Alabama
State Department
of Education

Data Preparation and Analysis Primer

RESEARCH
& DEVELOPMENT
ALABAMA STATE DEPARTMENT OF EDUCATION

SOUTHEAST COMPREHENSIVE CENTER
Advancing Research, Improving Education
SEDL
An Affiliate of the
American Institutes for Research

MAY 19, 2016

Table of Contents

Introduction

In 2013, the Alabama State Department of Education (ALSDE) submitted a request to the U.S. Department of Education pursuing a flexibility waiver from the Elementary and Secondary Education Act of 1965, a law amended by the No Child Left Behind Act of 2001 (U.S. Department of Education, 2015). The application proposed a state-determined model for accountability, Alabama Plan 2020 (Plan 2020), a strategic plan that envisions, "Every child a graduate–every graduate prepared for college/work/adulthood in the 21st century" (Alabama State Department of Education, 2016, p. 2). Plan 2020 consists of the following objectives (p. 5):

1. "All students perform at or above proficiency and show continuous improvement."
2. "All students succeed."
3. "Every student graduates from high school."
4. "Every student graduates high school prepared."

The Southeast Comprehensive Center (SECC) at SEDL, an affiliate of American Institutes for Research (AIR), partnered with ALSDE in 2014 to assist in evaluating Plan 2020, mainly focusing on learners and the graduation rate. SECC provided professional development (PD) and analytic technical assistance (TA) on evaluation statistical techniques to ALSDE research and development staff and later codeveloped the data primer in the process so that staff would have a record and instructional manual of the TA provided.

To assess whether students would graduate on time and college- and career-ready, ALSDE staff, with support from SECC, learned and utilized multiple regression as a methodological approach to predict the success of Plan 2020. Using this approach, ALSDE was able to use extant or existing school-level data (e.g., rate of free and reduced-price lunch, number of counselors per 100 students, and average daily attendance) to predict the state's graduation rate.

Using the above analysis as examples, the goal of the data primer is to provide an approach to addressing key research and evaluation questions in education. More specifically, the data primer is designed to enhance understanding and use of data to develop regression models for assessing school progress on the Plan 2020 priorities. The data primer is organized into three sections, *Cleaning and Merging Data*, *Analyzing Data*, and *Presenting Data,* each offering step-by-step instructions. In *Cleaning and Merging Data,* the data primer details how to prepare a data file for analysis using Microsoft Excel and IBM SPSS software, version 17.

In *Analyzing Data,* the process for running correlation and regression analyses is detailed, and lastly, in *Presenting Data*, considerations are offered for visually illustrating key findings from the analyses. Note that the contents of the data primer reflect instructions for using statewide data specific to the ALSDE's prediction analysis. They are not intended for use by other states or districts given their level of specificity. This manual may serve as an example of an approach other state and local education agencies (LEAs) could take with regard to using data analytics to explore educational issues.

Cleaning and Merging Data

One of the biggest challenges in data analysis is preparing the data file. Often, this may take upwards of 75 percent of the time needed to complete the analysis, depending on how "**clean**" the data are and how many issues you may have to address. Over time, as you get more familiar with your data, this process becomes much more efficient.

Hint: Don't underestimate how long this might take initially when scoping a project.

Preparing a Data File in Excel for Import into SPSS

The data file used in this analysis contained variables needed to predict graduation rates for high schools in Alabama. The independent variables were percentages of average daily attendance, counselor to teacher ratio, and students on free and reduced lunch (FRL) programs for the 2011 school year. The dependent variable was the 2011 graduation rate.

1. First sort by year, and delete years not wanted for analysis. In this case, we are only keeping 2013. In Excel, to sort the entire file (rather than a single variable or column), you must either select all columns (click on the blank cell where the row and column headers intersect on the upper left of the spreadsheet) or select a single cell.
2. Click Data > Sort.
3. Under Sort by, select the variable on which to sort the cases. Select and change the sort order if necessary.
4. Sort on school code (pushes the system/LEA-level data to bottom). Delete the rows of system/LEA data.
5. Create a unique identifier.
    a) Insert a new column (column A is a good place). Label the column heading to match the SPSS variable/column name (or name that will be used across Excel files, etc.)
    b) Concatenate the system code and school code to create the unique ID.
        i. For example, = concatenate (b2, c2).
    c) Copy concatenated values, and paste special values ONLY to save the Unique ID in a format that will keep leading zeros and match the format of the file into which the data will be merged.
6. Rename any variables or columns to help with identifying variables in the merged file (e.g., Avg. of English scale score to Explore English scale score 2013).
7. Save Excel file with new name (in case you need to go back to the original file). Note: If using SPSS version 17, save the file in Excel 97 to 2003 format (i.e., *.xls, NOT *.xlsx).

8. Convert Excel file to SPSS file.
    a) Open SPSS.
    b) Open data file (other type).
    c) Sort by Unique ID. Data > Sort Cases > Unique ID.
    d) Save as SPSS file.
    e) Close SPSS.

Merging an Excel file into an Existing SPSS Dataset

Once you clean up and prepare your datasets, you are ready to begin merging them. If you discover problems upon merging your datasets, you may have overlooked a particular data issue that still needs to be addressed. When you have successfully completed the steps for data clean-up and preparation, the next step is straightforward, with just a few point-and-clicks of the mouse.

1. Open SPSS file that data (i.e., Dataset A) will be merged into (e.g., Graduation Rate and Indicator Data).
2. Sort file (i.e., Dataset A) on UniqueID.
3. Open SPSS file from which new data (Dataset B) are being added (e.g., Explore 2013 data).
4. Sort file (i.e., Dataset B) on Unique ID.
5. Data > Merge Files > Add Variables.
    a) Select file (i.e., Dataset B) to merge into the existing data (i.e., Dataset A).
    b) Click Continue.
    c) Under New-active dataset, determine which new variables to keep or exclude. Variables can be excluded by selecting the variable and clicking the arrow to move the variable to Excluded Variables.
    d) Click Match cases on key variables in sorted files.
    e) Select Non-active dataset in keyed table.
    f) Under Excluded Variables, click on UniqueID and then click the arrow to move UniqueID to the Key Variables box.
    g) Click OK.
6. Remove any variables that are not desired from data file (e.g., subcategory).
7. Save file. Suggestion: Save with a new file name in case you have to go back to the original file.

Analyzing Data

Now comes the climactic point in the process—the statistical analysis. All the hard work in preparing and merging your data sets has paid off, and you are ready to begin crunching the numbers. You will start by checking your data using descriptive statistics and looking for possible outliers and normality. Then you will run a correlation analysis, consider which variables you will want to include in the model, and run a regression analysis in SPSS.

Data Checking/Exploration

1. Run descriptive stats on new variables.

   a) Select Analyze > Descriptive Stats > Frequencies.

   b) Select Variables.

   c) Click on Statistics. Select Mean, Median, Mode**,** **Standard Deviation**, Min, Max, Skewness, Kurtosis.

   d) Click Continue.

   e) Click Charts. Select Histograms, show normal curve.

   f) Click Continue.

   g) Click OK.

   h) Review the results to look for out of range or other inappropriate data values for each variable, e.g., some data points may be an anomaly or outliers and could be omitted from the analysis, such as the effect a natural disaster has on student attendance.

   i) Review results to determine if data is approximately **normally distributed** (an assumption that is required for regression analysis; if the data will not be used for regression analysis then this step may not be necessary). Figure 1 depicts a negative, positive, and normal distribution. When data are skewed, the result will show either a positive or negative skew.

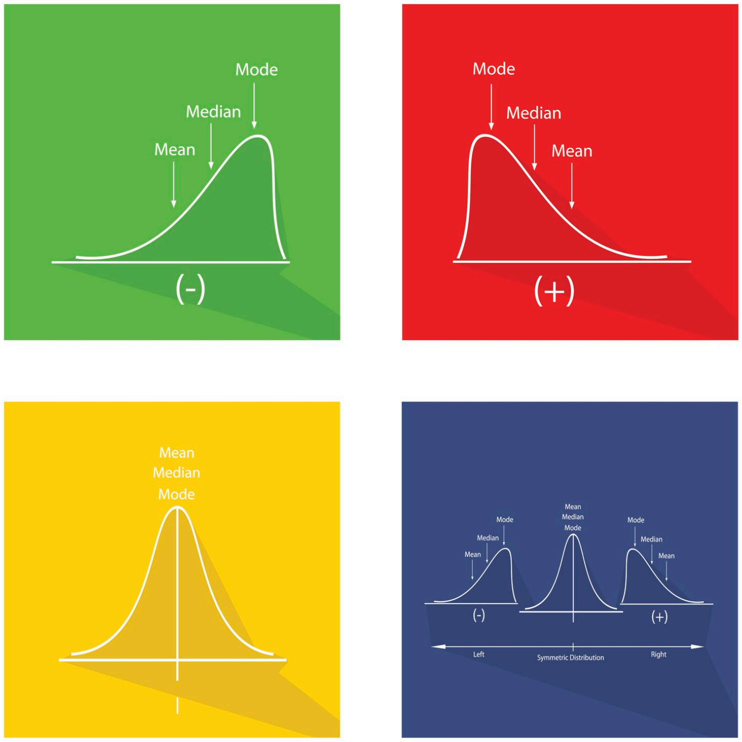


Figure 1. Distributions (source: Thinkstock)

2. Correlations

It may be helpful to examine the **correlations** between variables. Some variables may be so highly correlated that using both would not be necessary or useful. If this occurs, it is known as multi-collinearity.

a) Select Analyze > Correlate > Bivariate.

b) Select the variables to correlate.

Creating a New Variable Based on an Existing Variable

Preparing data for an analysis may first require creating a new variable to get the desired variable needed to run the analysis. For example, in the steps below, we want to include only schools that have a rate of 80 percent or higher FRL, but the data set includes all schools' FRL rate. In SPSS, we need to isolate only cases that have 80 percent or higher FRL rates which would constitute a new variable to include in the analysis. SPSS allows us to do this procedure easily. The steps below illustrate this process.

1. Select Transform > Compute variable.
2. In the Target Variable box, type the name for the new variable (e.g., FRL80).
3. In the Numeric Expression box, input the formula for computing the new variable based on one or more existing variables. For example, to place a 1 in the Target Variable column (FRL80) to indicate a school with a free/reduced lunch of 80 percent or higher, the formula entered in the Numeric Expression box would be: FRL100 >= 80. In this example, schools with an FRL rate of less than 80 percent would have a value of zero in the FRL80 column.
4. Click OK.

Selecting a Subset of Cases for Analysis

We now have the desired new variable to include in the prediction model. To select a subset of cases for analysis (e.g., only schools with an FRL rate of 80 percent or higher), follow these steps.

Select Data > Select Cases.

Under Select, select If condition is satisfied, and then click If…

1. Input the formula for computing, selecting the cases based on one or more existing variables. For example, to select only schools with an FRL rate of 80 percent or higher for analysis, the formula entered would be FRL100 >= 80.
2. Click Continue.
3. Click OK.
4. Any analyses conducted after this point will be based only on the selected cases. In the data view, the cases that are excluded are marked with a slash through row number. Also a temporary variable is displayed in the dataset with a name such as "filter_$."

Regression

Regression is used to determine the strength of relationships among and between variables (known as independent variables) and to predict other variables (the dependent variable). In this case, FRL, average daily attendance, and counselor/teacher ratio are analyzed for correlation and to predict the graduation rate for each school. It allows the researcher to determine which schools are over- or underperforming when comparing their actual outcomes to the predicted outcome. The steps below outline how this analysis is done.

1. Select Analyze > Regression > Linear.
   a) Select a dependent variable. A dependent variable is a measure that is impacted by or responds to an independent variable.
   b) Select independent(s) variable. An independent variable is not changed by the other variables that are measured. An independent variable can be thought of as the opposite of a dependent variable.
   c) Click OK and review the initial analysis.
2. If you decide you want to keep this analysis and determine which schools are over- or underperforming, then do the following:
   a) Select Analyze > Regression > Linear.
   b) Select a dependent variable.
   c) Select independent(s) variable.
   d) Click Save.
   e) Select Predicted Values Unstandardized.
   f) Select Residuals Unstandardized.
   g) Click Continue.
   h) Click OK.

Two new variables will be created in the data file; these will be the last two columns in the data file. Recommendation: Go to the SPSS Variable View and rename the new variables so they can be clearly identified in the future (e.g., PRE_2 = PredictedGradRatefromPLAN).

A positive residual means the school performed lower than predicted, and a negative residual means the school performed higher than predicted.

3. Save the merged file in SPSS. (Recommendation: Save the merged file with a new name in case you need or want to go back to the original file).
4. Save the file also as an Excel file for other purposes (e.g., additional conditional value color coding or for plotting variables.)

Plotting the Residuals Against Variable of Interest in Excel

An excellent way to present your data and one that audiences can typically understand is to visually show your audience what you expected versus the actual outcome by creating a scatterplot in Excel.

1. Select the two columns/variables of interest to be plotted. In Excel, press the control key, and click on the top of the columns you want to select. For example, click the ExploreCompositeScaleScore2013 column and the RES_2 column (i.e., the graduation rate residual column).

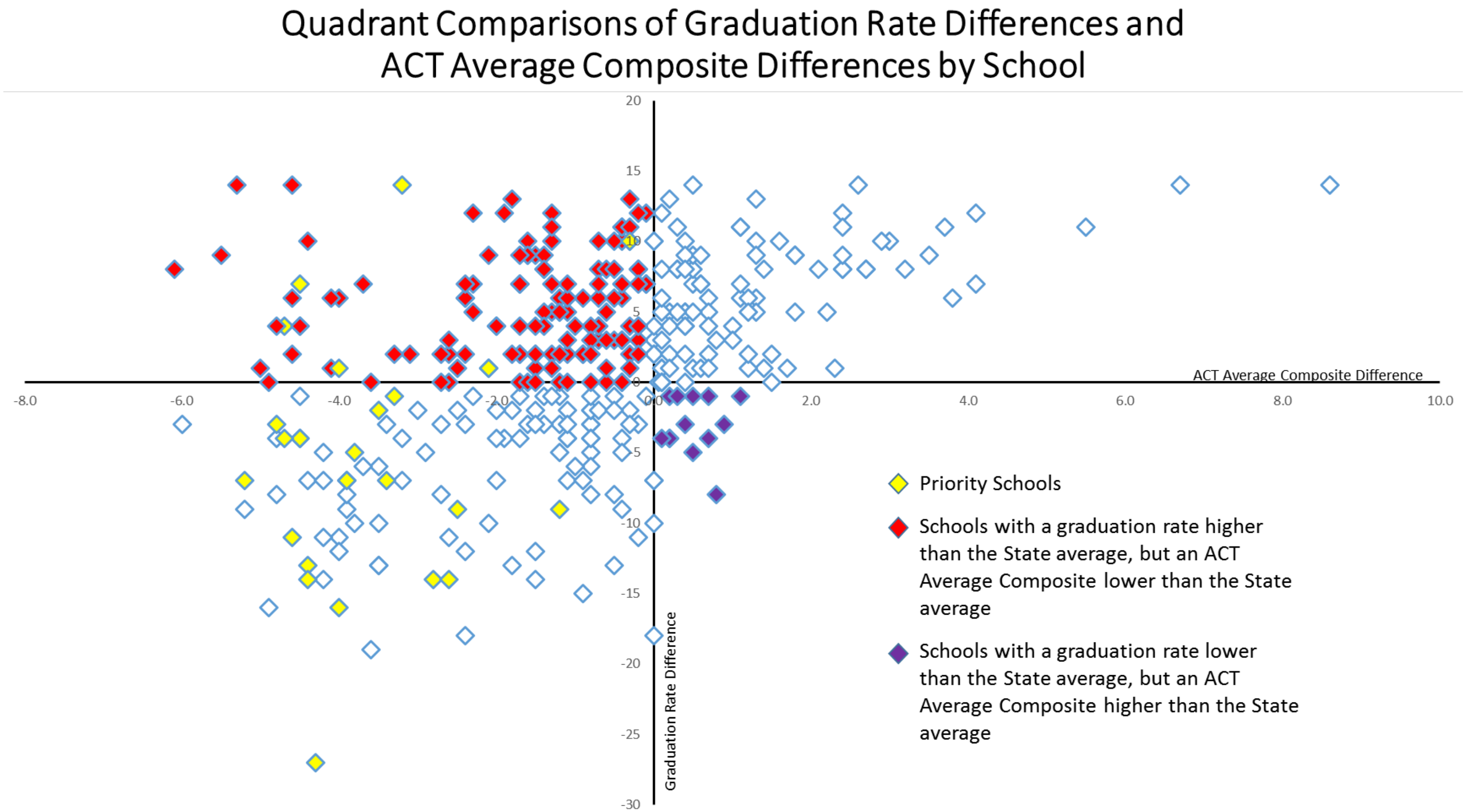
> **Hint:** If you do not want to see all the columns that are not being used or are not of interest, click on those columns by clicking and dragging or by holding the “Ctrl” key and clicking each column, then right click when the cursor is over one of the selected columns and select Hide. To unhide columns, select the entire spreadsheet (click on the blank cell where the row and column headers intersect on the upper left of the spreadsheet) and with the cursor at the top of one of the columns, right click and select Unhide. All the hidden columns will be unhidden.

2. Create the scatterplot.
    a) Select Insert.
    b) Select Scatter > Scatter with only Markers.
        - The scatterplot can be embedded in the spreadsheet or moved to a separate sheet/tab by right clicking on the plot and selecting Move Chart.
        - To get a better view of the data, it may be helpful to adjust the minimum and maximum values on one or both of the axes.
        - Labels can be added to the x-axis, y-axis, and title for the chart, so that the chart can be quickly identified and interpreted in the future.
        - Depending on your audience, however, you might want to try some innovative data visualization techniques to enhance communication of findings and emphasize the points you need to make.

Figure 2, the scatterplot on page 10, illustrates data presented by ALSDE. Using quadrant comparisons, it shows graduation rate differences and ACT average composite differences by school.

Results from analyses can be visualized by using a scatterplot, but depending on the point the presenter would like to make, there are different approaches to take when presenting data.

Figure 2. Quadrant Comparisons



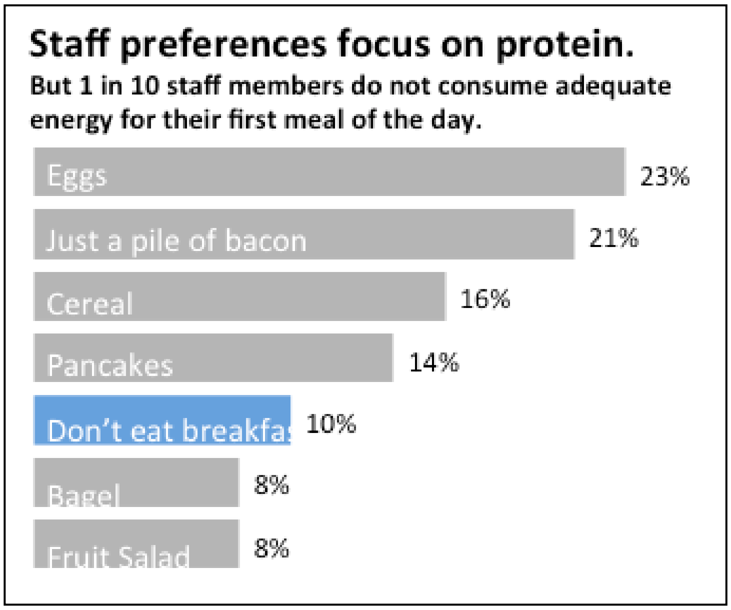
Presenting Results

Data Visualization

After all the work you have done up to this point, it is still essential that you consider how you will present the findings to your audience. The best run analysis will not be great if you do not ensure your audience will understand the findings in their own words. Here are a few suggestions:

- Keep it simple.
- Use visual and numeric elements.
- Make your point succinctly.

In Figure 3, descriptive data is depicted using techniques that follow the three points above.

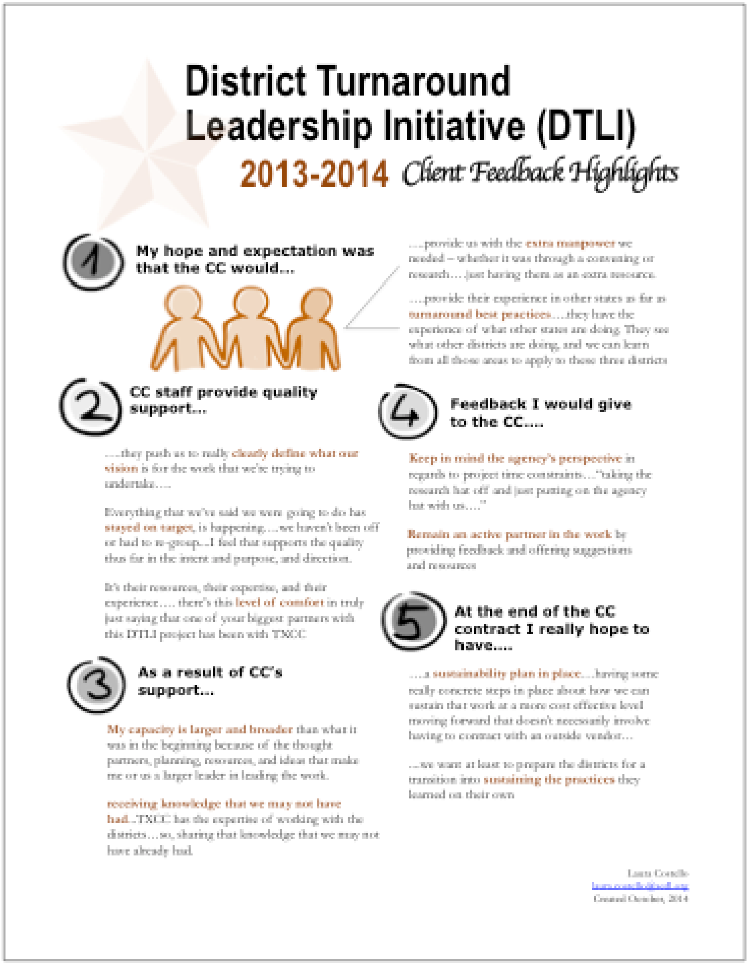


Other Points to Consider

1. Use real pictures to evoke emotions when appropriate.
2. Employ rule of thirds, the golden ratio.
3. Present data in logical order.
4. Use clean lines/font selection (serif/sanserif).
5. Use simple legends/labels as they can detract from the message.
6. State a finding, and include a subtitle as your title and text (see figure on page 6).
7. Be deliberate with colors; use only one color, and mute the rest.
8. Highlight keywords/ways to operationalize qualitative data to make it visual; use graphics to complement.

Data Dashboard

One way to share relevant information from a lengthy report is through a one-page data dashboard. Dashboards provide quick ways to help people reflect on progress in an easy to read document, rather than having to read a long report. Figures 4 and 5 below are examples of how data dashboards take form.



District Turnaround Leadership Initiative (DTLI)

2013-2014 *Client Feedback Highlights*

1 My hope and expectation was that the CC would...

….provide us with the extra manpower we needed – whether it was through a consulting or research….just having them as an extra resource.

….provide their experience in other states as far as turnaround best practices….they have the experience of what other states are doing. They see what other districts are doing, and we can learn from all those areas to apply to these three districts

2 CC staff provide quality support...

….they push us to really clearly define what our vision is for the work that we're trying to undertake….

Everything that we've said we were going to do has stayed on target, is happening….we haven't been off or had to re-group...I feel that supports the quality thus far in the intent and purpose, and direction.

It's their resources, their expertise, and their experience…. there's this level of comfort in truly just saying that one of your biggest partners with this DTLI project has been with TXCC

3 As a result of CC's support...

My capacity is larger and broader than what it was in the beginning because of the thought partners, planning, resources, and ideas that make me or us a larger leader in leading the work.

receiving knowledge that we may not have had...TXCC has the expertise of working with the districts….so, sharing that knowledge that we may not have already had.

4 Feedback I would give to the CC....

Keep in mind the agency's perspective in regards to project time constraints…"taking the research hat off and just putting on the agency hat with us…."

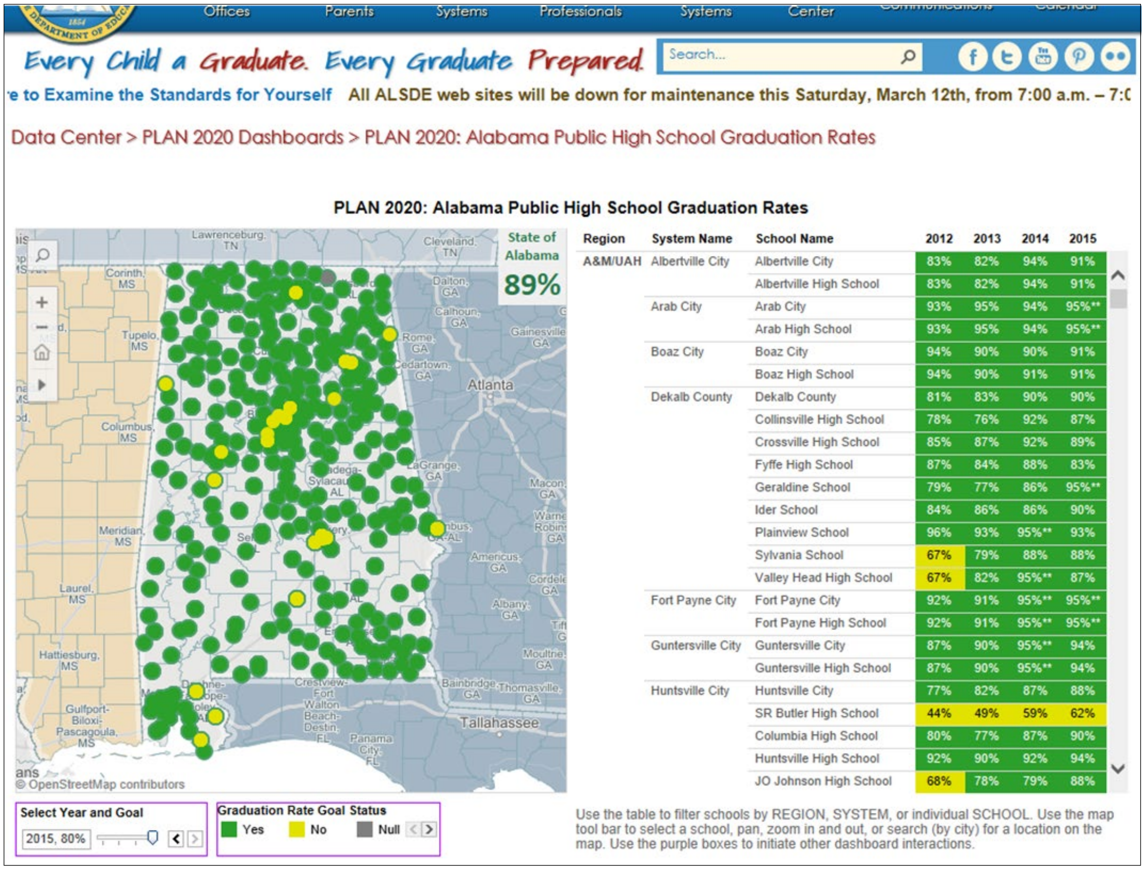
Remain an active partner in the work by providing feedback and offering suggestions and resources

5 At the end of the CC contract I really hope to have....

….a sustainability plan in place….having some really concrete steps in place about how we can sustain that work at a more cost effective level moving forward that doesn't necessarily involve having to contract with an outside vendor…

...we want at least to prepare the districts for a transition into sustaining the practices they learned on their own

Laura Costello
laura.costello@sedl.org
Created October, 2014



e to Examine the Standards for Yourself All ALSDE web sites will be down for maintenance this Saturday, March 12th, from 7:00 a.m. – 7:(

Data Center > PLAN 2020 Dashboards > PLAN 2020: Alabama Public High School Graduation Rates

PLAN 2020: Alabama Public High School Graduation Rates

Region	System Name	School Name	2012	2013	2014	2015
A&M/UAH	Albertville City	Albertville City	83%	82%	94%	91%
		Albertville High School	83%	82%	94%	91%
	Arab City	Arab City	93%	95%	94%	95%**
		Arab High School	93%	95%	94%	95%**
	Boaz City	Boaz City	94%	90%	90%	91%
		Boaz High School	94%	90%	91%	91%
	Dekalb County	Dekalb County	81%	83%	90%	90%
		Collinsville High School	78%	76%	92%	87%
		Crossville High School	85%	87%	92%	89%
		Fyffe High School	87%	84%	88%	83%
		Geraldine School	79%	77%	86%	95%**
		Ider School	84%	86%	86%	90%
		Plainview School	96%	93%	95%**	93%
		Sylvania School	67%	79%	88%	88%
		Valley Head High School	67%	82%	95%**	87%
	Fort Payne City	Fort Payne City	92%	91%	95%**	95%**
		Fort Payne High School	92%	91%	95%**	95%**
	Guntersville City	Guntersville City	87%	90%	95%**	94%
		Guntersville High School	87%	90%	95%**	94%
	Huntsville City	Huntsville City	77%	82%	87%	88%
		SR Butler High School	44%	49%	59%	62%
		Columbia High School	80%	77%	87%	90%
		Huntsville High School	92%	90%	92%	94%
		JO Johnson High School	68%	78%	79%	88%

Use the table to filter schools by REGION, SYSTEM, or individual SCHOOL. Use the map tool bar to select a school, pan, zoom in and out, or search (by city) for a location on the map. Use the purple boxes to initiate other dashboard interactions.

The graphic above is an example of a data dashboard that clearly illustrates the high school graduation rate using two colors for contrast.

Conclusion

To assess college and career readiness of students in Alabama, ALSDE explored indicators such as school attendance, poverty level, and counselor-teacher ratio to predict the graduation rate. In collaboration with SECC, the team used regression analyses to address its research questions. The team assessed availability of data, reviewed literature to determine key indicators, and collected extant data. After the lengthy process of cleaning the data in Excel, they used statistical software SPSS to conduct the analyses and generate output. The team interpreted findings from the output and considered innovative approaches to presenting them using best practices in data visualization. Using data analytics, such as regression, to evaluate Plan 2020, ALSDE has the capacity to examine future research questions that can inform state-level decision making and policymaking.

Glossary

Actual value –The observed value of the dependent variable (e.g., the graduation rate or test score).

Clean data – The term describes data that is ready to be analyzed. In most cases, the data has been merged properly, and missing cases have been examined and addressed.

Correlation – The association between two variables.

Correlation versus causation – A correlation between variables suggests that two events appear to occur at the same time; however, their association or relationship does not imply a causal relationship, which suggests that an action is responsible for another action.

Independent/dependent variables – A condition or trait in data that can influence another variable is referred to as the *independent variable*, while the variable that is affected is known as the *dependent variable*.

Line of best fit – The direction of the relationship as illustrated on scatterplot graph.

Missing data – This may refer to a situation where you have no information on individuals or you may be missing values for some variables.

Normal distribution – Skewness (i.e., symmetry of the distribution) and kurtosis (i.e., the peakedness of the distribution) of variables.

Predicted value –The predicted value is calculated from the estimated regression equation. The value forms the line of best fit.

Regression analysis – An analytical technique used to explore relationships between and among a set of variables.

Residual value –The difference between the observed and predicted value.

Standard deviation – The average distance from the mean.

Scatterplot – A graph that depicts a relationship between variables.

Significance – Also known as the p-value. In social science, anything at or below .05 is considered significant.

Z-scores and standardized coefficients – A z-score is a transformed score that allows for the comparison of different variables on the same scale. Variables standardized for comparing variables.

References

Alabama State Department of Education. (2016, August 24). Plan 2020. Retrieved from http://www.alsde.edu/sec/rd/Plan%202020/Alabama%20PLAN%202020.pdf

U.S. Department of Education. (2015, August 6). *Laws & guidance: Elementary & Secondary Education: Alabama—ESEA Flexibility Request.* Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/map/al.html

Southeast Comprehensive Center

at SEDL, an affiliate of American Institutes for Research

Report Team

Pam Bonsu, M.P.H., AIR Researcher; Heidi Goertzen, Ph.D., AIR Senior Technical Assistance (TA) Consultant; Beth Howard-Brown, Ed.D., Principal TA Consultant and SECC Director; Kris Kaase, Ph.D., Consultant; Jason LaTurner, Ph.D., AIR Principal TA Consultant, and TXCC Director; and Chris Times, M.B.A., SECC/AIR Senior Communications Specialist

State Liaisons

Alabama: Heidi Goertzen, Ph.D. (hgoertzen@air.org)
Georgia: Glenda Copeland, M.A. (gcopeland@air.org)
Mississippi: Debra Meibaum, M.A.T. (dmeibaum@air.org)
North Carolina: Shirley Carraway, Ed.D. (shirleycarraway@suddenlink.net)
South Carolina: Ursula Hill, Ed.D. (uhill@air.org)

ALABAMA STATE DEPARTMENT OF EDUCATION 1854

RESEARCH & DEVELOPMENT
ALABAMA STATE DEPARTMENT OF EDUCATION

SOUTHEAST COMPREHENSIVE CENTER
Advancing Research, Improving Education
SEDL
An Affiliate of the
American Institutes for Research