# **PART 3:** Processing the Data

After you receive the data you requested, you may find the next step most challenging of all: You must analyze it. As was clear in Part 2, many data requests can result in very large datasets that can be challenging to analyze.

This toolkit can't teach you how to conduct statistical data analysis, but it can help you understand, in broad terms, what data analysis entails. You should prepare for possible technical and legal hurdles in gaining access to the data, the challenges of organizing the data, and the implications of working with a partner to analyze the data or analyzing the data yourself.

This part of the toolkit takes you through two scenarios:

**1.** The Easiest Path: The good news is that the agency that provides the data can sometimes do most of the analysis for you. In this scenario, for example, the agency may send you a spreadsheet containing the percentage of students enrolled in arts courses, broken out by every school and district in the state. The agency has done most of the analytical work, yet you still need to:

- Understand the spreadsheet and make simple calculations.
- Anticipate the challenges of data that have already been analyzed.
- Ensure the quality of your data.

**2.** The More Challenging Path: You receive a raw, unanalyzed data set that contains individual student records. Unless you, or one of the partners you identified in Part 1, are an experienced data analyst with access to powerful statistical software, you need a professional with the expertise to help you:

- Get access to the data.
- Organize the data.
- Process the data.

If you are not an expert in data analysis, you may need help to pursue scenario one, and you will certainly need help to pursue scenario two. If you even need to ask whether you can do this work alone, then you probably can't. This part of the toolkit will help you understand what kind of help you need.

If you would like to explore data analysis in greater depth, see the following tools:

- **Tool I**: Sample Record Layout
- **Tool J**: Common Components of a Data-Sharing Agreement
- **Tool K**: Common Means of Getting Access to Large Data Files
- **Tool L**: Analyzing Data From Multiple Sources

# 1. The Easiest Path

In this scenario, the agency will probably send you the data in a widely used format like an Excel spreadsheet, and you may need to conduct only minimal additional analysis on your own.

## Understand the spreadsheet and make simple calculations.

Let's say the data provider provided a spreadsheet with data on arts course enrollments at every school in the state, broken out by grade level. Below is a small portion of that spreadsheet showing arts course enrollments in the (fictitious) Acme High School, which bears the (equally fictitious) state ID# 03476. In this case, each row (or "record" in data speak) gives enrollment counts and percentages for students in the school for each grade.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **SCHOOL NAME/ ID # (1,2)** | **GRADE (3)** | **TOTAL GRADE ENROLLMENT (4)** | **TOTAL ARTS ENROLLMENT (5)** | **VISUAL ARTS ENROLLMENT (6)** | **MUSIC ENROLLMENT (7)** | **DANCE ENROLLMENT (8)** | **THEATRE ENROLLMENT (9)** | **% ENROLLED IN ARTS (10)** |
| 2 | **ACME HIGH SCHOOL/03476** | 9 | 75 | 51 | 20 | 15 | 13 | 11 | 68.00 |
| 3 | **ACME HIGH SCHOOL/03476** | 10 | 66 | 61 | 39 | 19 | 15 | 0 | 92.42 |
| 4 | **ACME HIGH SCHOOL/03476** | 11 | 72 | 67 | 37 | 20 | 0 | 18 | 93.05 |
| 5 | **ACME HIGH SCHOOL/03476** | 12 | 88 | 64 | 35 | 23 | 14 | 2 | 72.73 |
| 6 | **ACME HIGH SCHOOL/03476** | ALL | 301 | 243 | 131 | 62 | 42 | 31 | 80.73 |

Most of the analysis is already done. Row 6, Column I tells you that roughly 81% of Acme High School students are enrolled in arts courses, for example. Using the spreadsheet's formula functions, you could perform simple additional analyses, such as calculating the percentage of Acme High School 10th graders enrolled in different disciplines. Moving across Row 3, you can determine the percentage of all 10th graders enrolled in any arts class (61/66 = 92.42%), visual arts classes (39/66 = 59.09%), music classes (19/66 = 28.79%) and so on.

TIMELY TIP

If the number of metrics you can report on seems limited, you may be able to perform simple calculations to develop additional metrics. For example, divide total school or district enrollments by the number of arts teachers to calculate average student/teacher ratios in the arts. Or divide the total number of students who take arts courses by the total number of students who have access to those courses to calculate a rough measure of what share of students embrace their schools' arts offerings.

To get percentages of students enrolled in the arts for an entire school district, simply add up enrollment numbers for each school in the district and divide the sum of arts enrollments by the sum of total enrollments. Spreadsheet programs like Microsoft Excel typically allow you to make such calculations in bulk.

STUMBLING BLOCK

Beware of subtotals when you add up enrollment numbers for all grades or all schools. In the example here, Rows 2 through 5 each contain data for one high school grade. Add them together, and you will get total enrollments for all grades in the school, which is already contained in Row 6. If you were to add Rows 2 through 6 together in this example, you would get twice the total enrollments. Repeat this mistake across hundreds or thousands of schools, and you will seriously skew your results.

Spreadsheet programs allow you to filter and sort your data quite easily to isolate the rows you need. If you are unsure of how to do this, ask someone who is proficient in Excel to help you.

## Anticipate the challenges of data that have already been analyzed.

Even if most of the analysis has been done, your spreadsheet may present challenges:

**THE COLUMN HEADINGS MAY CONFUSE YOU.** The headings of columns can often seem cryptic, because data systems often abbreviate them to save space. For example, "cbkaat" may mean the total count [or number] of black or African American students, and "caspit" may mean the total count of students who are Asian or Pacific Islanders.

Look for a spreadsheet or tab in your spreadsheet that contains a glossary or "record layout" that translates the column headings into standard English. This record layout will also define the type of data in each column — e.g., alphanumeric (which means text or numbers), numeric (which means numbers only), percentage, etc. If you can't find a spreadsheet or tab like that, ask the data provider to supply one. You have no room for error in interpreting your column headings. (To see a sample glossary or record layout for the spreadsheet above, see Tool J.)

**IF YOU HAVE LARGE DATASETS, YOU ARE MORE LIKELY TO MAKE MISTAKES.** Even in a small state like Rhode Island, which has roughly 300 schools, simply breaking data out by grade level will yield thousands of rows. In Illinois, with its more than 4,200 schools, breaking out the data in that way would yield tens of thousands of rows.

**DEVELOP A CLEAR PICTURE OF THE SPREADSHEET'S FORMAT AND ORGANIZATION BEFORE YOU BEGIN WORKING ON IT.** Before you can perform bulk calculations, you may have to sort the spreadsheet by grade bands, filter by key characteristics or add together certain rows while excluding others. Again, if you are uncomfortable manipulating spreadsheets, you should bring in someone with more expertise.

**DATA SUPPRESSIONS CAN LIMIT YOUR REPORTING OPTIONS AND THROW OFF YOUR ANALYSES.** In Part 2, we learned that state agencies suppress data in cells where the numbers are so small that publishing them could reveal private information about individual students. Examine your spreadsheet for evidence of such suppressions. In cells where data have been suppressed, you will see symbols like --, *, or ƚ in place of numbers.

You may find, for example, that much of the arts enrollment data for students of color at the school level are missing, which would make it impractical to report on individual schools' arts enrollment data by race or ethnicity. In addition, your calculations cannot include cells that contain suppression symbols, because the program will either misread them as zeros or simply refuse to perform the calculation.

**VERSION CONTROL IS CRITICAL.** Always keep original master copies of all your spreadsheets so that you can return to them in case you accidentally delete or corrupt any data during your calculations. Also, carefully label any spreadsheets that contain any of your own calculations, and keep clear notes of what analyses you've conducted. It is all too easy to forget what calculations you have already made, and all too difficult to reconstruct them.

**STUMBLING BLOCK**

Many datasets have fields that can be confused for one another if you are not careful in processing the file. For example, many state education agencies (SEAs) maintain at least two fields recording school IDs in their student file structures, because students often have a school that qualifies as their neighborhood school (their "home" school) and another school they actually attend (their "serving" school). If you do not distinguish these fields at the outset and determine how to include them in your analysis, you may get inaccurate results. (For more information on potential problems defining your research terms, see Tool H, defining terms for data requests and analyses.)

## Ensure the quality of your data.

The data set you receive may contain errors. Data providers can make mistakes in their analyses, technological glitches can corrupt data or school districts can enter the data incorrectly, among other challenges. Some errors are unavoidable, but there are simple steps you can take to test the reliability of your data before you begin any analysis.

To do so, look for these signs of trouble:

**ARE THERE ANY DUPLICATE RECORDS?** Every database should have a "key," a field or combination of fields that identify each record uniquely. No two records can have the same value in the key column. For a spreadsheet that primarily identifies characteristics of all schools in a state, for example, the key column would contain each school's unique ID. No two records should have the same ID. It is simple to check for this error. Software programs like Microsoft Excel allow you to sort by fields and check for duplications that can corrupt your analyses.

**IS THE DATA FILE COMPLETE?** An incomplete data file can seriously undermine your analysis. There are fields where the absence of a value will cause problems. For example, you should not have blank values in your key fields. Missing rows or columns can skew your results. Test the contents of the data file against what you know about education in your state. For example, does the number of school records in the file closely match the number of schools in your state? Do state, district or school enrollment counts line up with enrollment counts reported elsewhere? Does the number of teachers in your data file line up with the number of teachers the state publicly reports? If data are missing, contact your data provider.

**DO THE DATA APPEAR TO FOLLOW THE RULES SET FORTH IN THE DATA DICTIONARY OR GLOSSARY?** As noted above, you should have a record layout describing the format of what's in each field; do the data comply with what's in that record layout? If the record layout says the data in a field should be a number, for example, but the field contains text, you will have a problem.

**MOST IMPORTANT, DO THE DATA PASS THE "SMELL TEST"?** Some problems might be glaringly obvious — elementary school enrollments in the tens of thousands, percentages that far exceed 100% or ninth grade enrollments than exceed total school enrollments, for example.

Other problems may be subtler – not impossible, but in defiance of what you know about arts education. For example, is the percentage of teachers who are certified much lower than you expect? Are rural students much more likely than suburban students to have access to dance courses? Do the data for schools or districts you're familiar with contradict what you know about them? These could be startling new findings, but they could also be mistakes – the result of column headings that accidentally changed places, for example.

If you encounter such problems, check the record layout again to see if you are misreading the columns. If you aren't, then don't hesitate to contact your data provider to check the accuracy of the data.

You may introduce new errors if you conduct additional analysis of the data in the spreadsheets. After you finish your analysis, take some time to perform the quality tests above a second time.

**STUMBLING BLOCK**

Chances are, your data set will include some problems you cannot easily correct, because people have entered data incorrectly at the school or district level. Clerical staff in a school or district may enter incorrect course titles in the arts, for example, recording "chorus" as "music – general" or "jazz dance" as "social dance."

In most cases, you will still be able to produce useful findings about access to, and participation in, courses in the broad arts education disciplines, even if information about individual courses within those disciplines proves unreliable. In other words, your findings about participation in music classes could be mostly reliable, even if the data on "chorus" classes are wrong.

By publishing the results of your analysis, you may provide schools and districts an incentive to enter the data more accurately. In most states, data on arts education have never seen the light of day, so that incentive didn't exist. As information becomes public, schools and districts don't want to have their efforts misrepresented.

## 2. The More Challenging Path

If the agency providing the data does not conduct any analysis for you but provides raw data instead, your task will be much more complex. For example, if you receive data with records on individual students, your data set will probably contain millions of data points, require you to observe very strict privacy protocols, require complex statistical software and demand more sophisticated analysis.

Even if you plan to retain an outside vendor or partner to do this more detailed analysis, you should understand in general terms what it takes to get access to, organize and process the data. These steps will have an important impact on your timeline and project plan.

### Get access to the data.

Organizations conducting research with education data gain access to those data only after satisfying strict requirements to protect students' privacy. This is especially true if those data contain unit records, or records containing information about individual students. The privacy of minors is protected by the federal FERPA law and additional state laws.

If your data request is approved, then you've demonstrated to the data collector that your organization meets the legal requirements for gaining access to student information protected under that law. Only organizations that conduct research to improve education and thereby benefit schools can gain access to such data.

Privacy protections will have a profound impact on the data-sharing agreements you or your partner organizations must sign to receive the data, how the data will be transferred to you and how you must manage the data after you receive them.

To receive data containing private student records, you must:

### FINALIZE A DATA-SHARING AGREEMENT.

Data-sharing agreements are legally binding agreements between the organization that maintains the data and the organization or organizations requesting access to them. Typically, the organizations that maintain the data draft data-sharing agreements.

If a partner or vendor is conducting your data analysis, it will probably also manage the process of executing a data-sharing agreement. Still, you should be aware of what a data-sharing agreement is and what impact it might have on your plans and timeline. (For an overview of what a data sharing agreement commonly includes, see Tool J.)

Legal representatives from the data provider will have reviewed and approved the language of the agreement before you receive it. If you can, ask a lawyer representing your interest to review the document as well. You will have to confirm that the descriptions of data use, access and disclosure reflect your understanding of the agreement.

**STUMBLING BLOCK**

Allot substantial time for the data-sharing agreement and data acquisition phase of your project. It can be tricky to predict how long it will take, so discuss these phases of the work with your data collection partners early. The legal review and data acquisition processes in some states takes as little as two to three weeks, and in others it can extend to six months or even a year!

Stay connected to this process. Follow up with your team and the agency regularly. Don't be demanding, but do be persistent. Sometimes a data provider will expedite the process just because they know the project matters to you and because you've taken time to build a relationship with them.

Data-sharing agreements don't need to present an extraordinary obstacle to your arts education research project. Established data collection organizations usually have protocols in place for creating agreements, and your most experienced research partners know how to handle them.

**TIMELY TIP**

Some SEAs post their standard data-sharing agreements online. For example, the Ohio Department of Education has made an annotated example available on its website. You can find other templates for data-sharing agreements on state education agency websites for Arkansas, Colorado (Word document), Kansas, Louisiana, Minnesota, Rhode Island (Word document), South Dakota and Utah.

### DETERMINE HOW YOU WILL RECEIVE THE DATA.

After you execute a data-sharing agreement, you will need a mechanism for receiving the data. If a partner or vendor is conducting analysis for you, it will handle this process as well. Organizations with experience analyzing large government data sets should be comfortable with data-sharing protocols. Still, it pays to have a general sense of these protocols, which could affect the timeline and legal standing of your project.

The data files can be very large, and data security will be critical. There are several means of receiving the data:

- The agency with responsibility for the data transfers a file into your possession. Transfers like this can happen via email, a secure drop box or through a data warehouse or datamart.
- You gain direct access to the data. In this scenario, you would access and work with the data within systems established by the agency that houses the data.

For more information about these options for receiving the data, see Tool K.

## Organize the data.

After you receive the data, you should understand its format, prepare for formatting data challenges and check for errors. Your vendor or partner will probably take on these challenges, but it can help you to have a general sense of what it takes to ensure that the work proceeds smoothly and reliably.

### UNDERSTANDING THE FORMAT OF YOUR DATA

You may receive your data in different formats. The traditional standard for constructing data files, particularly those of moderate size, is in delimited format. Delimited files are files where each row contains a separate record and the fields in each record are each separated by a delimiting character. The delimiting character is most typically a comma or tab, but it could also be parentheses, brackets or other characters. Take the following example:

***Header row*** ➔ Ent_name, ttl_enr, enr_a, enr_b, enr_c, enr_d

***Line 1*** ➔ School A, 254, 32, 75, 56, 27

***Line 2*** ➔ School B, 1217, 456, 342, 176, 97

Lines 1 and 2 of the file contain records for individual schools, and commas separate different values. Above Line 1 is the header row, which contains labels for each of those values. Those labels are usually exported directly from the fields in the warehouse data storage system, so they might be difficult to interpret. The agency providing the data can explain the contents of each column.

When this file is imported into a data analysis software package, the analysis software will recognize this structure and import the data accordingly. Your data analyst or data analysis software will need to understand the structure of the data, so the record layout that defines the nature of the data in each field will be important.

The record layout will also identify which field or fields in the file represent the key, the field or fields that identify each record uniquely. If you have individual student record data, for example, the key would be unique student IDs, which could be social security numbers or other unique identifiers assigned by the state. Data analysis software cannot function without such unique identifiers.

**STUMBLING BLOCK**

A small mistake, like a missing delimiting character or a foreign character, can wreak havoc with your analysis. Data files must be handled with care.

You may receive data from multiple sources, which may complicate the process of formatting and analyzing the data. Tool L offers information on how data analysts confront that challenge.

### ASSESSING THE QUALITY OF YOUR DATA

The major principles for assessing the quality of data are the same whether you are examining a massive file of students' unit records or a much smaller Excel spreadsheet of aggregate data. That said, it is much more difficult to assess data quality in a massive file of student unit data — unless you have the right software and other statistical tools. If a partner or vendor is analyzing data for you, it will probably "clean" your dataset by using automated methods to detect errors in the data, such as missing data points, stray characters or statistical anomalies.

## Analyze the data.

Step-by-step instructions on how to analyze a large and complex data set are beyond the scope of this toolkit. Still, you should have a general understanding of the tools researchers use to conduct their analyses.

### COMMONLY AVAILABLE SPREADSHEET APPLICATIONS

Common spreadsheet applications like Microsoft Excel, Google Sheets and OpenOffice Calc have robust capacities — Excel can accommodate roughly 1,000,000 rows and 16,000 columns. Most also include essential statistical algorithms and charting features. If your data file is small and your calculations remain simple, these tools may serve you well. (See "The Easiest Path," above.)

### MORE SOPHISTICATED DATA ANALYSIS SOFTWARE

Although the calculations may be simple, the scope of your analysis or size of the data files may require more sophisticated software that can handle larger data sets, conduct more complex analysis and offer more options for data output.

There are many tools in the market that can support such analysis, including Stata, SPSS (which once stood for "Statistical Package for the Social Sciences") and SAS (formerly known as "Statistical Analysis System"). These programs are common in education research, but they are far from the only ones that can analyze large and complicated datasets. To use such programs, you need a solid background in statistical analysis.

### SOFTWARE PACKAGES THAT ALLOW YOU TO VISUALIZE DATA

Tableau and Power BI are examples of a recent trend in data analysis software that generates visual representations of data in the form of interactive maps or charts. Such software can make it much easier than ever before to create web-based data dashboards directly from your data set. Note that you may use Stata, SPSS or SAS to analyze your data before turning to Tableau or Power BI to present the data visually.

Even software like Tableau or Power BI, which are more intuitive than most statistical software, require some technical savvy and skill with data visualization. You may not have to hire a statistician or computer programmer to do this work, but you will need someone who is comfortable with technology.

## Summary

Analyzing large education datasets can demand substantial expertise and may require you to find partners or vendors who can do the work for you. No brief toolkit can supply the expertise you need if you don't have it already. Still, you will have an easier time planning your arts education data initiative if you understand what such data analysis typically entails. You should understand the impact of data privacy laws on your project and plan for the time and effort involved in analyzing large data files.

## NATIONAL ENDOWMENT FOR THE ARTS

Established by Congress in 1965, the National Endowment for the Arts is the independent federal agency whose funding and support gives Americans the opportunity to participate in the arts, exercise their imaginations and develop their creative capacities. Through partnerships with state arts agencies, local leaders, other federal agencies and the philanthropic sector, the Arts Endowment supports arts learning, affirms and celebrates America's rich and diverse cultural heritage, and extends its work to promote equal access to the arts in every community across America. Visit arts.gov to learn more.

## EDUCATION COMMISSION OF THE STATES

Education Commission of the States was created by states, for states, in 1965. It conducts comprehensive research, delivers evidence-based reports, provides expert counsel and convenes state leaders on the full spectrum of education policy issues, from early learning through the workforce. It is the only state-focused national organization to bring together governors, legislators, and K-12 and higher education chiefs, as well as other state education leaders. Learn more at ecs.org.