# GETTING MESSY WITH DATA: Tools and Strategies to Help Students Analyze and Interpret Complex Data Sources

Joshua Rosenberg[1], Alex Edwards[2], and Bodong Chen[3]

[1] *University of Tennessee, Knoxville*
[2] *Tate's School Knoxville*
[3] *University of Minnesota*

This is a preprint.
APA Citation: Rosenberg, J., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher, 87*(5), 30–34.

**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 for student learning, it is also challenging, particularly with respect to aspects of work with data that are not yet very common in schools, 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 to work with data in your class as well as three general strategies, including understanding how data is collected, experiencing the transformation of messy data sets in preparation for analysis, and modeling the data to answer a 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.

# 1 Introduction

Analyzing and interpreting data may present challenges to teachers and students because the Next Generation Science Standards emphasizes data analysis–related capabilities that are not often studied in the classroom, such as developing quantitative models (Kastens 2015). Even making ”simple” observations, such as the height of the school's flagpole, requires knowing what, how, and how many times students should measure and record observations in light of variation in the data (Lehrer, Kim, and Schauble 2007). Due to the shift called for in the NGSS from knowing about scientific theories and ideas to figuring out how the world works (Schwarz, Passmore and Reiser 2017), students can now learn about data not only as interpreters of quantitative models, but also as creators of those models themselves (Lehrer, Kim, and Jones 2011).

In addition, analyzing and interpreting newer sources of data, such as the “big” data sets collected and created by scientists and engineers, presents additional opportunities and challenges for science educators (Finzer, Busey, and Kochevar 2018; Kastens, Krumhansl, and Baker 2015; Lee and Wilkerson 2018). Traditionally, as part of the data-modeling approach, students use the data they have collected themselves. In the context of larger sources of data, students often use data originally collected for some other purpose, such as data from the city in which they lived (Wilkerson and Laina 2018). Facing large data sources, students also need to deal with analytical challenges with large data sources, such as the importance of structuring hierarchical data (Konold et al. 2017) and using technological tools (Finzer 2013).

In addition to what the educational tool can do, it is also essential to consider how it aligns with particular pedagogical aims, content area, and context (Mishra and Koehler 2006). Thus, we selected tools that we think exhibit some of the characteristics of effective data analysis platforms for learners (see McNamara 2015). Additionally, the tools for working with data we identify are those that are freely available (and do not require purchasing a license), browser-based (and so can easily be used across computer operating systems), and relatively easy to use, especially for students. This article also describes three strategies for analyzing and interpreting complex data using the Common Online Data Analysis Platform (CODAP).

# 2 Tools for working with data

## 2.1 Desmos

Mathematics and science teachers commonly use Desmos (see "On the web") with, or in replacement of, graphing calculators. 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. 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. Even for those who are not accustomed to writing an equation can easily write complex functions.

## 2.2 Google Sheets

Google Sheets (see "On the web") is widely used by science teachers and students, especially in school districts using 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. Unlike Excel, Google Sheets is browser-based, making it is easy for students to collaborate through a single Google Sheet. While many high school students may be familiar with Google Sheets, its 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) or fitting functions to data, likely requires 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.

## 2.3 JASP and R

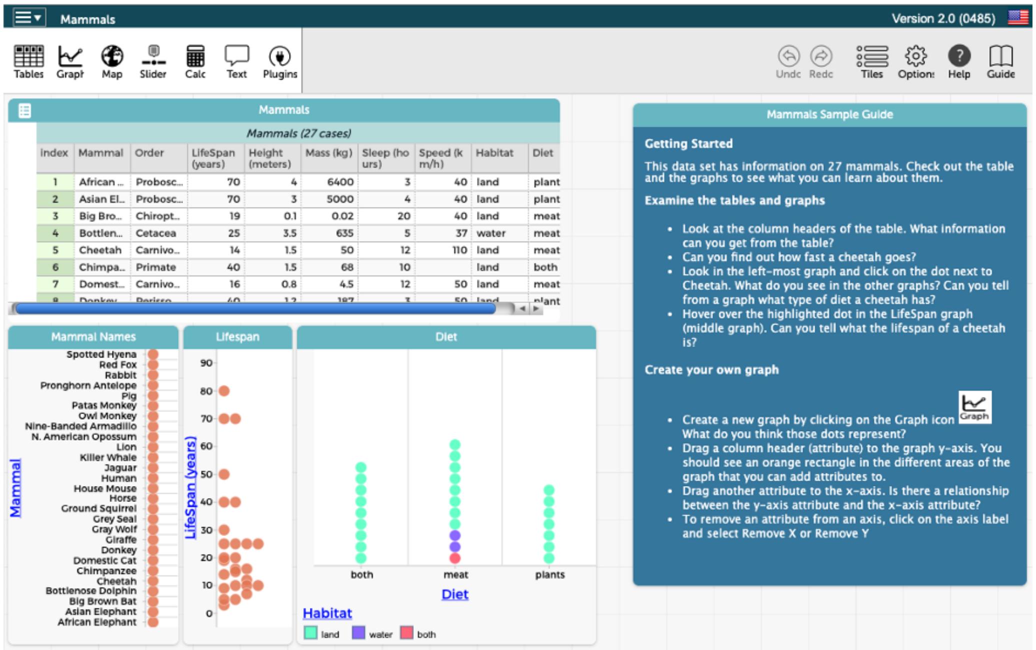
JASP is a statistical software program, based on R (R Core Team, 2019), that students can use for data analysis (see "On the web"). R is a programming language designed for data analysis. Unlike R, JASP has a point-and-click interface, through which it is possible to perform a wide array of statistical tests. JASP may be most useful for teachers who want students to conduct complicated analyses, such as t-tests for how two means (or averages) differ or multiple regression analyses. In addition, R is most commonly used via R Studio software, which executes R and provides additional functionality for enhanced data-analysis workflows. While challenging to use, in some advanced applications—such as for the use of methodologies for analyzing phylogenetic data—it may be useful to turn to R. JASP has both a desktop version and a browser-based version. There is a browser-based version of R Studio available, known as R Studio Cloud.

### 2.4 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 (see "On the web") draws upon past research and development of TinkerPlots and Fathom 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 scatterplot, can be clicked on to view the data to which they correspond (Figure 1). 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, such as the ability to fit quantitative models (i.e., linear models or models for simple linear regression). Additional resources include tutorials, example data sets, and activities.

## 3 CODAP strategies for analyzing and interpreting complex data sources

We have been engaged in research about 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. We have identified several research-based strategies that align with past



*Mammals (27 cases)*

| index | Mammal | Order | LifeSpan (years) | Height (meters) | Mass (kg) | Sleep (hours) | Speed (km/h) | Habitat | Diet |
|---|---|---|---|---|---|---|---|---|---|
| 1 | African ... | Probosc... | 70 | 4 | 6400 | 3 | 40 | land | plant |
| 2 | Asian El... | Probosc... | 70 | 3 | 5000 | 4 | 40 | land | plant |
| 3 | Big Bro... | Chiropt... | 19 | 0.1 | 0.02 | 20 | 40 | land | meat |
| 4 | Bottlen... | Cetacea | 25 | 3.5 | 635 | 5 | 37 | water | meat |
| 5 | Cheetah | Carnivo... | 14 | 1.5 | 50 | 12 | 110 | land | meat |
| 6 | Chimpa... | Primate | 40 | 1.5 | 68 | 10 |  | land | both |
| 7 | Domest... | Carnivo... | 16 | 0.8 | 4.5 | 12 | 50 | land | meat |
| 8 | Donkey | Perisso... | 40 | 1.2 | 187 | 3 | 50 | land | plant |

**Mammals Sample Guide**

**Getting Started**

This data set has information on 27 mammals. Check out the table and the graphs to see what you can learn about them.

**Examine the tables and graphs**

- Look at the column headers of the table. What information can you get from the table?
- Can you find out how fast a cheetah goes?
- Look in the left-most graph and click on the dot next to Cheetah. What do you see in the other graphs? Can you tell from a graph what type of diet a cheetah has?
- Hover over the highlighted dot in the LifeSpan graph (middle graph). Can you tell what the lifespan of a cheetah is?

**Create your own graph**

- Create a new graph by clicking on the Graph icon Graph What do you think those dots represent?
- Drag a column header (attribute) to the graph y-axis. You should see an orange rectangle in the different areas of the graph that you can add attributes to.
- Drag another attribute to the x-axis. Is there a relationship between the y-axis attribute and the x-axis attribute?
- To remove an attribute from an axis, click on the axis label and select Remove X or Remove Y

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

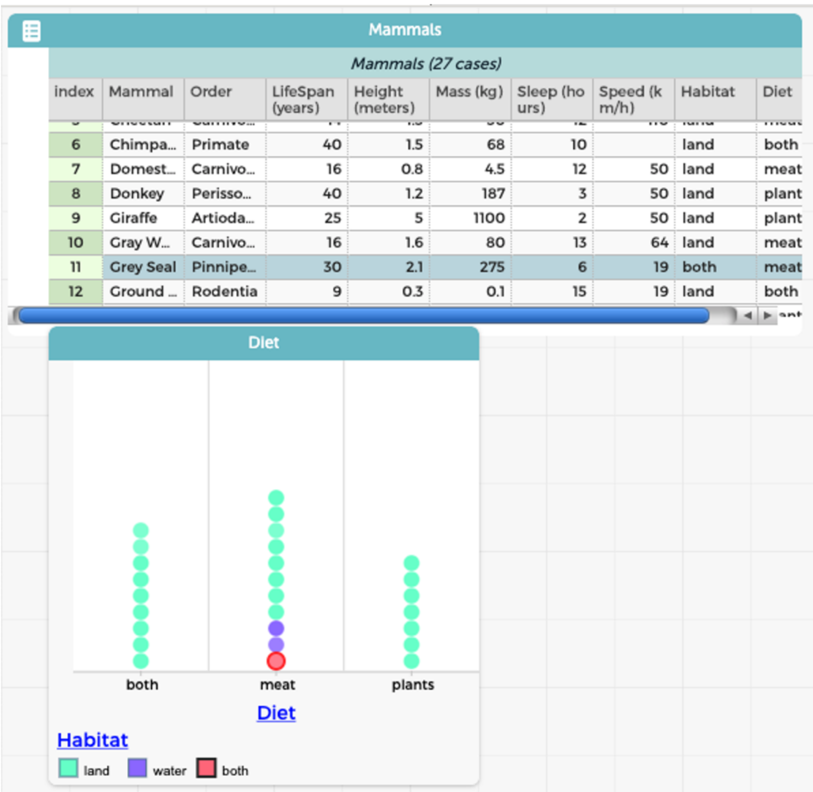
research, and that can be used as a part of longer investigations across a lesson sequence, over a unit, or as a part of a single lesson or class.

While we focus on how these strategies can be employed using CODAP, each could also be employed using other tools described in this article or 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 students first explore how the data were collected, then prepare a data set for analysis, and finally model the data in order to answer a driving question.

## 3.1 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, and 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, 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 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 (Hautea, Dasgupta, and Hill 2017).

To help students explore how the data were collected, 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 (see 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).



Mammals

Mammals (27 cases)

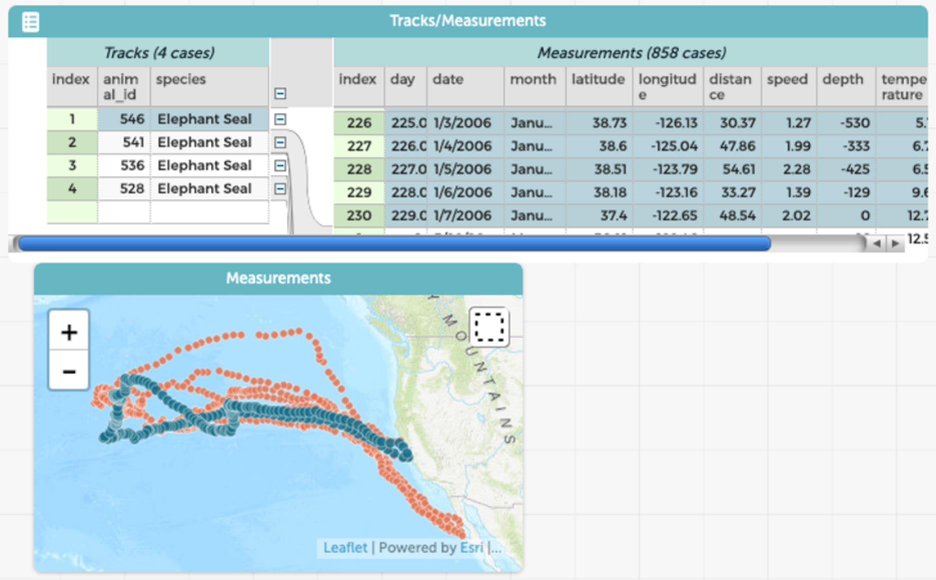
| index | Mammal | Order | LifeSpan (years) | Height (meters) | Mass (kg) | Sleep (hours) | Speed (km/h) | Habitat | Diet |
|---|---|---|---|---|---|---|---|---|---|
| 6 | Chimpa... | Primate | 40 | 1.5 | 68 | 10 | | land | both |
| 7 | Domest... | Carnivo... | 16 | 0.8 | 4.5 | 12 | 50 | land | meat |
| 8 | Donkey | Perisso... | 40 | 1.2 | 187 | 3 | 50 | land | plant |
| 9 | Giraffe | Artioda... | 25 | 5 | 1100 | 2 | 50 | land | plant |
| 10 | Gray W... | Carnivo... | 16 | 1.6 | 80 | 13 | 64 | land | meat |
| 11 | Grey Seal | Pinnipe... | 30 | 2.1 | 275 | 6 | 19 | both | meat |
| 12 | Ground ... | Rodentia | 9 | 0.3 | 0.1 | 15 | 19 | land | both |

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

## 3.2 Strategy 2: Analyze complex data

Working with clean, tidy data makes it easier for students to reach conclusions; however, 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 beneficial (Kjelvik and Schultheis 2019; Konold, Finzer, and Kreetong 2017; Schultheis and 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 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 to explore the data and to generate their own ideas about the data. This can be an especially useful way to expose 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 in "On the web" for structured activities that involve students in analyzing complex data from scientists).



Tracks/Measurements

Tracks (4 cases)

| index | animal_id | species |
|---|---|---|
| 1 | 546 | Elephant Seal |
| 2 | 541 | Elephant Seal |
| 3 | 536 | Elephant Seal |
| 4 | 528 | Elephant Seal |

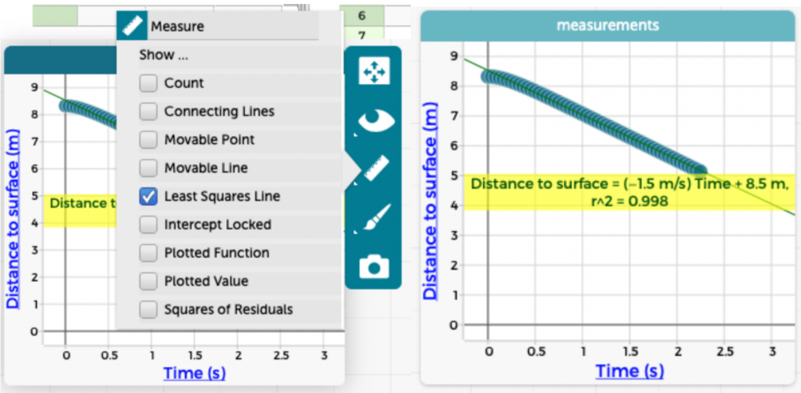
Measurements (858 cases)

| index | day | date | month | latitude | longitude | distance | speed | depth | temperature |
|---|---|---|---|---|---|---|---|---|---|
| 226 | 225.0 | 1/3/2006 | Janu... | 38.73 | -126.13 | 30.37 | 1.27 | -530 | 5. |
| 227 | 226.0 | 1/4/2006 | Janu... | 38.6 | -125.04 | 47.86 | 1.99 | -333 | 6. |
| 228 | 227.0 | 1/5/2006 | Janu... | 38.51 | -123.79 | 54.61 | 2.28 | -425 | 6.5 |
| 229 | 228.0 | 1/6/2006 | Janu... | 38.18 | -123.16 | 33.27 | 1.39 | -129 | 9.6 |
| 230 | 229.0 | 1/7/2006 | Janu... | 37.4 | -122.65 | 48.54 | 2.02 | 0 | 12. |

Figure 3: Analyzing hierarchical data in CODAP.

## 3.3 Strategy 3: 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 and Ben-Zvi 2019; Lehrer, Kim, and 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 determining what is going on with something concrete: a phenomenon.



☐ Count
☐ Connecting Lines
☐ Movable Point
☐ Movable Line
☑ Least Squares Line
☐ Intercept Locked
☐ Plotted Function
☐ Plotted Value
☐ Squares of Residuals

Distance to surface = (−1.5 m/s) Time + 8.5 m,
r^2 = 0.998

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

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. 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 add statistics, such as the mean and median to a graph (and to groups depicted within a graph). Students can also represent how spread out a variable is through the inclusion of statistics, such as the standard deviation or the range and through adding graphical representations of these statistics to a graph.

# 4 Conclusion

Working with data is essential to the practice of scientists as well as to science teaching and learning. As you consider these tools and strategies, we encourage you to think creatively: data do not have to be about something separate from students' investigations of the world. In many cases, data can come directly from students' experiences in your classroom—or their lives. We encourage you to not only seek out complex data sources that help your students demonstrate a performance expectation, or standard, and that also provide chances for students to investigate, critique, and share what they find about topics that are of interest, personal investment, or relevance to them. Doing so can support a shift away from students learning about the world toward figuring out how and why the world works in the ways it does.

# On the Web

CODAP: https://codap.concord.org/releases/latest/static/dg/en/cert/index.html https://codap.concord.org/for-educators/

Data Nuggets: http://datanuggets.org Desmos: www.desmos.com/calculator

Google Sheets: www.google.com/sheets

Fathom: https://fathom.concord.org

JASP: https://jasp-stats.org www.rollapp.com/app/jasp

R Studio: https://rstudio.cloud

Tinkerplots: www.tinkerplots.com

# 5 References

Aridor, K., and D. Ben-Zvi. 2019. Students' aggregate reasoning with covariation. In Topics and Trends in Current Statistics Education Research, 71–94. New York: Springer.

Finzer, W. 2013. The data science education dilemma. Technology Innovations in Statistics Education 7 (2): 1-9.

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

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

Hautea, S., S. Dasgupta, and B.M. Hill. 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, Kim. (2015, May). Data Use in the Next Generation Science Standards (revised edition) [White paper]. Waltham, MA: Oceans of Data Institute, Education Development Center, Inc. Retrieved from http://oceansofdata.edc.org/our-work/data-next-generation-science-standards

Kjelvik, M.K., and E.H. Schultheis. 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): 1–8.

Konold, C., W. Finzer, and K. Kreetong. 2017. Modeling as a core component of structuring data. Statistics Education Research Journal 16 (2): 191–212.

Lee, V. R., and M. Wilkerson. 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, DC.

Lehrer, R., M.J. Kim, and L. Schauble. 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., M.J. Kim, and R.S. Jones. 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

Mishra, P., and M.J. Koehler. 2006. Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record 108 (6):

1017–1054.

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

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

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schultheis, E.H., and M.K. Kjelvik. 2015. Data nuggets: Bringing real data into the classroom to unearth students' quantitative & inquiry skills. The American Biology Teacher 77 (1): 19–26.

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

Wilkerson, M.H., K.A. Lanouette, R.L. Shareff, T. Erickson, N. Bulalacao, J. Heller, and F. Reichsman. 2018. Data moves: Restructuring data for inquiry in a simulation and data analysis environment. In Making the Learning Sciences Count, 13th International Conference of the Learning Sciences, eds. J. Kay and R. Luckin, 1383-1384. London, UK: International Society of the Learning Sciences.