Tools and Strategies for Engaging Students in Analyzing and Interpreting Complex Data Sources

Joshua Rosenberg, University of Tennessee, Knoxville
Alex Edwards, Tate's School, Knoxville, TN
Bodong Chen, University of Minnesota, Twin Cities

Abstract

Analyzing and interpreting data is essential to the practice of scientists and is also an essential science and engineering practice for science teaching and learning. Although working with data has benefits in terms of student learning, it is also challenging, particularly with respect to aspects of work with data that are not yet very common, such as developing quantitative models, understanding variation in data, and using larger, complex data sources. In this article, we aim to describe tools for engaging students in work with data in your class as well as three general strategies, from understanding how the data were collected to how to include the messier parts of preparing a data set for analysis and then modeling the data in order to answer a driving question. We show how these strategies can be employed using the freely-available, browser-based Common Online Data Analysis Platform, and outline connections to curricular standards.

Background

Data and its creation, and the analysis associated with it, are essential parts of the practice of scientists as well as science teaching and learning (National Research Council, 2012; NGSS Lead States, 2013). In an analysis of the performance expectations from the *Next Generation Science Standards* (NGSS), Kastens (2012) reports that across all grade levels, analyzing and interpreting data appears more than all but three scientific and engineering practices (SEPs).

Providing opportunities for students to work with data "has an exceptionally high pay-off for children's scientific reasoning," but, curiously, analyzing and interpreting data as a scientific and engineering practice remains understudied (Lehrer & Schauble, 2015, p. 696). This practice may present challenges to teachers and students because the NGSS emphasizes data analysis-related capabilities that are not yet very common, such as developing quantitative models (Kastens, 2012). Even for more common capabilities such as making "simple" observations, such as of the height of the school's flagpole, requires negotiation not only of what to measure, but how and how many times to measure and record observations in light of how variable what is being measured is (Lehrer, Kim, & Schauble, 2007). Corresponding to the shift called for in the NGSS from knowing about scientific theories and ideas to figuring out how the world works (Schwarz, Passmore, & Reiser, 2017), it is essential for students to work with data not only as interpreters of the quantitative models that others collect and create, but also to create them themselves (Lehrer, Kim, & Jones, 2011).

In addition to these challenges, analyzing and interpreting newer sources of data, such as the "big" data sources collected and created by scientists and engineers, present additional, and new, opportunities and challenges for science educators (Finzer, Busey, & Kochevar, 2018; Kastens, Krumhansl, & Baker, 2015; Lee & Wilkerson, 2018). Traditionally, the data that

students use as a part of the data modeling approach is data that they collect themselves, whereas, in the context of larger sources of data, data originally collected for some other purpose is often used (Wilkerson & Laina, 2018). For example, Wilkerson and Laina (2018) explore how students repurposed locally-relevant data from the city in which they lived. Other scholars have explored other analytical challenges with large data sources, such as the importance of structuring hierarchical data (Konold et al., 2017) and using technological tools (Finzer, 2012).

Working with data, then, is potentially powerful in terms of student learning and frequently appears in the NGSS. But, analyzing data, particularly as the data available are increasingly larger and more complex, presents opportunities and challenges. In this article, we aim to describe freely-available tools to engage students in work with data in your class. We also articulate three general strategies that you can use to guide your use of these tools in your teaching.

Tools for Working With Data

Like many other educational technologies, in addition to what the tool itself can do, it is also essential to consider how it fits with particular pedagogical aims (and your content area) as well as your particular context. Thus, we selected tools that we think exhibit some of the characteristics of effective data analysis platforms for learners (see McNamara, 2015). Also, the tools for working with data we identify are those that are freely-available and browser-based, and so should be relatively easy-to-use.

Desmos

Desmos is commonly used in addition to (or in replacement of) graphing calculators by mathematics teachers and students, and it also has some uses in the science classroom. Like graphing calculators, Desmos works well with functions that do not require a data table (such as the use of the function $f(x) = \sin(x)$ to display the form of that function). But, it also works well with datasets. Data can be typed directly into Desmos or can be copied from Google Sheets or other spreadsheet software. Then, functions, such as a sin, linear, or quadratic function, can be estimated based on the data (and added to a graph). A distinctive feature of this tool is the ease with which functions (even complex functions) can be written, even by those not accustomed to writing out an equation. Desmos is available via browser without the need to log in[1].

Google Sheets

Google Sheets is widely known to (and used by) science teachers and students especially for school districts using the Google Suite. A benefit of Google Sheets is that it bears similarities to other, widely-used tools, namely, Microsoft Excel! This may make it easy for students to begin to use this tool. Compared to Excel, because Google Sheets is browser-based, it is easy for students to collaborate through a single Google Sheet. While students at the high school level

[1] https://www.desmos.com/calculator

may be familiar with Google Sheets, advanced functionality--such as writing commands to populate cells with values that rely on other cells (i.e., to create the mean of multiple variables)--likely requires additional support. Moreover, while fitting complex functions to data is possible as a part of editing a graph, it will likely require additional support. Finally, while easy to use, sometimes Google Sheets can make it so easy to create a figure that students may not have the opportunity to think carefully about what each part of the figure represents. Google Sheets is accessible via web-browser with a log-in[2].

JASP and R

JASP is statistical software for professionals to do data analysis (as well as students at the undergraduate level and above), but it can also have a role in high school classes. JASP is based on the R software; unlike R, JASP has a point-and-click interface, through which it is possible to carry out a wide array of statistical tests. Thus, JASP may be most useful for teachers of students needing to carry out analyses more complicated than those that are feasible or easy to carry out using other tools. In addition to using JASP, some students may be interested in using R. R is most commonly used via the R Studio software, which executes R and provides some additional functionality for enhanced data-analysis workflows. While challenging to use, in some advanced applications, it may be useful to turn to R. JASP has a desktop version[3] and a browser-based version[4]. There is a browser-based version or R Studio available, R Studio Cloud[5].

The Common Online Data Analysis Platform (CODAP)

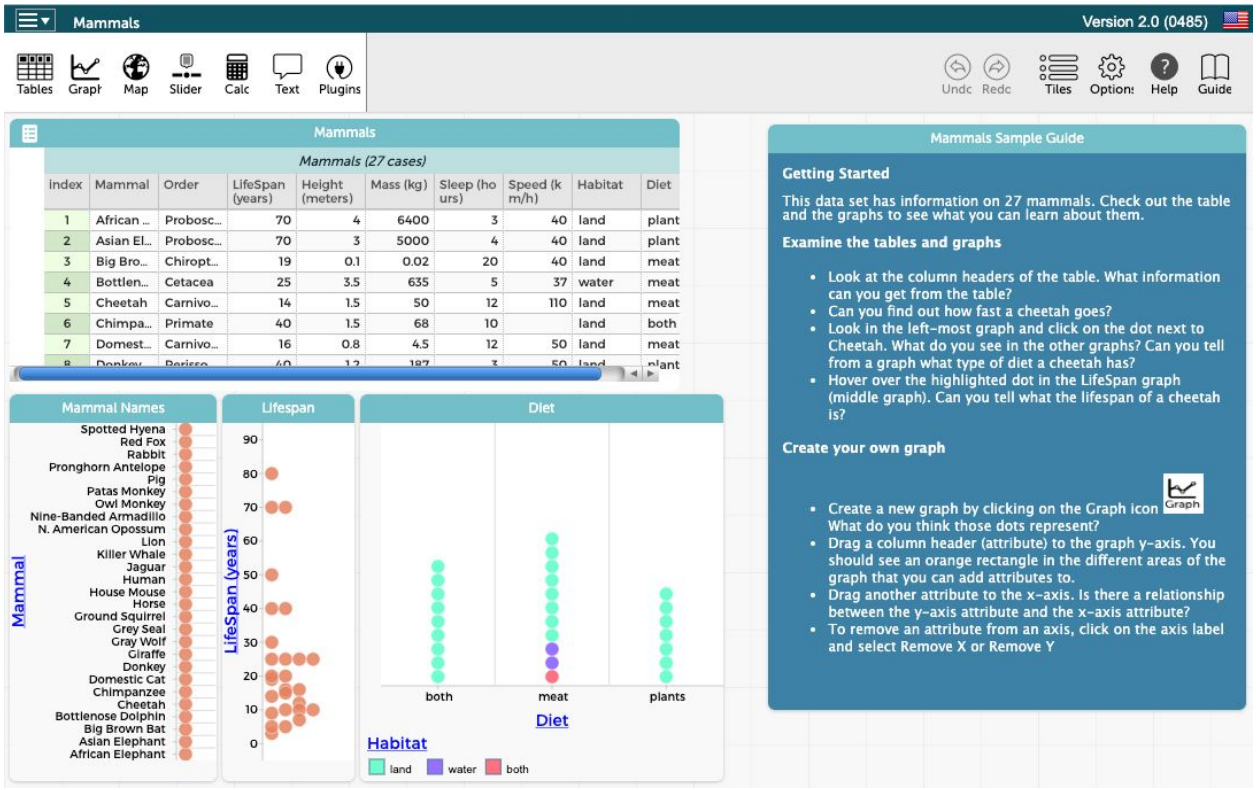
The Common Online Data Analysis Platform (CODAP) provides a distinctive interface to view, transform, and analyze data and create and interpret graphs. Developed by the Concord Consortium, CODAP draws upon past research into and development of the TinkerPlots[6] and Fathom[7] statistical software. One distinctive feature, related to how both data and graphs can be viewed together, is that elements of graphs, such as dots on a scatter plot, can be clicked on to view which data they correspond to. Another distinctive feature of CODAP is its drag-and-drop interface: for example, to create a graph, columns from a data table can be dragged to the *x*- or *y*-axes or the grid of the graph, to color points based on the values in the column. It is also easy to load data (as long as you can save the data as a CSV file, which can be done in Google Sheets or other software). A CSV file can be dragged into the window to load the file as a table. In addition to being easy to use, CODAP has more advanced functionality, as well, such as the ability to fit quantitative models (i.e., linear models or regression models). CODAP is available via browser without the need to log in[8]. Additional resources, including tutorials and example

---

[2] https://docs.google.com/spreadsheets
[3] https://jasp-stats.org/
[4] https://www.rollapp.com/app/jasp
[5] https://rstudio.cloud/
[6] https://www.tinkerplots.com/
[7] https://fathom.concord.org/
[8] https://codap.concord.org/releases/latest/static/dg/en/cert/index.html

data sets and activities, are also available[9]. Because of the affordances of CODAP, we focus on the use of this tool in the next section.

Figure 1. A screenshot of the freely-available, browser-based CODAP software.



Note. The flag in the upper-right corner indicates that this is the English-language version; Spanish, German, Hebrew, Greek, Turkish, and Chinese are also supported.

Three Strategies for Analyzing and Interpreting Complex Data Sources

We have been engaged in research around how tools such as those described above—and in particular CODAP, because of its distinctive features—can be used to support student learning in the context of the NGSS. From this work, we have identified several strategies. We summarize three that align with other past research and that we found can be used as a part of other, longer investigations (such as those that take place across a lesson sequence or a unit) or as a part of a single lesson, or class. Thus, these are strategies that we encourage you to consider as you support students to analyze and interpret data in a variety of contexts[10]. While we focus on how these strategies can be employed using CODAP, each could also be employed using another of the tools (or, through the use of tools other than those described here). Finally, while the strategies can be considered on their own, they may best be considered as a part of a cycle, where, for instance, students first explore how the data were collected, next engage in the

[9] https://codap.concord.org/for-educators/

[10] For specific activities, consider Kastens et al. (2015) and Finzer et al.'s (2018) articles in *The Science Teacher*

messier parts of preparing a data set for analysis, and, then, model the data in order to answer a driving question.

Strategy #1: Explore How The Data Were Collected or Created

Creating or collecting data is an essential step in the data analysis process (Hancock, Kaput, & Goldsmith, 1992). This step can also serve as an introduction to working with data, particularly for students who are familiar with the practice. When students record observations themselves, then, they have the opportunity to consider how the data gets created, and may be more confident when analyzing it later on. When students use already-collected data, or secondary data, there are still benefits to considering how the data came to be. Thus, when students are analyzing already-collected data it is still important - and maybe more important than when students collect the data themselves - for students to have the chance to think about how the data were originally collected or created. These discussions may lead students to question how and why the data were collected and to consider sources of bias (deliberate or unintentional) that change the nature of the data, which can be seen as an example of critical data literacy (Hauteau, Dasgupta, & Hill, 2017).

A specific way to support students to explore how the data were collected to created is to start with data that represents a single *case*. Often, the data that students are analyzing are *aggregates* of individual cases of data (such as when a data set includes a column representing the mean of a measurement collected multiple times). In CODAP, this is supported by the connections between the data points and the figure, as in Figure 2. Another way is to talk through, with students, what the data collection process was like, or what it could have been like, as facilitated through a discussion of a description of a study associated with the data, a codebook describing what the variables are, or a data collection instrument (or a description of one).

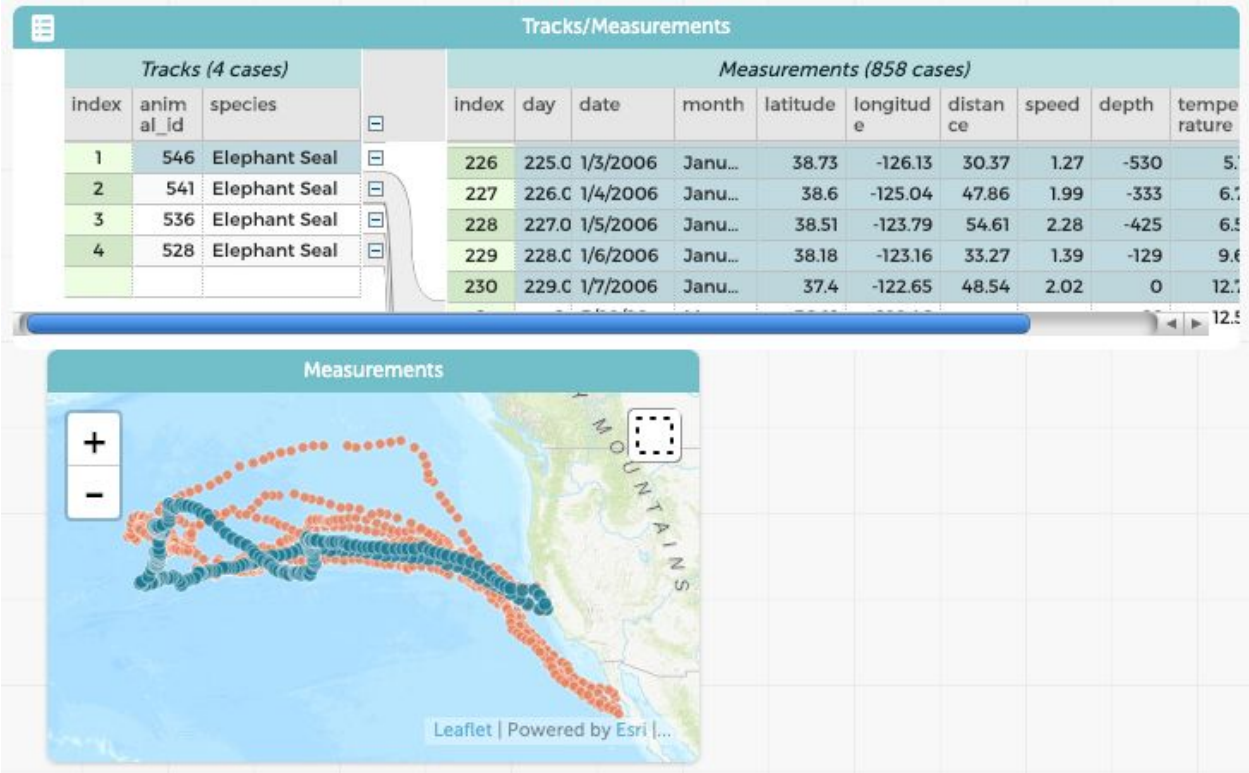
Figure 2. Which mammal eats meat and lives on both land and water? Corresponding data points and their place in figures.

Strategy #2: Involve Students in the Messier Parts of Analyzing Complex Data

Working with clean, tidy data makes it easier for students to reach conclusions, but, particularly with complex sources of data, the need to think about and work through the messier parts of the process--such as renaming and selecting variables and joining together multiple datasets--can also be valuable (Kjelvik & Schultheis, 2019; Konold, Finzer, & Kreetong, 2017; Schultheis & Kjelvik, 2015; Wilkerson-Jerde et al. 2017). In CODAP, it is easy to include data sets that are hierarchically structured or to create nested data structures. In this way, students can see and explore the connections between data at multiple levels. Figure 3 depicts how all of the observations associated with one elephant seal are grouped. Another way to engage students in the messier parts of data analysis is even more simple: allow some time for students time to explore the data and to generate their own intuition for and ideas about the data. This can be especially useful as a part of exposing students to raw, messy data, akin to the kinds that

scientists create and use, but which may also require greater time and effort than is required in typical data analyses (see Data Nuggets[11] for structured activities that involve students in analyzing complex data from scientists).

Figure 3. Analyzing hierarchical data in CODAP.



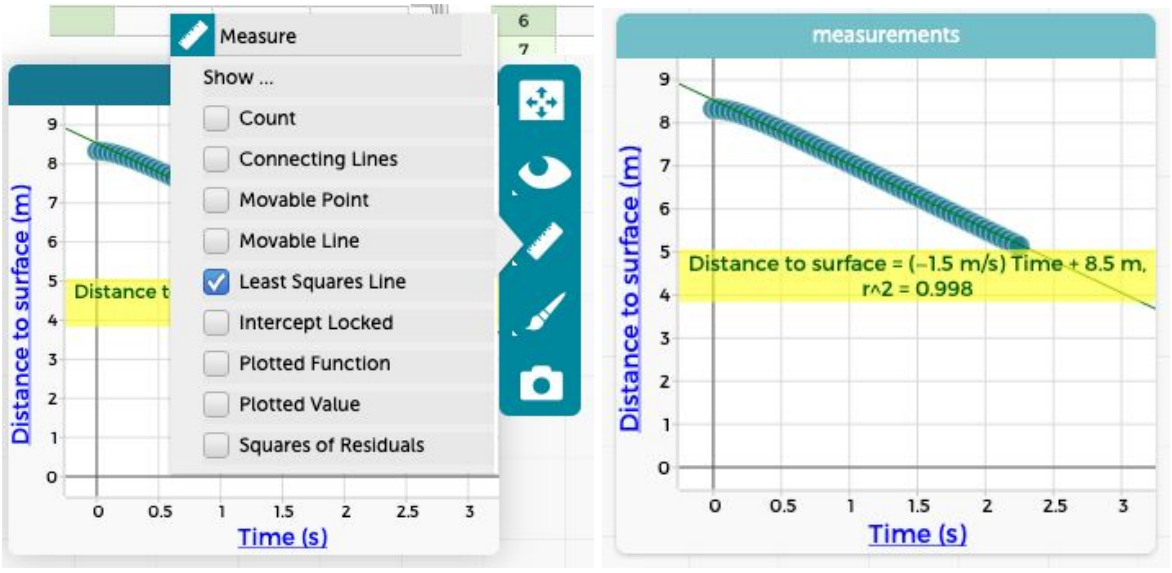
Strategy #3: Push Students to Model and Explain Variability in the Data to Answer a Question and Solve a Problem

Finally, a central goal of statistical models - and statistics - is to understand what is going on in light of variation in the data (Aridor & Ben-Zvi, 2019; Lehrer, Kim, & Schauble, 2007). Importantly, explaining variability does not need to involve highly complex models: even a *mean* or a *median* can be an important summary. A key part of using this strategy is recognizing that it is not essential for students to learn about the mean or the median; it *is* important that students have the opportunity to use statistics that are useful for figuring out what is going on with something concrete or able to be inferred from something concrete in the world: a phenomenon. When using this strategy, it is important to push students to reach and to defend their conclusions in light of variability to answer an authentic question, such as a driving question that allows students the opportunity to answer the question in multiple ways--as well as share and revise their answers--about what they think is happening. In CODAP, modeling and explaining variability is easy to do by clicking on an already-created figure, as demonstrated in Figure 4. In addition to adding a model such as the linear model depicted in Figure 4, students can also easily, for example, add statistics such as the mean and median to a (and to groups depicted

[11] http://datanuggets.org/

within a graph) and represent how spread out a variable is through the calculation of statistics such as the standard deviation or the range and through adding graphical representations of these statistics to a graph.

Figure 4. Adding a regression (linear model) in CODAP.



Note. The left screenshot shows how to add a least squares line. The right screenshot shows the results in terms of both the graphical representation of the model as well as the equation and R^2 term.

Conclusion

In this article, we described some of the freely-available, browser-based tools that have promise for supporting working with data in your classroom. We also articulated three strategies, focused around the use of CODAP: 1) understanding how (already-created) data were generated, 2) involving students in some of the messier parts of working with data, and 3) pushing students to explain variability in data to answer a question or to solve a problem. As you consider these tools and strategies to support students to engage in the practice of analyzing and interpreting data, we encourage you to seek out data that both helps your students to progress toward demonstrating a PE and data that has the potential to spark their interest, such as data about phenomena, locations, or problems that are of interest and are relevant to your students.

Connections to the NGSS

Dimensions	**Name**	**Specific connections to classroom activity**
Science and Engineering Practices	Using mathematics and computational thinking	Learning about and using statistics, such as the mean or the median and the standard

		deviation or the variance.
	Analyzing and interpreting data	Making and recording observations or considering how a measure was constructed and used; renaming and selecting variable, joining multiple data set, and calculating summary statistics; and creating statistical output and communicating findings.
	Developing and using models	Using statistical models and methods, such as a linear model (regression) and a comparison of means using a *t*-test and interpreting the results of their use.
Crosscutting Concepts	Patterns	Understanding how variation and variation between two variables (covariation) may reveal underlying processes or mechanisms that can be used to explain phenomena or to solve problems.
	Scale, proportion, and quantity	Distinguishing between variables at different levels or time scales and understanding how some quantities can be better understand and compared in relation to the whole (as proportions).

Standard and Name	**Specific connections to classroom activity**
CCSS.MATH.PRACTICE.MP2: Reason abstractly and quantitatively.	Using statistical models (such as a linear model, or a regression) to relate two or more variables at an abstract level.

CCSS.MATH.PRACTICE.MP4: Model with mathematics.	Understanding how different functions can be used to explain the patterns in data.
CCSS.MATH.PRACTICE.MP5: Use appropriate tools strategically.	Recognizing that tools for analyzing and interpreting data have specific features as well as drawbacks, and so should be selected based upon the specific question or problem.
CCSS.MATH.PRACTICE.MP6: Attend to precision.	When communicating results from a data analysis, considering the units associated with variables and the appropriate degree of precision needed.

References

Aridor, K., & Ben-Zvi, D. (2019). *Students' aggregate reasoning with covariation.* In Topics and Trends in Current Statistics Education Research (pp. 71-94). Springer.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2).

Finzer, W., Busey, A., & Kochevar, R. (2018). Data-driven inquiry in the PBL classroom. *The Science Teacher, 86*(1), 28-34.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337-364.

Hautea, S., Dasgupta, S., & Hill, B. M. (2017). Youth perspectives on critical data literacies. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 919–930. https://doi.org/10.1145/3025453.3025823

Kastens, K., Krumhansl, R., & Baker, I. (2015). Thinking big. *The Science Teacher, 82*(5), 25.

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: exploring the potential of using data from scientific research to support student data literacy. *CBE—Life Sciences Education, 18*(2), es2 (1-8).

Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal, 16*(2).

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects.* Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, D.C. https://works.bepress.com/victor_lee/43/

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195-216.

Lehrer, R., Kim, M. J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM, 43*(5), 723-736.

McNamara, A. (2015). *Bridging the gap between tools for learning and for doing statistics* [doctoral dissertation]. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt1mm9303x/qt1mm9303x.pdf

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press

Schultheis, E. H., & Kjelvik, M. K. (2015). Data nuggets: Bringing real data into the classroom to unearth students' quantitative & inquiry skills. *The American Biology Teacher, 77*(1), 19-29.

Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM, 50*(7), 1223-1235.

Wilkerson, M. H., Lanouette, K. A., Shareff, R. L., Erickson, T., Bulalacao, N., Heller, J., & Reichsman, F. (2018). *Data transformations: Restructuring data for inquiry in a simulation and data analysis environment.* In J. Kay & R. Luckin (Eds.), Making the Learning Sciences Count, 13th International Conference of the Learning Sciences. London, UK: International Society of the Learning Sciences.