1. INTRODUCTION

Our research team has been developing, with NSF support (IIS-1530578), introductory data science materials for high school students. In comparison to typical introductory statistics tasks, our tasks use large, rich datasets and emphasize exploration, transformation, pattern finding, and model-building as opposed to inference. This approach is consistent with recent developments in statistics education research that emphasize modeling and the social nature of data work, though with a decidedly exploratory focus (e.g., Pfannkuch, Ben-Zvi, & Budgett, 2018). It also reflects an emerging interest among statistics educators to prepare students to understand the increasingly pervasive and computational nature of data (e.g., Ridgway, Nicholson, Borovcnik, & Pfannkuch, 2017).

When novices have access to large, rich datasets, a variety of different questions and interests emerge for them to pursue. Thus no matter how data are originally structured, novices often need to manipulate the data in order to work effectively (this has been called “data wrangling” in the Information Sciences; e.g., Kandel, et al. 2011). For example, a dataset may require filtering because it contains extraneous information unrelated to the students’ goals. It may need to be merged with other datasets. Or, students may wish to use the available data to create new groupings or construct new measures in order to conduct their analysis. Though such actions are common, they are not typically taught as an essential component of data analysis. We call these actions *data moves*.

Though we have found that some novices perform data moves easily, most are unaware that these actions are an available and appropriate way to move their analysis forward. This is not the fault of students: many of these data moves and their purposes are not explicitly taught—indeed, not even made visible or necessary—in typical introductory statistics instruction. Statistics curricula usually give students specific tasks and cleaned, pre-structured datasets that contain exactly the information they need to accomplish only that goal (such as to describe the relationship between two named variables). Or, they might tell students to perform certain data moves in advance of analysis without justifying the need for those moves, or considering their implications.

While such activities have value, they do not give learners the chance to, for example, consider for themselves whether particular cases should be excluded from a given dataset. This in turn means they do not have opportunities to consider *why* students may decide to exclude those cases, *when* such exclusion might be appropriate, or *how* to do so efficiently and systematically. Nor do such activities allow students to consider the impact of these moves, or what analyses may be possible with the same dataset if

they perform other manipulations such as recoding data, calculating auxiliary variables, and so forth. Without such reflection, students may see datasets as objective and static, rather than as socially constructed and transformable.

We argue that data moves should be a central component of what it means to engage in data analysis. In this paper, we first attempt to define what constitutes a data move, and then describe the moves we have identified thus far in our own research. We exemplify how data moves emerge during goal-driven analysis through an example using public data from the National Health and Nutrition Examination Survey ("NHANES"; CDC 2003). Finally, we consider how attending to data moves can inform the planning of curriculum and instruction not just in high school statistics or data science education, but in any situation where learners explore rich datasets.

Digital technologies are important—in many cases even necessary—to conduct the types of analyses we describe. However, we conceptualize data moves as independent of any specific analysis tool. Instead, data moves are a set of actions made possible by a broad class of emerging digital tools designed to facilitate the manipulation and analysis of large, complex datasets (such as R, Tableau, or the Python Data Analysis Library). In this paper, our data-analysis technology of choice is the Common Online Data Analysis Platform ("CODAP"; Finzer 2014), which we use frequently in our work with middle and secondary students. CODAP is a freely available web-based visual tool designed especially for novice analysts; its features parallel many functions of tools designed for professional data science practitioners.

2. DEFINING DATA MOVES

We use the word *case* to refer to a single observation in a dataset. A case is typically represented as a single row in a data table. For example, in the NHANES dataset we use in this paper, each case is a single subject. We use the word *attribute* to refer to a specific piece of information one may collect about a case. An attribute is typically represented as a column in a data table; attributes may be known to readers as variables or parameters. Attributes in the NHANES dataset include weight, height, marital status, and other information about each subject. Finally, *value* refers to the information recorded for a specific attribute for a specific case. A value is typically represented as the contents of a particular cell in a data table. The given marital status or observed height of a specific subject in the NHANES dataset is a value.

We define a data move as *an action that alters a dataset's contents, structure, or values.* Altering a dataset's *contents* means changing the cases or attributes already present in the dataset: adding or removing rows or columns. Altering a dataset's *structure* means changing the way that cases, attributes, and values are related to one another. Altering a dataset's *values* is simply changing the values in the cells. Some data moves, such as *merging*, may alter both contents and structure. Other data moves—which we discuss in detail below—include *filtering*, *grouping*, *summarizing*, and *calculating* new attributes.

Understanding how data moves can alter a dataset, and the circumstances under which such an alteration might be useful, is not trivial. Consider *filtering,* a data move that removes cases from the dataset we are working with. As part of an initial investigation, an analyst might want to look at a subset of the data, say only the health records for 12-year-olds. This filtering action might be only temporary, in order to see how the data are structured. It may be a way for them to "get a feel" for the data by focusing on a population they know more about, so they can check the dataset's face validity or envision how to make calculations using the available attributes. Or, the analyst may filter the data more purposefully because they are only interested in the health status of preadolescent children.

Unfortunately, filtering *per se* is often backgrounded. If it is taught at all, it is described simply as a required step for a larger task or mentioned as a side note. These brief mentions also obscure or omit discussion of the circumstances under which filtering may be a useful step in analysis. The instructor might assume that learners recognize the *need* for filtering, and simply mention it as part of a list of relevant commands or preparatory sequence. However, just as with other practices such as measurement or representation (e.g. Lehrer, Kim, & Schauble, 2007), taking data moves for granted downplays important interpretive and purpose-driven aspects of statistical work. Even students who understand filtering implicitly could

benefit from putting a name to it, recognizing the many ways filtering can be used in the data analysis process, and seeing various ways to implement filtering with technological tools.

Wild and Pfannkuch's (1999) discussion of transnumeration, a "dynamic process of changing representations to engender understanding" (p. 227), foreshadows these ideas in the study of statistical thinking. Our conceptualization differs from theirs in that data moves may be performed not only to engender understanding but also to prepare or simplify a dataset, remove extraneous cases, reorganize it, or even complexify it in a way that reflects their goals. More recently, in the "tidyverse" (inhabited by users of R), Grolemund and Wickham (2017) have developed a comprehensive and well-documented set of tools to do what they call data wrangling; their "verbs"—which are functions in R—often correspond quite closely to our data moves.

2.1. Data Moves and Visualizations

Many researchers have noted the implicit relationship between data moves and visualization (e.g., Chick, 2004; Chick, Pfannkuch, & Watson, 2005; Lee, et al. 2014). Exploring Wild and Pfannkuch's (1999) notion of transnumeration, Chick and colleagues emphasized the transnumeration process as specifically changing how data are represented in order to gain new insights—whether in graphs, tables, or other forms. Lee et al. (2014) further highlight this connection by exploring the production of graphs as one form of transnumerative activity often practiced by pre-service teachers.

Despite this interplay between visualization and data moves, the two are qualitatively different. For example, in a course about data analysis, learning about grouping, summarizing, and hierarchical structures seems to be a completely different animal from the myriad considerations surrounding data graphics. Grolemund and Wickham (2017) make this same distinction between transformation and visualization in their framework. This decision is also contextual—in other work, when studying how people make sense of data, we have included "making a graph" as a data move to simplify terminology (e.g., Wilkerson, et al. 2018).

This same argument applies to other data products as well. For example: we might use data moves to prepare for models and to analyze the results of simulations. For similar reasons we also exclude "data cleaning" steps that are vital to real-world data analysis.

Even without visualizations, models, simulations, and cleaning, however, there is plenty of material left, as we shall see. In this paper, we will focus only on the data moves themselves, the actions that alter the data, as well as their purposes and potential.

2.2. A Possibly Useful Metaphor: Decks of Cards, with Labels

To better understand the class of actions that we consider data moves, we offer a brief metaphor. Let us return to the example above, using a dataset from NHANES which includes health-related attributes for a large representative sample of people from the United States. Imagine that Lynn has a deck of 800 cards, and each card represents one person in the dataset. Suppose Lynn wants to explore whether there are differences between the heights of 12-year-old children with different gender identities[1]. Lynn might…

- look through the deck to pick out all the 12-year-olds,
- separate the 12-year-olds into stacks by gender, and finally,
- calculate an average of the heights on the cards in each stack.

Lynn's first step, picking out the cards for the 12-year-olds, is an example of *filtering*. The second, separating by gender, is *grouping*. The final step, calculating an average, we call *summarizing*. To extend the metaphor of stacks of cards, during the grouping step Lynn may choose to take a blank card to act as a *label* for each stack, giving each group an explicit name. And, since Lynn is interested in the mean heights of each group, they may put something like "meanHT: 157.4." on the label card.

We previously defined data moves as altering a dataset's *contents, structure*, or *values*. Now think about the deck of cards, Lynn's data. Conceptually, Lynn first altered the card deck's *contents* by restricting which cards to look at. They changed the card deck's *structure* by separating it into groups and adding the labels. And they computed new data *values*—the mean heights for each group.

During the process of performing each of these actions, Lynn also may be prompted to critically reflect on the nature of the dataset itself. For example, while *grouping* Lynn may reflect that Male and Female are the only reported genders in the dataset. This may lead them to wonder how "Gender" is determined; is it really the person's gender identity—or something else? The construction of the dataset determines how gender can be used in analysis, with potentially serious implications depending on Lynn's driving questions. Similarly, while *summarizing*, Lynn needs to decide which measure of center (mean, median, mode, etc.) would be most appropriate for describing the heights of each group. Depending on the attribute, this may require further inspection of the data and reflection on the purpose of the summary.

Imagine Lynn wants to present their results about 12-year-olds (for example, as a dot plot with the means marked; See Figure 1). All the information they

[1] In NHANES, the value of Gender for a 12-year-old is initially set in an interview with the parent as either Male or Female. Other datasets available from NHANES include more detailed information about sexual orientation (but not gender identity) for subjects 18 or older; the National Health Interview Survey (NHIS) asks about identity.

need is written on the cards, either on the labels (which become mean markers), or on the original cards (each of which is represented by a dot on the plot). The *x*-axis describes the groups of the cards while the *y*-axis describes values on each card.

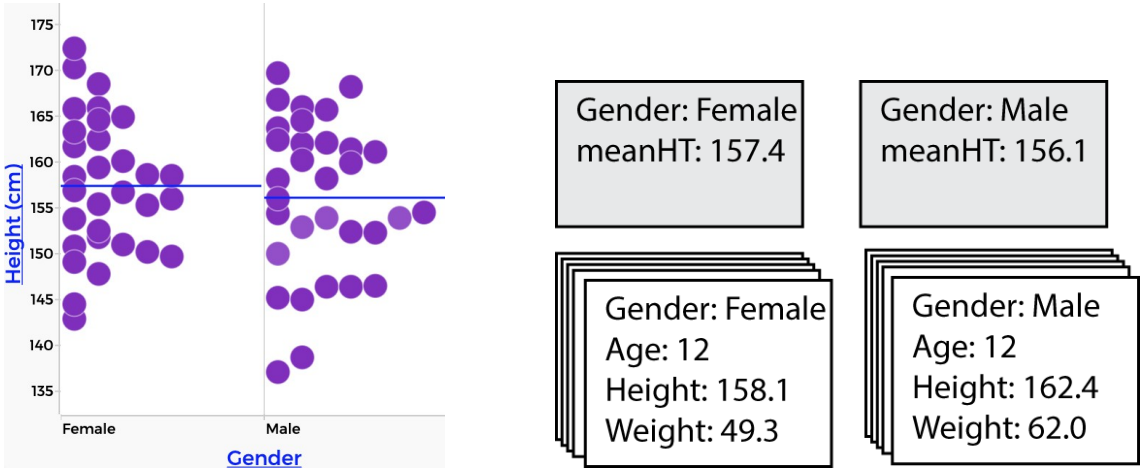


Figure 1. Left: Heights of 12-year-olds, with means by gender. Right: what the cards might look like. The shaded labels have each group name, and also the mean height for that group.

Now suppose Lynn wants to explore, more generally, the growth patterns of young people. What might they do? There are many approaches, but one possible result of such an investigation might be a graph like the one on the left side of Figure 2, which shows how average height changes with age and gender.

Lynn would get the data they need to make that graph in two steps:

- Separate the 800 cards into piles—by Gender and then by Age—and make labels. (*Grouping)*
- For each pile, look though the cards and compute the mean of Height. Write it down on the label. (*Summarizing*)

The resulting piles of cards might look like the ones in Figure 2 (additional, older-kid piles extend to the right). In this example, it’s clearer how grouping has imposed structure on the dataset—and how we keep track of the structure with the labels.

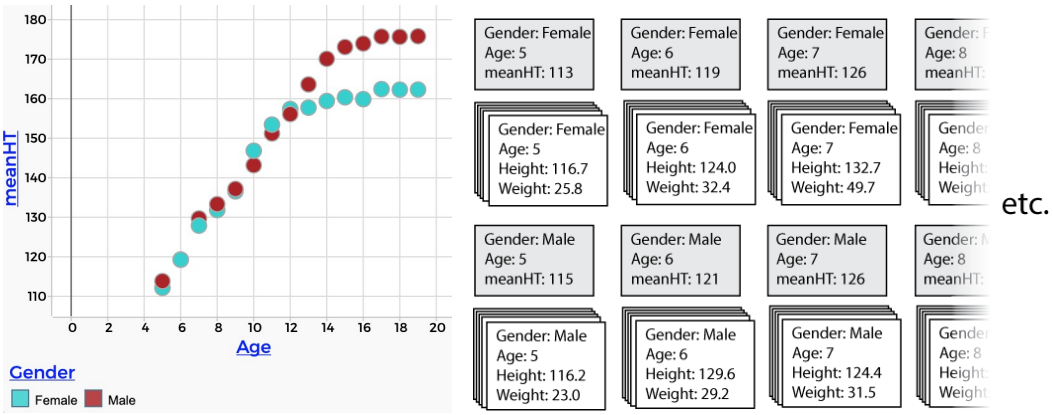


Figure 2. Left: a graph of mean height by age, separated by gender. Right: Piles of cards by gender and age, with labels for each pile. Once again, each label shows the identity of the group plus the summary value.

Again, Lynn's data moves—*grouping* and *summarizing*—have prepared all of the values they need in order to make the new graph. This time, the points on the graph represent groups rather than individuals. Both of the graph's axes (age and mean height) appear on the labels rather than on the original cards.

The alert reader may have noticed that there is really no underlying difference between the form of a label and the form of a card. This is no accident. We will return to this when we discuss the relationship between grouping and hierarchical data structures.

We hope that imagining the decks of cards helps clarify what we mean by data moves. At the very least, it emphasizes data moves as three basic types of action: adding or eliminating cards (contents); moving cards around (structure); and writing things on them (values). The cards metaphor also clarifies what a data move can and cannot do; it highlights how certain data moves satisfy specific analytic needs and even helps illuminate how datasets are constructed. Although every metaphor breaks down somewhere, thinking about the cards emphasizes a focus on *data* moves as opposed to other important actions.

3. THE CORE DATA MOVES

Now, we will explore six specific data moves in further detail, namely, *filtering*, *grouping*, *summarizing*, *calculating*, *merging/joining*, and *making hierarchy*. While not an exhaustive list, these seem useful to examine as a core set of data moves as we have consistently observed these six moves in our own work. They are also included in many different data analysis tools.

Although we will use figures from CODAP to clarify the narrative, our focus here is on the moves themselves and the ways they affect data, rather than on how to perform moves using CODAP or any particular tool. Making this

clean separation is not always possible, as becomes evident when we talk about data hierarchy.

3.1. Filtering

As we have written earlier, *filtering* produces a subset of data, as when Lynn selected only the 12-year-olds in the NHANES dataset above. Although filtering is conceptually simple, it serves at least two important purposes.

- If a dataset includes extraneous cases, filtering removes the irrelevant ones. This is sometimes called *scoping*—reducing the scope of the investigation—or *focusing*.
- Filtering may be used in order to reduce the complexity or quantity of data in order to gain insight. Sometimes this is called *slicing*.

Here is an example of filtering a dataset to gain insight using data from a forest research station in California (Thomson, 2018). Each case in this time-series dataset reports many attributes including air temperature and the temperature of the soil 5 cm deep measured at a given time. For the entire year 2000, these attributes were measured and reported every 30 minutes—over 17,500 cases total.

Plotting air temperature and soil temperature as x and y generates the leftmost graph in Figure 3. There is clearly an association, which makes sense: the warmer the air, the warmer the soil. Imagine that time is plotted on the z axis, with more recent measurements stacked on top of one another as if they are coming out of the page. A “slice” of these data parallel to the page would represent a shorter interval of time. For example, the data points highlighted in the center group in Figure 3 represent the 48 cases from a single day: May 19, 2000.

Filtering the dataset to isolate these 48 cases generates the rightmost graph. This graph reveals an underlying diurnal structure in the data that was completely obscured when the entire dataset was plotted. The slice reveals structure and facilitates a more nuanced interpretation of the data. In this case, the loop likely signifies a time lag between quickly warming and cooling air and more slowly warming and cooling soil.

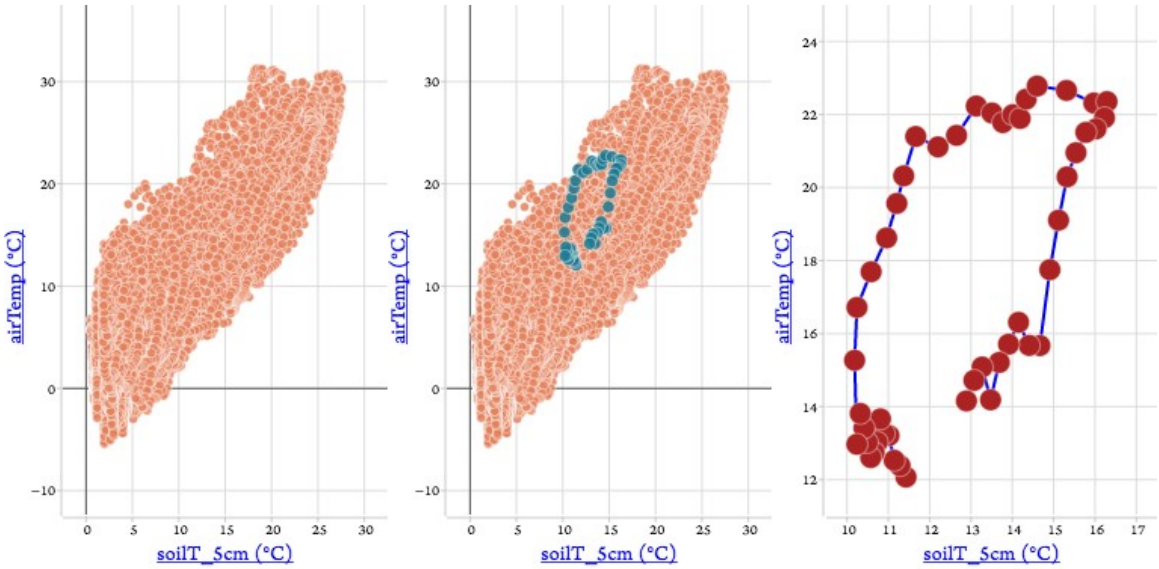


Figure 3. Left: Air temperature versus soil temperature (5 cm depth) every 30 minutes for the year 2000. Center: we have highlighted the data from May 19. Right: showing only the data from May 19, 2000, rescaled to fill the plot. URL: http://short.concord.org/b5

This forest example demonstrates how filtering is not just for getting a subset of data from a larger dataset; it's also a tool for exploring data and revealing patterns. This kind of filtering—to obtain a "slice" of data—simplifies the data by reducing its dimensionality. Understanding slicing in this way might also be a help in understanding more sophisticated methods of dimension reduction such as principal components analysis. Notice how it's also related to a crucial science move: controlling variables.

3.2. Grouping

Grouping is typically used to set up a comparison among different subgroups of a dataset. In an earlier example, Lynn used grouping to separate the deck of cards into groups by age and reported gender. Just as filtering restricts analysis to a single subset, grouping divides a dataset into multiple subsets. This division is guided by the available value(s) of some attribute or attributes so that, among the cases within each resulting group, the values of these "grouping" attributes are the same.

Binning is a special type of grouping that uses ranges of continuous values (bins or classes) to determine group membership. Imagine Lynn decides to prepare to make a histogram with the data cards described earlier. They could organize cards into groups—heights 120 to 129.9, 130 to 139.9, etc., labeled, and counted to determine the heights of the histogram bars. Depending on the way in which groups are defined (groups with a 5 cm range versus those with a 20 cm range), the resulting histogram may reveal very different types of patterns.

3.3. Summarizing

Analysts often compute values that summarize a group (even if the group is the entire dataset). *Summarizing* is the process of producing and recording a summary or aggregate value, i.e., a statistic. In the cards metaphor, Lynn summarized each age/gender group by computing the mean of the relevant heights; wrote the new values on the group labels; then used those values to make a graph (Figure 2, left-hand side).

Although the mean is the go-to summary function for many applications, there are a wide variety of summary measures, and "summary" does not necessarily mean "numerical" or "typical." Often, the point of summarizing is not even the chosen aggregate measure, or the results of that measure across groups. The purpose may be deeper: The value of an aggregate measure summarizes a group, and that summary value can then be used as data in further analysis.

Grouping and summarizing work together to help an analyst get a simpler display or dataset—many fewer points!—that more clearly shows an overall pattern. (Later in this paper, we discuss how the move of making a table hierarchical can facilitate grouping and summarizing in unexpected and powerful ways.)

Note, though, that consolidation into simpler distinct categories leads to a reduction of information. For example, when a display shows only measures of center, variability is lost. Furthermore, the anticipation of grouping and summarizing can lead to data collection design that may oversimplify reality; one example of this is in NHANES's reporting of gender identity as binary.

3.4. Calculating

Another data move is to create a new attribute, often represented by a new column in a data table. Because this typically involves calculating the values in this new attribute using a formula, in this paper, we call this data move *calculating*. Statisticians sometimes refer to calculating as "mutating" or "transforming."

Many new attributes are calculated using the values from one or more existing attributes. A good example within the NHANES context is Body Mass Index (BMI), which combines an individual's height and weight values to create a measure that some medical professionals find to be more informative than weight alone.

In the data cards metaphor, Lynn could go through every card and write in a new attribute with its value according to a formula or set of rules. Summary measures function as new, conceptual attributes as well; the difference is that they appear on group labels rather than individual data cards, and they use summary functions (such as mean) that use values from the whole stack

of cards. In this way, the data move of *summarizing* is a special case of *calculating*.

In addition to conceptual attributes, *calculating* can also be used to create convenience attributes. For example, one may wish to create a categorical attribute whose value is “tall” if an individual’s height is greater than the mean height for their age, and “short” otherwise. The new attribute can then be used in further data moves, for example, *filtering* to study only children who have been identified as tall, or *grouping* to compare children in the tall and short categories.

Convenience attributes are quite common. Other examples we can imagine with NHANES include:

- Creating a new column that converts heights to inches instead of centimeters.
- Using birth dates included in a dataset to compute subjects’ ages.
- *Recoding* an education attribute from several categories (e.g., “GED,” “high-school graduate,” “one year of college,” “bachelor’s degree,” etc.) to fewer (perhaps, “completed high school,” “completed college”).

3.5. Merging/Joining

Merging lets analysts combine multiple datasets into one. The simplest form of merging concatenates datasets about the same phenomenon but from different sources, for example, combining height data from two different classrooms to make a larger dataset. An equivalent action using the card deck analogy is simply to put two decks together. Combining datasets often requires additional preparation such as making sure that the names of the attributes are the same and that the codes and units are compatible. Such preparation is likely to involve additional data moves such as calculating, grouping, and filtering.

Joining is a more complex form of merging. It does not add new cases, but rather adds more information—new attributes—about existing cases from a separate dataset. The NHANES data we have been discussing throughout this paper is an example of this. As originally downloaded, Gender, Age, and Height are stored in different data tables depending on how they were collected. Gender and Age are from the “demography” table, while Height is from “body measurements.” Each table has a “sequence number” attribute whose values act as unique IDs for individuals. This sequence number is used to connect the datasets, letting the software, for example, add the Height attribute to a dataset that has Age and Gender, and fill in the correct corresponding values. (This is fundamental to using a relational database.)

In the card deck metaphor, joining can be imagined as operating on two different decks of cards: the “demography” deck with ID, Gender, and Age on

each card; and the “measurements” deck with ID, Height, and other such measurements on each card. One way to proceed is to go through the demography deck, and for each card, find the card in the “measurements” deck with the same ID. The Height value (and any other values of interest) could be copied from the second card to the first one, so that this information is now available in first deck for further analysis.

3.6. Making Hierarchy

It is often the case that data are nested within multiple levels. For example, the *American Community Survey* supplies records of individuals as one dataset with attributes including age, race, and income. It also supplies a separate dataset with records of households including when each home was built, whether it’s rented or owned, whether it’s on a farm, and so forth. The datasets are linked through the household to which each individual belongs. This is a relational structure, for every individual, we use an index to find associated household information.

The structure is also *hierarchical*, sometimes called “nested” by statisticians. Every individual belongs to exactly one household and every household may contain one or more individuals. In that sense, the household is superordinate or a “parent” to the individual. As a result, it’s possible to explore relationships between the attributes of the households and the attributes of the individuals.

As a simple example of this, Figure 4 shows the relationship between household size and whether the dwelling is rented or owned, for 1109 Wisconsin residents in 2016. (Data from IPUMS: Ruggles et al., 2017.) It appears that the mean number of people in a household is higher for owners than for renters.

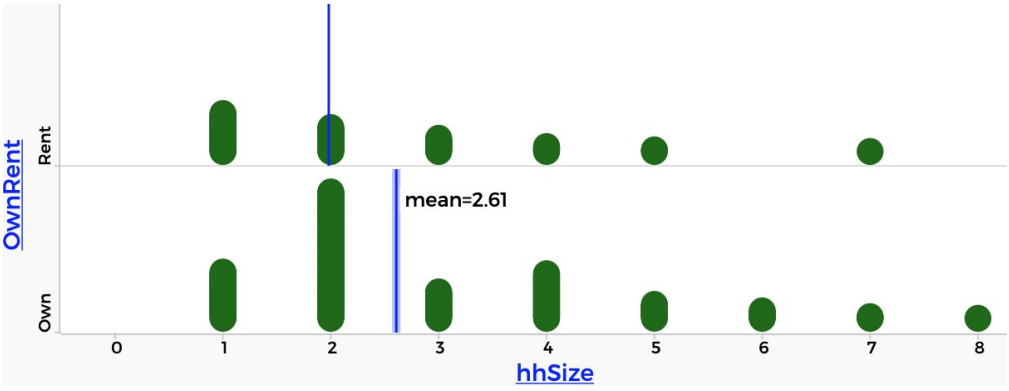
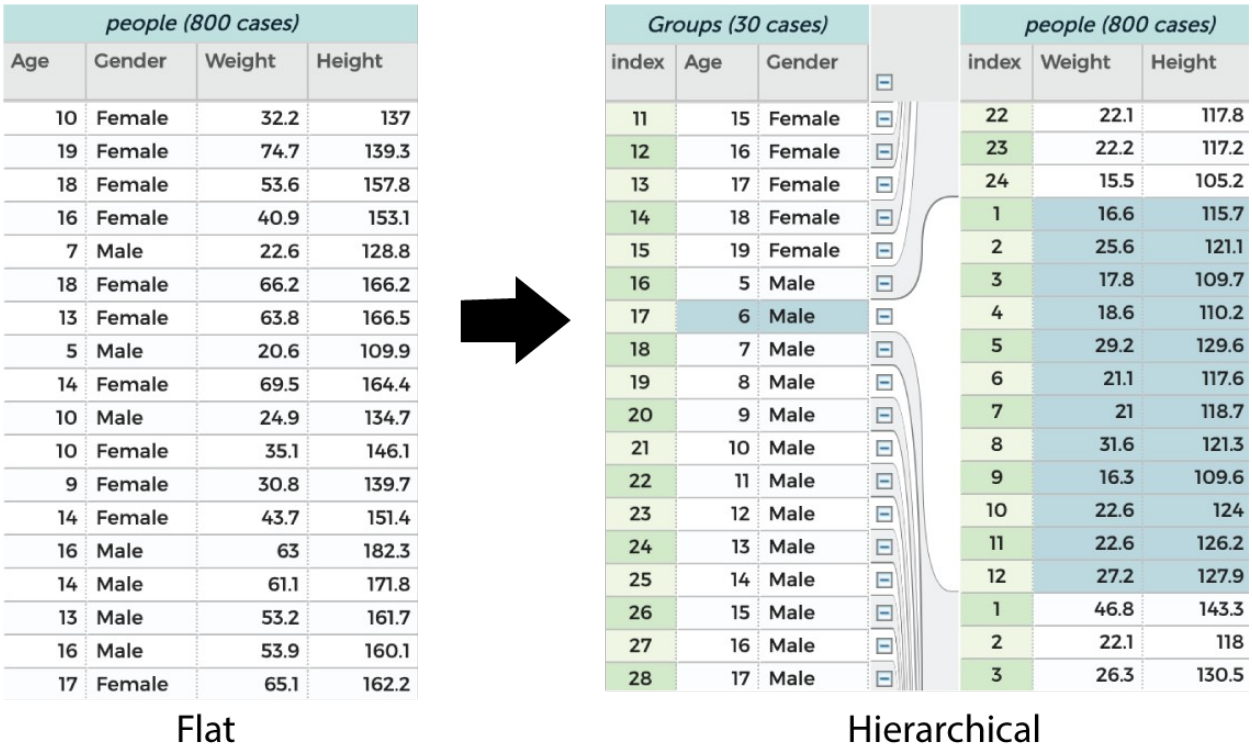


Figure 4. Relationship between household size and whether the dwelling is rented or owned. The lines show mean household size. URL: http://short.concord.org/b6

In the cards metaphor, there are two decks—people and households. An analyst would lay out all of the household cards and then place each person card next to its associated household. Now each household has a stack. The analyst counts the stack—and writes that number down on the household card. In this way, the household cards become the labels for groups.

That example used two datasets that have a hierarchical relationship to one another. Putting the datasets together to connect the number of individuals per household to whether that household is owned or rented is an example of *joining*.

But a single dataset can also be *made* hierarchical, even if it is not originally structured as such. Consider again the NHANES data from which we made the graph in Figure 2 showing mean height for various values of Age and reported Gender. That dataset was “flat”: it can be displayed as a rectangular array of cells, one case per row, one attribute per column. An analyst may, however, choose to move Age and Gender to a higher level in the hierarchy; part of the resulting table appears at the right in Figure 5.



people (800 cases)

Age	Gender	Weight	Height
10	Female	32.2	137
19	Female	74.7	139.3
18	Female	53.6	157.8
16	Female	40.9	153.1
7	Male	22.6	128.8
18	Female	66.2	166.2
13	Female	63.8	166.5
5	Male	20.6	109.9
14	Female	69.5	164.4
10	Male	24.9	134.7
10	Female	35.1	146.1
9	Female	30.8	139.7
14	Female	43.7	151.4
16	Male	63	182.3
14	Male	61.1	171.8
13	Male	53.2	161.7
16	Male	53.9	160.1
17	Female	65.1	162.2

Groups (30 cases)

index	Age	Gender
11	15	Female
12	16	Female
13	17	Female
14	18	Female
15	19	Female
16	5	Male
17	6	Male
18	7	Male
19	8	Male
20	9	Male
21	10	Male
22	11	Male
23	12	Male
24	13	Male
25	14	Male
26	15	Male
27	16	Male
28	17	Male

people (800 cases)

index	Weight	Height
22	22.1	117.8
23	22.2	117.2
24	15.5	105.2
1	16.6	115.7
2	25.6	121.1
3	17.8	109.7
4	18.6	110.2
5	29.2	129.6
6	21.1	117.6
7	21	118.7
8	31.6	121.3
9	16.3	109.6
10	22.6	124
11	22.6	126.2
12	27.2	127.9
1	46.8	143.3
2	22.1	118
3	26.3	130.5

Flat

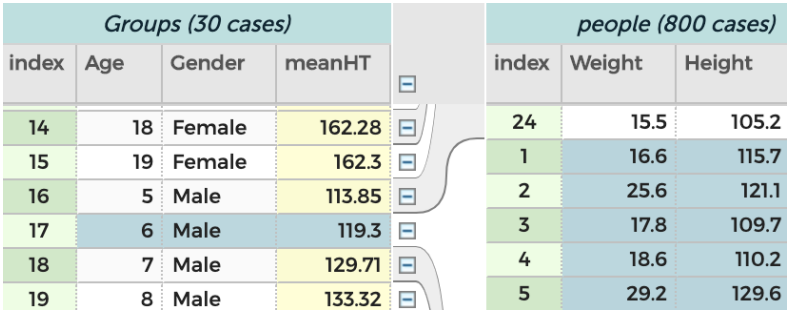
Hierarchical

Figure 5. Left: Flat table of the heights data. Right: hierarchical table of the same data. The 6-year-old males are highlighted. URL: http://short.concord.org/b7

On the left of Figure 5, in the flat table under “people (800 cases),” each case has Age, Gender, Weight, and Height. There are 800 cases in the table —800 rows.

The right side of Figure 5 has two linked tables, side by side. The table under “Groups (30 cases),” has only Age and reported Gender. There are 30 different combinations, so that table has 30 rows. The table at the far right still has the 800 cases, but with only Weight and Height. The two tables are dynamically linked so that when the 6-year-old male row in “Groups” is selected, the 12 corresponding cases in the “people” table are highlighted.

The large dataset is thus split into 30 smaller ones that appear one above the other. And now the payoff: a new attribute at the “Groups” level can contain summary calculations that apply to each of the groups separately. For example, a new attribute (mean height or “meanHT”) can be added to Groups. The resulting table appears in Figure 6.



Groups (30 cases)

index	Age	Gender	meanHT
14	18	Female	162.28
15	19	Female	162.3
16	5	Male	113.85
17	6	Male	119.3
18	7	Male	129.71
19	8	Male	133.32

people (800 cases)

index	Weight	Height
24	15.5	105.2
1	16.6	115.7
2	25.6	121.1
3	17.8	109.7
4	18.6	110.2
5	29.2	129.6

Figure 6. Hierarchical table with mean height (meanHT) at the “Groups” level.

Notice that the hierarchical table in Figure 6 is sufficient to create the height-and-age graph shown in Figure 2; it’s parallel to the way Lynn did it with cards. In this way, hierarchy offers an elegant, alternative way of thinking about grouping and making summary calculations. The “parent” cases are the groups, and a summary measure is an attribute—a calculated column—at that level of the hierarchy.

Making hierarchy is especially powerful as a data move because analysts can manipulate the structure of a dataset for a purpose rather than simply coping with the structure they were given. We argue it can give analysts a deeper understanding of *grouping* and *summarizing*, and how aggregate, calculated attributes can be created as “first class” variables for use in analysis. For these reasons, we treat *making hierarchy* as distinct from *grouping*. One can group without using the conceptual or technical machinery of hierarchy. Furthermore, even though some research (Haldar et al., 2018; Konold et al. 2014) has shown that quite young students understand and can use hierarchy, it seems a qualitatively different approach worth keeping separate for now.

3.7. Beyond the Core Data Moves: Connecting to a Changing Curriculum

We initially proposed the six data moves described above to be included in a core set, distinct from related activities such as graphing. However, as data science continues to enter the statistics curriculum and related disciplines, we expect additional data moves and related activities to become increasingly relevant. In this section, we examine the relationship between data moves and related activities such as visualization and simulation in more detail, and explore a few candidate moves to support those related activities. We hope that the broader statistics education and data science communities will help us recognize and characterize others as appropriate.

3.7.1. Visualizations and Data Moves

We made a point about how, in this paper, we do not consider making a graph to be a data move. However, data moves and visualizations are closely related. Some graphs are impossible without certain data moves. For example, to create a typical bar chart one needs to group data and then summarize the groups by counting the number of cases or performing some other summary calculation. If you want those bars ordered by height, you need *sorting*, which could easily be a data move with many uses. Thus the state of the data determines what graphs are possible; and a desire for a certain type of graph demands particular data moves.

The importance of the relationship between data moves and visualization becomes even more clear when—as is often the case in a data science application—*none* of the common graph types meet an analyst's needs. Understanding data moves opens a space of nonstandard alternatives. e.g.: "What if we make parallel box plots of height for each age, but instead of plotting the upper and lower quartiles, let's make the rectangles extend from the 10th to the 90th percentiles in income?" To make that graphic requires the output of new summary statistics, grouped appropriately. Put another way, knowing about data moves helps practitioners be creative with visualizations.

3.7.2. Cleaning Data

Cleaning data is often a prerequisite for performing data moves. When someone first gets data from a source, it usually needs work to be truly usable. There are several types of fixes that may need to be employed. Such fixes include wrangling data formats (e.g., decimal characters); fixing pathologies (e.g., using "N/A" or 999 for missing data); and coping with special data types such as dates and geocoding data. Cleaning might also involve recoding, as described above.

Sometimes data are recorded in a way that is inconvenient for computing. For example, a table of heights might have one column listing the heights for boys and another listing those for girls. But more functions become available

in many analysis packages if datasets follow a different convention with one row per case, and separate columns for each attribute (i.e., *tidy* data; Grolemund & Wickham, 2017). In the example given, each row would correspond to an individual, and there would be two columns, one for gender identity and one for height. Converting between these two data structures, sometimes called *stacking* or *re-structuring*, could be a data move.

Although this phase of data analysis is an important part of real-world data analysis, it's not clear how much of it should be part of our treatment of data moves. One could argue, as we did with graphing, that these might be a category of moves of their own that are qualitatively different from the moves that are the focus of this paper.

3.7.3. Sampling, Simulations, and Special-Purpose Data Moves

We envision *sampling* as a move related to *filtering*, but where the choice of cases is usually random rather than purposeful. This is important in simulation and machine learning. For example, simulation-based inference—e.g., randomization tests, bootstrapping, and the jackknife—involves creating a sampling distribution, which, in turn, benefits from understanding *grouping* or *hierarchy*: each sample is a new group, and its statistic is a summary measure at the "parent" level of the hierarchy. In machine learning, one can use *sampling* to create a training set.

4. DISCUSSION

As explained at the beginning of this paper, we want to help students engage productively with larger, richer datasets. We identified characteristic data moves that are often necessary in order to conduct such data explorations. We also briefly considered when such moves are useful, with the assertion that not only knowing *how* to make moves, but *why* to make them is important. In this section, we will summarize the moves so far, then address some issues and give suggestions.

4.1. Summary, with Connections among Data Moves

Data moves are related in complicated ways. Here, we reflect on the purposes for and connections among our core data moves as a vehicle for summarizing them.

- *Filtering* is used for scoping and exploration. It is conceptually a prerequisite for *grouping* and *sampling*.
- *Grouping* is fundamental for comparing. One could argue that grouping is really just repeatedly *filtering*, but it's so common that we give it its own name.
- *Summarizing* creates aggregate measures that describe a *group* (which could be the entire dataset).

- A *hierarchical organization* can be equivalent to *grouping*, and provides an alternative way of thinking about *summarizing*.
- *Calculating* a new attribute involves describing a new idea in terms of other attributes. If the new idea is at a "parent" level in a *hierarchy*, and uses the values in the "children," this is equivalent to *summarizing*. Sometimes a calculated attribute facilitates *grouping* or *filtering*.
- *Merging* or *joining* datasets through a relation is a way to add cases or *computed attributes* from outside the original dataset. Some hierarchical data sets are "hooked up" by joining.

As data science and related activities such as visualization become a more central focus of introductory courses, other data moves may join this core set of six. For example:

- As suggested above, data moves are intimately connected to visualizations; students will need data moves to facilitate creating new ways of displaying data.
- Simply *sorting* a dataset can give important insights into patterns in the data, and is an essential element in helping some visualizations communicate more effectively.
- There are a number of ways to go about cleaning a dataset so that it can be imported to an analysis package. This can include *merging, calculating* (especially *recoding*), and *stacking*.
- *Sampling* and related processes are important to simulation-based inference. Sampling is related to *filtering*. A sampling distribution collects *summary* values from these samples, and can be facilitated by *hierarchy*.

4.2. Data Moves: Missing in Traditional Introductory Statistics

Data moves seem to play a larger role in the data science curriculum than in introductory statistics courses. Statistics students are seldom asked to engage in the reasoning or actions described here. Why is that? One reason may be that the data used in introductory statistics courses tends to be pre-aggregated; a textbook problem often provides the means and standard deviations of the two groups and asks students to choose a test. Even if students have the original case-level data, those data and the tools students are expected to use for analysis are set up so that students don't have to think too hard about the data moves they are making. If they are asked to compare income by education, students get data only for income and education. It is relatively straightforward to decide in this set up which one is the grouping variable. As a result, students can get a correct answer without really understanding why grouping was a useful move.

Another barrier for engaging with data moves in introductory statistics courses is time. If students were to encounter more complicated data situations, they would need computer skills to resolve them, and it would require class time to teach those skills in addition to statistics. Consequently, introductory statistics students tend to work with sanitized data and analyses in order to keep the pedagogical focus on the "important statistical content."

All of this begs the question: if not in an introductory statistics course, where are students expected to learn data moves? Many statistics educators seem to assume that issues of "altering a dataset's contents, structure, and values," in addition to not being actual statistics content, are something students would address later in the curriculum. Although these issues can be messy, they may not be perceived as cognitively demanding, essential, and foundational when compared to, for example, understanding the true meaning of a confidence interval. Or perhaps the computational side was assumed to be too difficult, the province of graduate students and faculty, too abstruse for high schoolers and undergraduates.

That made sense in 1990. But since then, rich datasets and their attendant computational needs have become ubiquitous and more accessible. This has, in turn, given rise to data science, computational statistics, machine learning, dynamic visualizations, and allied fields. Data science is changing the nature of work, science, and society; it is used for everything from public health and climate change mitigation to targeted campaigning and warfare. We are all, as citizens, both beneficiaries and victims of data science, and lack of attention to the ways data are constructed and manipulated have already reproduced and exacerbated ongoing injustices (Noble, 2018; O'Niell, 2017). This changes the educational landscape. At least it should.

4.3. Four Recommendations

Given the increased role of data in today's world, we think attention to data moves would build students' understanding and ability to work with increasingly complex datasets. Here are four concrete recommendations for statistics educators to consider as curriculum and practice evolve.

First, include data moves explicitly as a part of data analysis. Students with different levels of experience with coding and statistics will come with very different understandings of underlying concepts like grouping, summarizing, and hierarchy. Especially in early assignments, leaving data moves implicit may be asking some students to learn too many new things at once. The equity component is important here; assignments that highlight data moves might help level the playing field (adopting the strategy that led to the College Board's (2017) AP Computer Science Principles), or at least give the class common experiences and vocabulary. Another approach would be to acknowledge, justify, and name data moves as they appear when the instructor uses them, so that students see how the data moves recur and fit

together. Instructors and course designers would likely also benefit from recognizing data moves and thinking about which ones students are expected to use, and in what sequence.

Second, early assignments should be more computationally demanding—and less "sanitized." Large, rich, datasets can be very interesting and motivating for students (see Erickson 2012 as an example at the high-school level), and are worth class time even if they do not always illuminate orthodox frequentist inference. If data moves are prioritized in the structure of such lessons, whatever the students learn will be useful later as well. Statistics topics may have to readjust to include more computational thinking; this is a serious undertaking, and data moves could help organize the journey. Kaplan (2017) offers a thoughtful, compatible take on updating the introductory course.

Third, use data moves to help students transition between tools. McNamara (2015) and Rubin and Erickson (2018) have explored transitioning between "learning" software (such as CODAP) and "professional" software (RStudio, Jupyter Notebook) when the former is no longer powerful enough. Referring to data moves could help students understand that many R commands are equivalent to the drag-and-drop actions they know from CODAP. Tool designers could also use this formulation in order to see how or whether their tools implement these moves. (Many of the data moves we have described correspond quite explicitly to features in more powerful systems. In the tidyverse, for example, one uses `filter()` for filtering; `mutate()` for calculating new values; `group_by()` to do grouping, and so forth.)

Fourth, consider data moves as part of data literacy. We argue that it is important for more of the general public to understand what is possible in the interaction between computation and data. Data moves make clear how data can be manipulated, and how decisions about data can enable or constrain the types of investigations that are possible. Equity is here again a concern—understanding how data and its construction and manipulation can reproduce or interrupt harmful discourses, even about students themselves (Philip, Schuler-Brown & Way, 2013). By experiencing and manipulating data more directly, students have more opportunities to interrogate and change datasets, to ask questions, to encounter and explore constraints, and to learn to use data to explore what interests them most.

4.4. What Else Should We Think About and Do?

Identifying data moves and using them as a way to think about computation and data is an intriguing idea. But, thus far it is based largely on conjecture and limited experiences. Do data moves productively capture students' data analysis activities? Is use of data moves associated with richer or more ambitious investigations; with a sense of ownership and authorship with datasets? What types of instruction and supports enable students to gain confidence in manipulating data? Some studies (e.g., Wilkerson & Laina,

2018) are emerging, but we need more. To save the reader from even more exposition—for the time being—we will simply present a slew of potential questions that could be explored:

- How do data moves relate to frameworks for data analysis such as GAISE (Franklin et al., 2005) and PPDAC (MacKay and Oldford 1994, Wild and Pfannkuch 1999)?
- Where do students typically have trouble with data moves? Some quick suggestions: (a) they sometimes have trouble—as suggested in the sections on *summarizing* and *stacking*—distinguishing the names of attributes from their values, and that can lead to trouble with grouping; (b) making lots of groups (e.g., 365 days in a year) feels very different to the novice from making two—and may not even seem like grouping; and (c) in general, students may be good at learning specific steps and tools, but using data moves well means being flexible and understanding how a move may address the specific needs of a given problem.
- How should we teach about data moves? They are very abstract as we have presented them here—not the best exposition for beginners. But calling them out as they unfold during instruction, and noting the consequences, can't hurt: *Look! When I group the data by age and summarize by taking the mean, I can make a graph with only 15 points instead of 800. Not only that, now you can really see how the growth patterns for teenagers are different from the children*. Also, the cards metaphor might be useful for explanations.
- What's the difference between a good data move and a great data move? How can we tell? Wild and Pfannkuch (1999) discuss quality in statistical thinking. How is this different?
- What's the connection between data moves and computational thinking (e.g., in the K–12 Computer Science Framework, 2016)? If data science requires computational thinking, are data moves evidence of it?
- Should we do this kind of analysis for modeling and visualizations? Perhaps we will find *graphics moves* and *modeling moves*. For visualizations, we can begin in the tidyverse with the "grammar of graphics" in Wilkinson (2005) or Wickham (2010). That grammar is a deep and thorough structuring of data graphics from the computational point of view. Would it look any different coming from statistics education as opposed to computer science?
- Data moves are not neutral or value-free. As suggested above, people of color, LGBTQ+ people, and other oppressed groups are disproportionately harmed and marginalized by biases in study design, data construction and manipulation, and analysis. How shall we, as data science educators,

illuminate and fight those biases in a way that results in meaningful change?

- How do habits of mind, dispositions, practices of data analysis, and elements of data craft—none of which are captured in this analysis—relate to data moves?

5. CONCLUSIONS

When analysts carry out rich investigations with data—including data science tasks—they use *data moves* to prepare data, derive new data values, and organize data, in order to produce visualizations and other results. In this paper, we have tried to enumerate and characterize data moves independent of how they are implemented using a particular language or tool.

This paper defines data moves narrowly, to be about the data—and not, for example, about visualizations or modeling. Even so, data moves encompass a wide variety of actions that we employ for a wide variety of purposes, and they connect to and build upon one another. Indeed, we would contend that data moves do not exist in isolation from the investigatory contexts—including the datasets, tools, motivating questions, and learners' goals—that necessitate their use.

We suggest that educators should consider data moves as they think about students' opportunities to learn, curriculum materials, and assessment techniques. We also acknowledge that this is a very early analysis of this topic in a field that is changing rapidly; we look forward to continuing the conversation.

6. REFERENCES

Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2003). *National Health and Nutrition Examination Survey data*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved from http://www.eeps.com/zoo/nhanes/source/choose.php, January 2018.

Chick, H. (2004). Tools for transnumeration: Early stages in the art of data representation. In *Mathematics Education for the Third Millennium: Towards 2010. Proceedings of the Twenty-seventh Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 167-174)

Chick, H., Pfannkuch, M., & Watson, J. (2005). Transnumerative thinking: Finding and telling stories within data. *Curriculum Matters*, *1*, 87–109.

The College Board. (2017). *AP computer science principles: Course and exam description*. New York: CollegeBoard. https://apcentral.collegeboard.org/pdf/ap-computer-science-principles-course-and-exam-description.pdf

Erickson, T. (ed.) (2012). *Signs of change: History revealed in U. S. Census data*. Oakland: eeps media.

Erickson, T. (2017). Blog post, *More about data moves and R*. https://bestcase.wordpress.com/2017/07/18/more-about-data-moves-and-r/.

Finzer, W. (2014). Common Online Data Analysis Platform [Computer Software] (CODAP). Concord, MA: The Concord Consortium. https://codap.concord.org/releases/latest/static/dg/en/cert/index.html

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report*. Alexandria, VA; American Statistical Association. http://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics* (4th edition). New York: W W Norton.

Grolemund, G. and Wickham, H. (2017). *R for Data Science*. O'Reilly. http://r4ds.had.co.nz/

Haldar, L., Wong, N., Heller, J. and Konold, C. (2018). "Students making sense of multilevel data." *Technological Innovations in Statistics Education*, 11(1). Retrieved from https://escholarship.org/uc/item/7x28z96b

Ham, K. (2013). "OpenRefine (version 2.5, now 2.8). http://openrefine.org. Free, open-source tool for cleaning and transforming data," *Journal of the Medical Library Association: JMLA*, 101(3), 233–234. doi: 10.3163/1536-5050.101.3.020

K–12 Computer science framework (2016). Retrieved from http://www.k12cs.org.

Kaplan, D. (2017). "Teaching stats for data science," *American Statistician*. In press. doi: 10.1080/00031305.2017.1398107

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, B., & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization, 10*(4), 271–288. doi: 10.1177/1473871611415994

Konold, C., Finzer, W., and Kreetong, K. (2014), "Students' methods of recording and organizing data," Paper presented at the annual

meeting of the American Educational Research Association, New Orleans, LA.

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. doi: 10.1007/s10758-007-9122-2

MacKay, R.J. & Oldford, W. (1994). *Stat 231 course notes fall 1994*. Waterloo: University of Waterloo.

McNamara, A. (2015). Ph.D. Dissertation: *Bridging the gap between tools for learning and for doing statistics*. UCLA.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press: New York, NY, USA.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books: New York, NY, USA.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM - Mathematics Education*. doi: 10.1007/s11858-018-0989-2

Philip, T. M., Schuler-Brown, S., & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning*, *18*(3), 103-120.

Ridgway, J., Nicholson, J, Borovcnik, M., & Pfannkuch, M (2017). Statistical literacy [Special issue]. *Statistics Education Research Journal, 16*(1).

Rubin, A. and Erickson, T. (2018). “Panel summary: Tools, best practices, and research-based reminders,” in Biehler, R. et al. (eds.). *Paderborn Symposium on Data Science Education 2017: The Collected Extended Abstracts*. Paderborn: University of Paderborn.

Ruggles, S., Genadek, K., Goeken, R., Grover, J. and Sobek, M. (2017). *Integrated Public Use Microdata Series: Version 7.0* [dataset]. Minneapolis, MN: University of Minnesota. doi: 10.18128/D010.V7.0

Thomson, A. (2018). Private communication. *Blodgett Forest data*, University of California. Visit http://ucanr.edu/sites/cff/Blodgett_Forest_Research_Station/

Wickham, H. (2010). “A layered grammar of graphics,” *Journal of Computational and Graphical Statistics*, **19 (1)**. 3–28. doi: 10.1198/jcgs.2009.07098

Wild, C. and Pfannkuch, M. (1999). “Statistical Thinking in Empirical Inquiry,” *International Statistical Review* **67(3)**, 223–265. https://iase-web.org/documents/intstatreview/99.Wild.Pfannkuch.pdf

Wilkerson, M., Lanouette, K., Shareff, R. L., Erickson, T., Bulalacao, N., Heller, J., St. Clair, N., Finzer, W., & Reichsman, F. (2018). “Data moves: Restructuring data for inquiry in a simulation and data analysis environment.” In J. Kay & R. Luckin (Eds.), *Rethinking learning in the digital age: Making the learning sciences count*, *Proceedings of the 13th International Conference for the Learning* Sciences (ICLS 2018) (Vol. 2, pp. 1383–1384). London, England: ISLS.

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.

Wilkinson, L. (2005). *The grammar of graphics*. New York: Springer-Verlag.