Expansion of the data moves framework to support data processing, analysis and modeling

Stephanie Casey[1], Gemma F. Mojica[2], Hollylynne Lee[2] and Rick Hudson[3]
[1]Eastern Michigan University, US
[2]NC State University, US
[3]University of Southern Indiana, US
scasey1@emich.edu

Erickson et al. (2019) defined a data move as an action that alters a dataset's contents, structure, or values. They also proposed a framework of six data moves from their work with CODAP. We have expanded their framework through reflection on the use of data moves by ourselves and teachers. Our framework has added new data moves, restructured the relationship of data moves, and added purposes for each data move. We will explicate our expanded framework and how it supports data processing, analysis, and modeling through exploration of a dataset involving recent adult obesity levels in Western Pacific and South-East Asian Countries. We will also discuss how teachers' intentional use of the framework can enhance students' understanding of data science.

INTRODUCTION

There are many innovative tools used in educational settings to support students' engagement with large multivariate datasets (c.f., Israel-Fishelson et al., 2023). Although there are multipurpose tools such as spreadsheets and programming environments based on Python and R, other data tools are specifically designed for educational purposes (e.g., CODAP, TinkerPlots, Tuva, INZight). These tools have low/no code requirements, drag-and-drop interfaces, and support linked representations. A common characteristic of all data tools is an interface for users to engage in various actions with data. In 2019, Erickson and colleagues proposed the term 'data move' to describe those actions which alter "a dataset's contents, structure, or values" (p.3) and named six core data moves based on their work with CODAP: Filtering, Grouping, Summarizing, Calculating, Merging/Joining, and Making a Hierarchy. They further called for other researchers to recognize and describe additional data moves as appropriate, and our research group has taken on this task.

METHOD

In a prior research study (Hudson et al., 2025), we began by reflecting on data moves we have used ourselves, as well as those used by teachers we have worked with. We noted *why* those moves were done, what they afforded in data processing, analysis, and modeling, and how different actions resulted in similar re-organizations of data. Next, we analyzed screencasts from 30 teachers' CODAP-based investigations of education data. We used an iterative process of reflection and consideration of screencasts to develop an expanded data moves framework; see Hudson et al. (2025) for further description of the methods. In this paper, we describe our 'Data Moves with a Purpose' framework more generally and illustrate how such moves can support work with data in different data tools. The research in Hudson et al. (2025) led to identifying seven data moves: Ordering, Linking, Inspecting, Grouping, Summarizing, Filtering, and Expanding.

DATA MOVES WITH A PURPOSE FRAMEWORK

Below we describe the seven data moves and explicate how they can purposefully be used to explore a dataset regarding adult obesity levels in Western Pacific and South-East Asian Countries in the years 2000 and 2022. Examples of this exploration will be illustrated with actions in CODAP (The Concord Consortium, https://concord.org), Excel (Microsoft Corporation), and R. The dataset was created using the World Health Organization Plugin in CODAP; read about this plugin and more in Miller (2025).

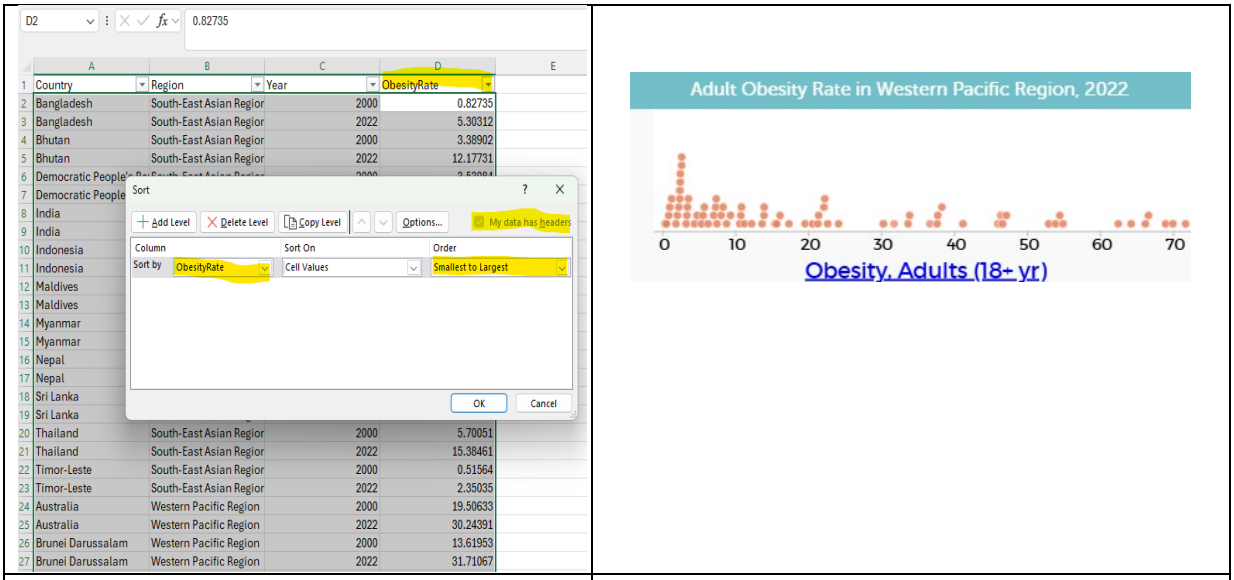
Ordering

Ordering refers to actions that sort numerical or categorical data into a particular order. Examples of ordering actions are:

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*.

- Organizing cases by sorting by an attribute in a table (e.g., using a "sort" action in a spreadsheet to order values in ascending or alphabetical order)
- On a graph axis, reordering categorical attribute names that makes sense for the context (e.g., traffic light data ordered by green, yellow, red)
- Placing a quantitative attribute on an axis to sort cases from low to high
- Creating a scatterplot or time series to order two quantitative attributes on a horizontal and vertical axis and visualize their coordinated location in the plane

In exploring the data from the Western Pacific Region countries in 2022, an example of an ordering data move would be to sort the data based on the obesity values in a spreadsheet (Figure 1a) or to make a dotplot of percent of adults classified as obese in each country (Figure 1b).



a. Using "Sort" in a spreadsheet to order cases based on one column

b. Creating a dotplot in CODAP to order cases from smallest to largest obesity rate

Figure 1. Example of ordering cases in a tabular or graphic view.

The action of ordering a dataset in a table allows for quickly finding the lowest or highest value as well as other information about those cases. Ordering cases from smallest to largest in a graph affords finding extreme values, but also supports reasoning about the range, shape, and clusters in the distribution of adult obesity rates.

Linking

Actions which select case(s) in a representation to identify corresponding case(s) in another representation are considered instances of a linking data move. Linking is often used to coordinate trends across different views of data or to find out more contextual information about a selection of cases. In data tools that support linking, a user typically selects specific cases in one representation (e.g., graph, table) and the corresponding cases are highlighted for easy identification in another representation (e.g., different graph, map). Data tools created for education like CODAP and INZight are designed to support dynamically linking data across representations. Spreadsheets are severely limited in their data analysis capabilities because they do not support linking. In tools like RStudio, using the ggplot or plotly libraries can help in creating interactive plots that allow for linking and inspecting case values within a graph.

For instance, to understand which countries in our dataset have the lowest obesity rates, one could select the countries with the lowest obesity rates in a histogram in CODAP and identify those linked cases in the map and table (Figure 2). An affordance of the linking data move in this instance is seeing in the map that the lowest obesity rates occur in countries in South-East Asia.

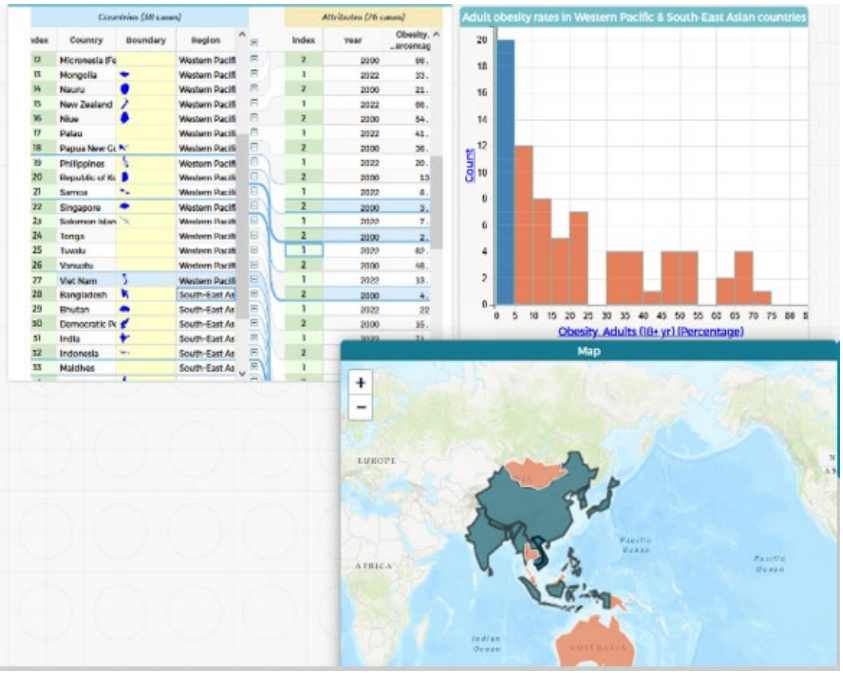


Figure 2. Linking by selecting cases in a graph in CODAP in order to identify them in a table and map.

Inspecting

The inspecting data move involves actions of hovering or clicking on an object to gain additional information about a case, a group, or measure markers shown within a graph. This can be done for geospatial representations by clicking on a map to locate and situate cases based on their location. Other actions that would be considered inspecting include:

- Hovering on an attribute name in any representation to obtain a definition or formula
- Accessing a tooltip in a graph or map by hovering or clicking on a case or geographic region
- Hovering on a measure in a graph or part of a graph to obtain its value

Consider the changes in adult obesity rates from 2000 to 2022 for these countries. Figure 3 presents a graph one could make in CODAP to study these changes. When reading the graph, Samoa stands out as a country whose obesity rate has increased substantially. Hence, determining obesity rates in Samoa in these years is a natural next step. In CODAP, one can inspect a case to find that information directly in the graph by hovering on a case; Figure 3 shows the tooltip for the case of Samoa in 2000.

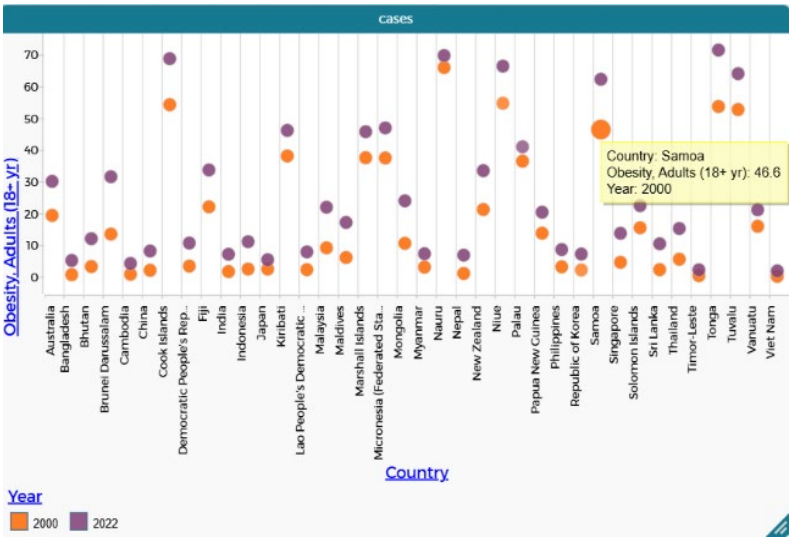


Figure 3. Example of inspecting a case through its tooltip in a graph in CODAP.

In the graph shown in Figure 3, one can immediately read that the adult obesity rate for Samoa in 2000 was 46.6%. Inspecting the case for Samoa in 2022 through a similar process displays its adult obesity rate as 62.4%, making for an increase of 15.8%. By having the cases colored by year, you can also observe that in every country, the purple dot (Year 2022) is higher than the orange one (Year 2000), leading to a strong claim that across all countries, there has been an increase in obesity rates for adults in these 22 years. Coloring the cases by year, treated in Figure 3 as a categorical attribute, is a grouping data move, which is further described next.

Grouping

Grouping involves actions that reorganize the data into subsets to facilitate comparisons between groups. We identified three distinct ways one can make a grouping data move: using existing subsets (e.g., group data by a categorical attribute in the dataset); creating user-defined subsets; or highlighting subsets to consider as group. Actions one can take to make a grouping data move include:

- Separating data in a graph with a categorical attribute on an axis
- Using a categorical attribute as a legend to a graph or a map (as in Figure 3)
- Grouping quantitative data into bins and changing bin size
- Creating a new attribute to group existing cases into categories
- Highlighting a collection of cases

A fruitful use of the grouping data move for our example is using the existing categorical attribute of Region to create a stacked dotplot of adult obesity rates by region for 2022 (Figure 4). To further compare, one can find the percent of countries in each region that had an obesity rate higher than 25%. A moveable line groups each distribution into two subsets, and adding counts and percents shows the number and percentage of cases for each subset of the distribution.

Figure 4. Example of grouping data in CODAP by existing subsets in a graph (left) and separating distributions to find the count and percent of countries with more than 25% obesity rate (right).

This data move makes apparent the stark difference in the distributions of adult obesity rate in the two regions, with the countries in the South-East Asian Region generally having lower and more consistent adult obesity rates (all are below 25%).

Summarizing

Summarizing data moves use computations to describe characteristics of a dataset. Actions that allow one to summarize include:

- Adding visual overlays of statistical measures (e.g., mean, standard deviation) to a graph
- Adding counts or percents to a grouping of cases in a graph
- Showing a boxplot to display the five-number summary visually

Figure 5 shows the addition of visual overlays (a boxplot and the mean) in a spreadsheet and CODAP to provide numerical computations that support analysis of the difference in adult obesity rates across the two regions.

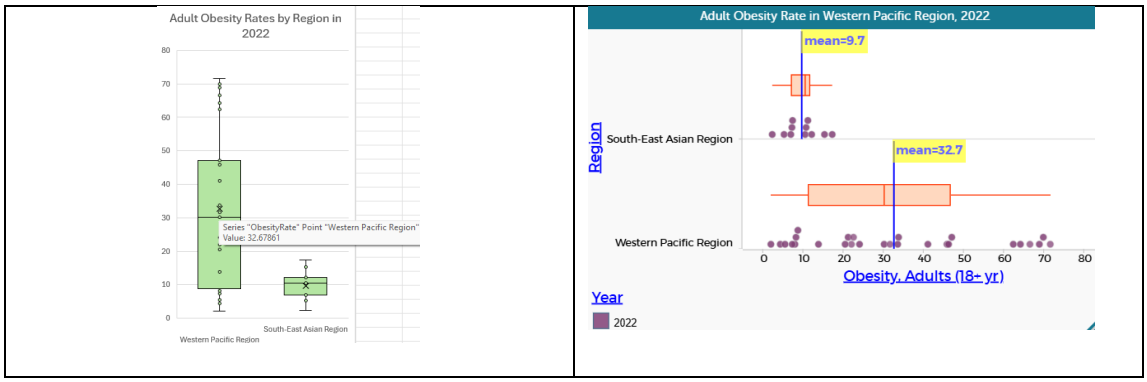


Figure 5. Example of summarizing with boxplots and adding a mean in a spreadsheet and CODAP.

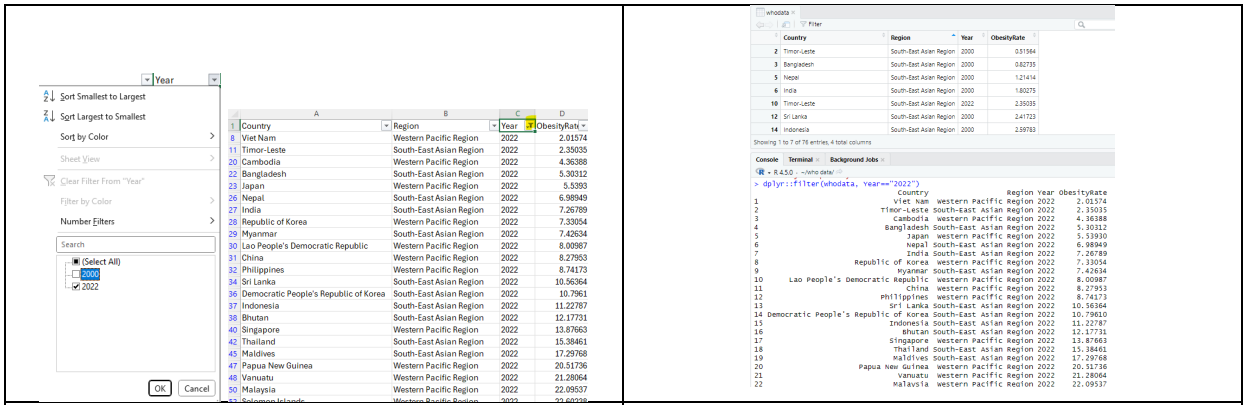
The differing lengths of the boxes (IQR) and their whiskers highlight the difference in the variability of the two distributions, with greater variation shown in the Western Pacific Region. The difference in the placement of the median and mean across the two regions shows that on average, the Western Pacific Region countries have a substantially larger adult obesity rate than countries in the South-East Asian Region. Figure 5a also shows the use of an inspecting move where the user hovered over the mean for the Western Pacific Region in a spreadsheet boxplot to find its value. The values for the mean in CODAP could be seen through inspecting, but the mean values can also be permanently shown on the graph, as shown in Figure 5b.

Filtering

A filtering data move reduces a dataset to only include a subset of its original cases. Actions that one can take to filter the data include:

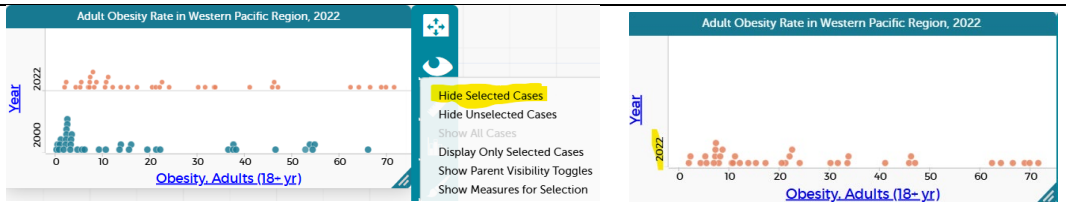
- Removing cases, either individually or as a collection of cases that belong to a certain subset
- Hiding cases in a graph or table (e.g., certain groups, cases, outliers)
- Deleting a case (e.g., an outlier), either from a table or a graph
- Choosing a random sample from a dataset to obtain a smaller dataset

Often, filtering is done to focus on a subset of the data based on certain categories within a categorical variable (like a specific year, or country) but could also be used to examine data within a certain date range or range of values (e.g., countries with obesity rates greater than 50%). For instance, one may want to focus on the more recent data from the year 2022 regarding adult obesity rates in the Western Pacific Region. Filtering is easily done in a spreadsheet (Figure 6a) as well as in R with a filter function such as filter (dataframe, condition) (Figure 6b). To accomplish this using a graph in CODAP, filtering can be done by selecting all cases in the year 2000 and then hiding them (Figure 6c).



a. Using a spreadsheet's "Filter" feature

b. Applying a filter function in *RStudio*



c. Using a "Hide Selected Cases" feature in CODAP to filter data in a graph

Figure 6. Example of filtering the data in a spreadsheet table (a), RStudio (b), and in a CODAP graph (c).

Expanding

Expanding data moves add additional data to the original dataset through three possible methods: adding data values (e.g., adding a missing value manually), merging with another dataset with the same attributes to add multiple new cases, or joining with a different dataset to append new attributes to the original one. Actions one can take to make an expanding data move include:

- Entering data values in a data table, often to clean data or add a missing value
- Adding a new attribute and entering values for that attribute manually
- Using existing attributes to calculate a new attribute (e.g., a rate comparing two existing attributes)
- If using a data portal or sampler plugin, collect or sample new cases

An example of the last action is pertinent to our example: one could expand the dataset by returning to the World Health Organization Plugin in CODAP to add data for additional geographic regions. Figure 7 demonstrates adding countries from the European Region, as shown in the table and map. Alternatively, a new attribute could be added to the dataset by computing whether or not the country has an obesity rate less than 25% or more than 25%. Because this computation is taking a quantitative attribute and using a logic statement to create two groups in a new attribute, this is both an expanding and grouping data move. This is illustrated in CODAP, a spreadsheet and R, in Figure 8. Notice the similar structure of the computation, though the syntax is different across tools.

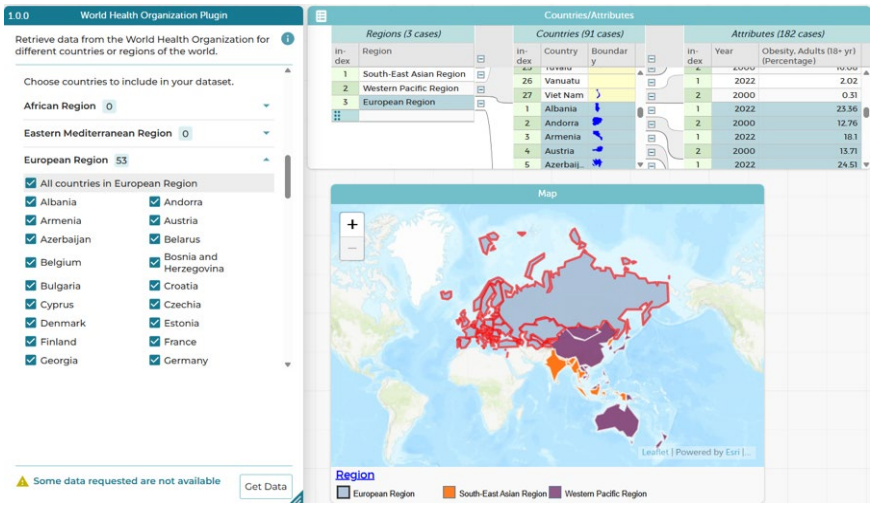
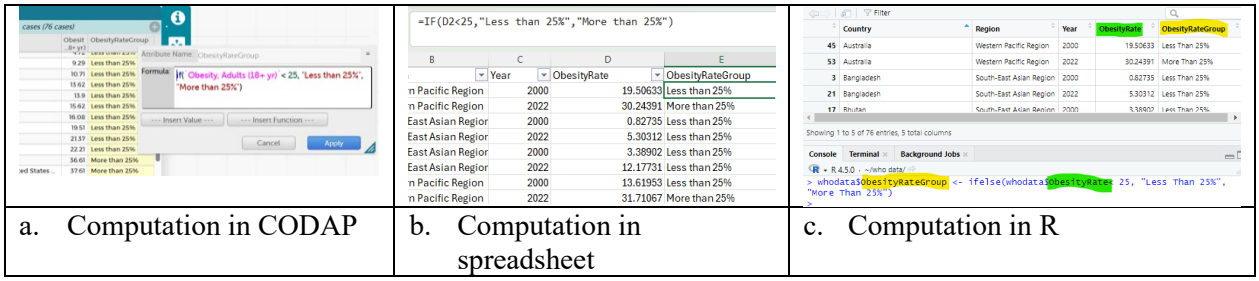


Figure 7. Example of expanding the dataset through use of a plugin in CODAP.



a. Computation in CODAP | b. Computation in spreadsheet | c. Computation in R

Figure 8. Expanding the dataset with a new attribute that creates two groups.

ENHANCING DATA SCIENCE EDUCATION THROUGH USE OF THE FRAMEWORK

By making data moves terminology and purposes an explicit part of data science education, students' learning of data processing, analysis, and modeling can be better supported. Consistent use of data moves terminology across students' learning experiences (different subjects and different years in school) and in working across various platforms (e.g., CODAP, Excel, R) can enhance the teaching and learning of data science in schools. Common language and purposes can assist students to see how particular actions in different data tools are the same data moves. Addition of purposes to the data moves framework supports a deeper understanding of the motivation for using each of the moves, an important aspect for helping students know when to utilize each move. Additionally, the framework can support teachers in scaffolding students' work with data to move them towards more complex analyses. For instance, students who are not disaggregating their data into subsets can be encouraged to use a grouping data move to facilitate making comparisons across subsets. In summary, we advocate for use of the framework in these ways to enhance data science education.

FUNDING

This work was supported by the National Science Foundation (NSF) under grants DUE 1625713, DUE 2141727, DUE 2141716, and DUE 2141724 awarded to NC State University, Eastern Michigan University, and University of Southern Indiana. Any opinions, findings, and conclusions or recommendations expressed herein are those of the principal investigators and do not necessarily reflect the views of the NSF. See http://go.ncsu.edu/esteemhub for more information.

REFERENCES

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12*(1). https://doi.org/10.5070/T5121038001.

Hudson, R. A., Mojica, G. F., Lee, H. S. & Casey, S. (2025). Data moves as a focusing lens for learning to teach with CODAP. *Computers in Schools, 42*(3), 276-301. https://doi.org/10.1080/07380569.2024.2411705.

Israel-Fishelson, R., Moon, P. F., Tabak, R., & Weintrop, D. (2023). Preparing students to meet their data: An evaluation of K-12 data science tools. *Behaviour & Information Technology*, *44*(5), 934-953. https://doi.org/10.1080/0144929X.2023.2295956.

Miller, K. (February 11, 2025). Exploring data in social studies using two new CODAP plugins. *Concord Consortium Blog*. https://concord.org/blog/exploring-data-in-social-studies-using-two-new-codap-plugins