of obtaining the data begins. This process can present additional, unanticipated challenges, even if the researcher has already deeply investigated how to access the data. Challenges may include delays in receiving the data, limitations in the amount of data that can be requested, and restrictions to protect the privacy of the data.

If the organization providing the data does not have a streamlined process for sharing the data, the time it takes to receive the data can increase significantly. For instance, the organization may have limited resources to devote to this effort. Even when organizations have an established data sharing process, delays may still occur due to review boards that meet infrequently or requested revisions to the research application.

The breadth of the data request can also pose problems if the organization considers it too large. For instance, some of the state SLDS systems limit data requests to specific student groups and do not share data on students from all grades. In addition, organizations may have concerns that possessing large amounts of data on individual students enable personal identification. Similarly, organizations may restrict or delay data due to privacy concerns. For instance, some SLDS systems require a separate request under the Freedom of Information Law (FOIL), which allows members of the public to access records from government agencies. This process requires multiple signatures within the department which further delays the data request.

Given these potential, unexpected challenges, it is important to allocate extra time for acquiring the data. In addition, talking to other researchers who have already used the data can help troubleshoot issues before they arise. This also helps set expectations regarding the data acquisition process. For more details on acquiring SLDS data, see the Census project case study in the Appendix.

Challenge 5: Assessing the Quality of the Data

The large majority of these alternative data sources are gathered for purposes other than research, therefore, once the data are received, researchers need to evaluate and re-purpose the data for their own research (Keller et al., 2016). Figure 3 outlines the steps involved in re-

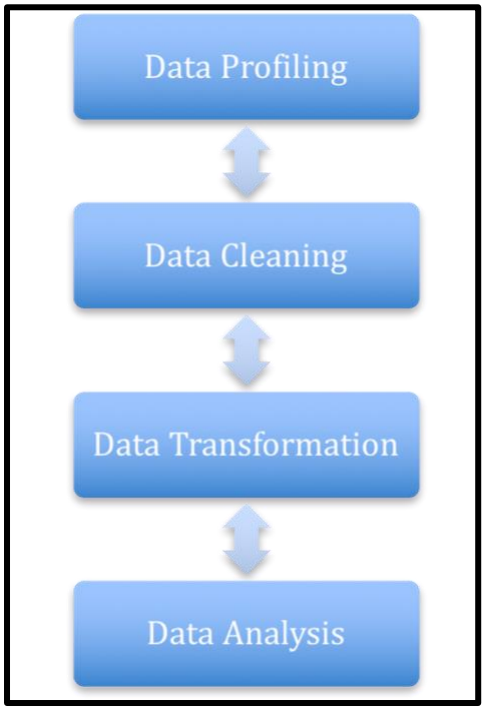
purposing the data and although the steps appear in linear order, the process is generally iterative.

Data profiling involves determining the quality of the data for the intended research purposes. In this data model, only the variables needed to answer the research questions undergo this iterative process. In this stage, the researcher discovers and documents any issues with the data, but does not actually fix or clean the data. This involves assessing the completeness (extent of missing data), value validity (extent that the data had proper values), consistency (the degree to which data values align across years or other datasets), uniqueness (number of unique values), and duplication (replication of observations) of the data (see Keller et al., 2016 for further details). The researcher often needs to communicate with the data provider to address any irregularities in the data.

Data cleaning involves identifying which issues needed to be fixed and deciding on the data cleaning rules (e.g., decision rule for removing duplicates). Data transformation involves restructuring the data in a format that is conducive the statistical analysis. For example, if separate datasets exist for each academic year, the researcher will need to merge the datasets to conduct longitudinal analyses. Only after the profiling, cleaning, and transforming of the data can researchers conduct the data analyses. For an example of the data quality process, see the

case study in the Appendix.

Figure 3. The steps involved in working with the data after acquisition.



Conclusions

This paper presents potential challenges when using alternative data sources to traditional statistical surveys or other research data collection, as well as guidelines for addressing these challenges. Examples from an education project with the U.S. Census Bureau highlight the data discovery, acquisition, and processing steps. In addition, the SLDS serves as an example of an alternative education data source.

The project research questions must guide all steps of the process, from the identification of relevant data sources to determining which data source to acquire. Tools such as data inventory, development of a metric, and collecting additional detailed information can help researchers filter the data sources to determine which source(s) to target. Even when researchers

identify the data source to acquire, additional challenges may occur, such as a lengthy acquisition process or denied data requests. Successful data requests can still result in unexpected surprises in the data quality and structure. Overall, alternative education data, such as SLDS, represent a rich and relatively untapped resource that researchers can harness. The steps presented in this paper address several challenges that researchers may encounter when using these data and present methods to better harness this opportunity.

References

American Association for Public Opinion Research. (2015). AAPOR Report on Big Data. AAPOR Big Data Task Force.

Iwig, W., Berning, M., Marck, P., & Prell, M. (2013). Data Quality Assessment Tool for Administrative Data. Washington, DC: Federal Committee on Statistical Methodology.

Keller, Sallie, Stephanie Shipp, Mark Orr, Dave Higdon, Gizem Korkmaz, Aaron Schroeder, Emily Molfino, Bianica Pires, Kathryn Ziemer, and Daniel Weinberg. (2016). Leveraging External Data Sources to Enhance Official Statistics and Products. Report prepared for the U.S. Census Bureau. Social and Decision Analytics Division (SDAD), Biocomplexity Institute of Virginia Tech.

United Nations. (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team.

Appendix – Case Study of Acquiring and Processing SLDS Data

The following case study illustrates the outcomes of working through the five challenges for acquiring alternative education data. The case study is based on a collaborative research project conducted by the Social and Decision Analytics Division (SDAD) and the U.S. Census Bureau (Keller et al., 2016).

Challenge 1. Identifying Relevant Data

The data source discovery process was driven by the research question, "For which American Community Survey (ACS) education questions can non-federally collected alternative data sources be obtained?" Online searches revealed that statewide longitudinal data systems (SLDS) contained administrative data with similar education variables as those in the ACS, such as student enrollment and educational attainment. However, there were other data sources as well, such as Location, Inc., that also had potentially useful information for the project. Therefore, we needed to determine which data source from our discovery process would be the most useful.

Challenge 2: Determining Usefulness of the Data

A short inventory was completed for each of the potential data sources (see Table 1 above for the criteria used). Gathering this information was an iterative process. Web searches often provided sufficient information to answer questions related to the purpose and description of the data. However, more detailed information, especially regarding the accessibility of the data and the process for obtaining the data, was often not provided in the online information. In these cases, we contacted the company or organization directly, often engaging in multiple

conversations with different staff members to obtain the necessary information. Based on the short inventory, we determined that the SLDS data sources would be the most useful since they had the most complete education information for comparison to the ACS. However, the 50 states and the District of Columbia each have their own SLDS system(s), so we needed a way to determine which SLDS data source to use.

Challenge 3: Choosing Among the Data Sources

In order to narrow the options for the SLDS data sources, we created a metric from the short inventory and collected additional information using a long inventory. The creation of the metric is described in the paper and highlighted in Table 3 and Figure 2. The additional information collected is illustrated in Table 4. Although all states except for two have a functional SLDS systems, we found that 14 states (including the District of Columbia) presented challenges for researchers to access student-level data (identified by the purple diamonds in Figure 2). Some of these challenges included not releasing student-level data or releasing very limited data, providing on-site access only, requiring an internal sponsor (e.g., collaborator within the state education department and only accepting data requests that aligned with the state's research goals, such as in the Commonwealth of Virginia). Moreover, some states informed us that they were not currently taking any researcher data requests (e.g., District of Columbia, Maryland, Maine), but likely would in the future.

The categorization of SLDS systems and the long inventory helped us determine which data sources to pursue for our project. For instance, we did not pursue states that had limited data accessibility for researchers. In addition, the cost and timeliness of receiving the data drove which states to pursue so that we could stay within our budget and complete our project on time.

We discovered that states could not always deliver the data within the time frame promised due to their prioritization of requests or requirements to obtain signatures from many agencies before releasing the data. For these reasons, we decided to pursue acquiring data from Kentucky, North Carolina, Texas, Virginia, and Washington state SLDS systems. The process for Kentucky and Texas are described below.

Challenge 4: Acquiring the Data

Kentucky required us to select files and variables for our study. Choosing these variables was an iterative process with the Kentucky staff recommending what data we could obtain within the timeframe of our project. Upon signing a Memorandum of Understanding, Kentucky provided us with the requested student-level data for pre-school through higher education for the years of 2009-2014. None of the data were suppressed or removed, therefore the dataset represented the complete population of students enrolled in public school in the state during that time period. The higher education data included students from both in-state public school as well as independent institutions. Personally identifiable information (e.g., birthday, names) was not included. However, students had unique identification (ID) numbers that allowed for linking individuals across years. Overall, we received 20 separate datasets (see Table 5). The Master Demographics dataset could be linked with the other datasets via the unique student ID. In addition to the student ID, all of the preschool and K-12 datasets also included the school name, school district, academic year, and grade of the student. The higher education datasets included the institution name and academic year.

Table 5. Datasets and examples of variables received from Kentucky

Dataset	Example Variables
Preschool	Enrollment start/end date, Head Start indicator
Early Childhood	Assessment type, Assessment score at exit
K-12 Annual Person	Days enrolled, Dropout code, Graduation code
K-12 Assessment Scores	Assessment score, Score percentile, Date of assessment
K-12 Courses	Course name, Grade score, Honors indicator
K-12 Person Enrollments	Enrollment status (e.g., first time in state), Start date
K-12 Schools	NCES school number, Superintendent, County name
K-12 Special Education	Primary disability code, Full funding eligible indicator
K-12 Title I	Enrollment start/end date
K-12 Transcript	Course name, Course grade, GPA, Score percentage
Higher Education Annual Person	High school of graduation, High school GPA
Higher Education Cohort	Degree sought, Entry age, Residency status
Higher Education Degree	Year degree earned, Degree level, Major
Higher Education Enrollments	Residency, Degree sought, First time student indicator
Higher Education Financial Aid	Expected family contribution, Total income
Higher Education Institution	Institution, Sector code (4-year, 2-year)
Higher Education Readiness	ACT score, SAT score, Readiness status
Technical Education Credentials	Credentials (e.g., associate degree, GED)
Technical Education Enrollment	Education level, Attendance hours, High school name
Master Demographics	Race/Ethnicity, Gender, Birth year

For Texas, we obtained selected variables for student-level data for preschool through 12th grade for the years 2009-2013. The Texas staff guided us to select as few variables as possible to maximize the amount of data we could receive. Texas' interpretation of the Family Educational Rights and Privacy Act (FERPA) suppresses cells with counts fewer than 5 students. As a result, we received approximately 1/5th of the total public school population in the state. Moreover, the data received were not a representative sample across the state since the selection was based on suppression rules rather than random sampling. We could have received a larger and more representative sample of students if we had requested fewer student-level demographic variables. Table 6 includes the datasets and examples of the variables that we received from Texas. The Student Demographics, Discipline, and three Assessment datasets included student-level data and unique

student IDs that could be linked across years. The Class, Employment, and Non-Class Employment datasets included detailed information about school staff, such as names and salaries.

Table 6. Datasets and examples of variables received from Texas

Dataset	Example Variables
Assessment – Academic Readiness	Student ID, Grade, Sex, Ethnicity, Scores
Assessment – End of Course	Student ID, Grade, Sex, Ethnicity, Scores
Assessment – Knowledge and Skills	Student ID, Grade, Sex, Ethnicity, Scores
Class	Teacher Name, Course Taught, Type of Student Served
Course	Course Name, Year Course Offered
Discipline	Student ID, Type of Disciplinary Action
Employment	Staff Name, Tenure, Total Pay, Employment Type
Non-Class Employment	Staff Name, Tenure, Total Pay, Employment Type
Student Demographics	Student ID, Grade, Race/Ethnicity, Gender

Challenge 5: Assessing the Quality of the Data

Our research questions for our project with the Census Bureau guided the choice of variables targeted for profiling, cleaning, transforming, and analyzing. For Kentucky, we focused on the variables of race/ethnicity, gender, birth year, grade, student identification number, district code, year, high school dropout code, high school dropout reason, high school graduation indicator, and limited English proficiency indicator. These variables corresponded to similar information collected by the Census Bureau for education. Overall, the variables were of high quality with less than 0.2% duplicates and less than 1% missing values. Two transformations were performed: we computed age based on students' birth years and added counties to the dataset by matching school districts with the county Federal Information Processing Standard (FIPS) code.

For the Texas data, we profiled the same variables with the addition of a variable for economically disadvantaged status. Since we received only 1/5th of the total public school population, we weighted the data to make it more representative. However, the weights did not

solve the problem because the weights could not apply to empty cells. Of the data received, the variables were of high quality with fewer than 2% duplicates and less than 1% missing values. The transformations included adding county information.

About Virginia Tech's Social and Decision Analytics Division

The **Social and Decision Analytics Division (SDAD)** is a leading Division in the Biocomplexity Institute at the University of Virginia. The Biocomplexity Institute is at the forefront of a scientific evolution, applying a deeply contextual approach to answering some of the most pressing challenges to human health and well-being within our changing environment. SDAD was created in the fall of 2013 to extend Biocomplexity Institute's capabilities in social informatics, policy analytics, and program evaluation. The researchers at SDAD form a multidisciplinary team, with expertise in statistics, policy and program evaluation, economics, political science, psychology, computational social science, and data governance and information architecture. SDAD's mission is to embrace today's data revolution, developing evidence-based research and quantitative methods to inform policy decision-making and evaluation.