Data Representations and Visualizations in Educational Research

Ting Dai [a]

Joshua M. Rosenberg [b]

Michael Lawson [c]

[a] Educational Psychology, College of Education, University of Illinois at Chicago

[b] Department of Theory and Practice in Teacher Education, College of Education, Health, and Human Sciences, University of Tennessee, Knoxville

[c] Mathematics Education, Rossier School of Education, University of Southern California

Corresponding Author: Ting Dai, University of Illinois at Chicago, 1040 W Harrison St., MC 147, Chicago, IL 60607 USA. Email: tdai@uic.edu. Fax: 312-996-5651.

The graphical representation of quantitative information is not a modern development, but rather it can be traced back to the earliest map-making and, later, thematic cartography and statistical graphics (Friendly, 2008). The early 19th century witnessed the invention of all major forms of statistical graphics, including the ever so popular pie and bar charts, histograms, line graphs, and scatterplots. At this time, data from a wide variety of domains (e.g., economic, social, medical, physical) began to be depicted, and a wide range of novel techniques were used to facilitate data representation. At the same time, graphical analyses of natural and physical phenomena made regular appearances in scientific publications. In the second half of the 19th century, there was a rapid growth in the visualization of data: the importance of numerical information for public policy, industry, and health was acknowledged, and the various applications of statistical theory and methods made it easier to make sense of large bodies of data. This period has been referred to "the Golden Age" of data visualization (Friendly, 2008, pp. 12-13).

Another historically critical period of the development of data visualization is between 1950 and 1975 (Friendly, 2008). In this period, data analysis began being recognized as a distinct branch of statistics by the international research community and significant advances were made in the area of computer processing of statistical data, interactive statistical applications, and digital graphic technologies. Since the mid-1970's, data visualization has blossomed into a vibrant multi-disciplinary research area. It features characteristics such as highly interactive statistical computing systems, advanced visualizations of high-dimensional data, substantially increased attention to the cognitive and perceptual aspects of data display.

Data representations and visualizations have also become commonplace in the applied work of a variety of professions. Scientists, for example, use data visualizations to make sense of trends within their research that employs mathematical and statical models of phenomena and make such results understandable by others. Engineers use data representations to monitor environmental, commercial, and industrial processes. Historians and journalists also utilize data representations and visualizations to communicate information from a myriad of sources, including textual data. Finally, more recently, individuals—even students—have begun to use data representations and visualizations to understand aspects of their lives, such as their wellness and finances (Lee, 2019). Not left behind the data revolution are educational researchers, who use data representations and visualizations, which we use synonymously in this article, for many of the same reasons as other professionals and non-professionals—to understand and communicate results effectively.

While many engage with data representations and visualizations, a focus on the effectiveness of their design has often been ignored (Wilkinson, 2005). However, after a period where data representations and visualizations were seen by statisticians as, “a minor subfield and are not well-integrated with larger themes of modeling and inference” (Gelman & Unwin, 2013, p. 1), many professionals are beginning to take representation and visualization seriously. This is evidenced by the recent theoretical and practical work that is being done by the likes of Healy (2018), Wickham (2016), and Wilkinson (2005). Moreover, there is research and recent work in the broader fields of computer science, statistics, and sociology, to name a few, that can inform how we, as educational researchers, go about creating data representations and visualizations effectively. Finally, as we begin considering effective ways to represent and visualize data, it is important to consider findings from educational, psychological, and developmental research on how people interpret data representations and visualizations as we make related decisions. Thus, there is, presently, a greater focus on the effectiveness of data representations and visualizations, a focus which we aim to highlight and demonstrate through this article.

Aims of this Article

In this article, we summarize research on designing data representations and visualizations to communicate educational research findings effectively. Particularly, we focus on providing recommendations to make informed decisions about effective data representations and visualizations.

First, we focus on surveying the field of data representation and visualization. Here, we discuss various perspectives on the aesthetics of data representations, principles of designing data visualizations, and the grammar of data representations and visualizations. We then focus on five visualization categories for key points educational researchers make. In particular, we focus upon the following five categories of visualizations: 1) using visualizations of distributions, 2) comparisons of groups, 3) relationships between data, 4) trends over time, and 5) a few other directions such as infographics. We then bring the ideas from the field together with our understanding of these categories to draw attention to making decisions about data representation and visualization.

Developments and Past Research on Data Visualization

What makes a good graph? This question has flummoxed many over the years, as the phrase “good graph” can take up multiple meanings in a variety of settings, where a multiplicity of data representations and visualizations jockey for precious space in manuscripts. For this article, we explore the literature on *good data representations and visualizations* by focusing on the aesthetics, principles, and language of data representations and visualizations that communicate our data in an effective way to an intended audience.

Aesthetics and Principles of Data Representation and Visualization

The aesthetics of our data representations and visualizations force us to consider its underlying beauty. In a seminal piece on visualizing data, Tufte (1983) takes up these considerations by using a minimalist perspective to examine the aesthetic properties of data representations and visualizations. For instance, one of Tufte's principles is that our data representations need to have high *data-to-ink ratios*, where data-ink is the ink on a graph that represents data. This principle implies that when displaying data, we maximize data-ink and minimize non-data-ink, such that a data visualization does not act as, "content free decoration" in a report (Tufte, 1983, p. 177). While Tufte's minimalist perspective can lead to meaningful and beautiful visualizations, one issue with this work is that Tufte relies heavily on graphic design tools often inaccessible to educational scholars and provides little guidance for how data visualization designers should heed his recommendations. In addition, there exist many texts that support our learning and understanding of statistical analyses (e.g., Tabachnick & Fidell, 2013; Gravetter & Wallnau, 2013), but these texts often forget or minimize discussion about effective practices and principles to communicate one's findings effectively to an audience. As Wickham (2013) notes, "it is interesting to see how long statisticians have been pointing out problems with other people's visualisations [sic], and how little impact it has had" (p. 39). Wickham's observation is poignant as conversations about aesthetics and principles of data representations and visualizations have often existed on the periphery, but by communicating to diverse audiences, the conversation is becoming pertinent.

A Grammar of Graphics

In his 2005 text, *The Grammar of Graphics*, Leland Wilkinson, and his colleagues using SPSS, continues a discussion that shifts the conversation from focusing on representing and visualizing data to the actual *communication of information*. As Wilkinson notes, “Grammar makes language expressive... By specifying how words are combined in statements, a grammar expresses a language’s scope… The grammar of graphics takes us beyond a limited set of charts (words) to an almost unlimited world of graphical forms (statements)” (2005, p. 1). Wilkinson also makes the case that visualizations of data should be called graphics, rather than charts, because charts suggest a collection - or typology - of common representations; as such a collection “will have no deep structure” and will “inevitably offer fewer charts than people want” (2005, p. 2).

Healy (2018) builds off Wilkinson’s work and articulates the importance of connecting the structure of one’s data to the specific ways we represent it:

> Some approaches work better for reasons that have less to do with one’s sense of what looks good and more to do with how human visual perception works. When starting out, it is easier to grasp these perceptual aspects of data visualization than it is to get a reliable, taste-based feel for what works. For this reason, it is better to begin by thinking about the relationship between the structure of your data and the perceptual features of your graphics (p. 14).

Here, Healy makes the case that it is the structure of one’s data that should have a large impact on how we decide about representation and visualization. Structure here pertains not only to storing particular types of data (e.g., nominal, categorical) in particular ways but also how the data table or spreadsheet is organized. Longitudinal data, for example, can be stored such that each time point from each individual is stored in a row, with columns indicating with what time point the measure is associated and the value of the measure (“long” form) or with the value of the measures stored in distinct columns (“wide” form). How longitudinal data is structured impacts how the data is later visualized or modeled.

Healy also argues that psychological research on visual perception can, within the constraint of the structure of one’s data, be used to inform the creation of effective data representations and visualizations (Healy, 2018). Together, Healy and Wilkinson share a perspective that it is important for statistical software to invite users to consider the aesthetics of data representation and visualization by specifying the structure of one’s data and making decisions about how to represent the data in terms of size, shape, and color.

Summary

It is in this section that we aimed to situate this article within the discourse about effective data representation and visualization practices. We discussed designing representations that are aesthetically pleasing while still making a clear point, and while Tufte and Wilkinson’s ideas are important as we begin considering principles and best practices for representing data with the goal of communicating information, it takes some work to begin implementing these ideas in statistical software other than *SPSS* (Wickham, 2011, 2016). In his work, Wickham implemented principles from the Grammar of Graphics through the freely available and widely used statistical software *R*, in the software *ggplot2*. In the grammar of graphics, users must use some programming in order to specify the structure of one’s data and how to represent it, and once this process is understood, it is possible to make graphic representations and visualizations that are customized to one’s data and purpose. In addition, *ggplot2* implements graphics in a way that recognizes research on how people perceive graphs (see the above Grammar of Graphics section) and incorporates sensible default options for graph features such as default colors, gridlines, and tick marks (Healy, 2018). Thus, *ggplot2* has become one of the most popular packages in *R*, and educational and educational psychology scholars increasingly use R in their research (and teaching). In this article, we use the *R* to present data representations and visualizations using simulated data sets (which we make available for others along with the code needed to reproduce every data representation and visualization here: https://osf.io/3phcz/).

Visualizations Widely Used in Educational Research

As educational researchers, we often use data to convey an overarching finding from statistical analyses. In communicating these findings, it is often pertinent to provide a visualization to aid in understanding its meaning. In this section, we discuss data visualizations and representations that effectively attend to the points below and describe each graph (created using *R*) in detail.

For figures representing groups *a-d*, we use the research context of an elementary classroom and after-school intervention to provide an example of how different visualizations and representations answer different questions and interests. In this sample research context, there are two fifth-grade classrooms from four schools (groups *a-d*) in an urban southeastern United States school district.

In the research study, each class completes an inquiry-based lesson sequence, and within each class, you have some students making up an intervention group, those who attend an after-school program where students engage with inquiry-based learning activities, and other students making up the control group, those who do not attend the after-school program. In this case, the research team has collected data, particularly, the scores of student responses to an inquiry-based task and audio of classroom discourse during the activity. As we introduce data representations in the following sub-sections, we keep with this research context and consider potential research questions that align with the particular data representations and visualizations in the figures to follow.

1. Visualizing Data to Understand a Distribution

One of the keywords of statistics is distribution, how a collection of quantitative data varies. There are a number of distributions that statisticians' study and have summarized and utilized their key features and values in statistical testing, for instance, the normal distribution, *t*-distribution, Poisson distribution, *F*-distribution, and chi-square distribution. However, sample data that researchers encounter do not always conform to one of these known distributions. It is difficult for researchers to observe the characteristics of a sample data distribution without a proper visualization. It is also not intuitive to convey in word key features of a distribution to the intended audience without presenting the visual.

Box plot. The box plot, also known as the box-and-whisker plot, is extremely useful for displaying data distribution. At the center of the plot is the median (or the second quartile, Q_2) showing the value above and below which there are 50% of the scores. In this sense, median is the true middle point of the data. Unlike the mean, median is not sensitive to outliers, which makes it an equally important central tendency measure of data's distribution. A box plot shows not only the median (or Q_2) but also other quartiles of data's distribution. Surrounding the median line is a box, whose bottom line marks the first quartile (or Q_1) and the top line is the third quartile (or Q_3), representing where the middle 50% of the scores fall, also known as the interquartile range (IQR). Extending from the top and bottom of the box are two straight lines (or the whiskers), whose length is 1.5 times of IQR above the top and below the bottom of the box. The whiskers show the fuller range of the scores. Whiskers are useful in visually summarizing a distribution particularly for understanding the symmetry of a distribution: if the whiskers are of different lengths, it indicates an asymmetrical distribution (e.g., box plot for group b in Figure 1.1). Finally, the outliers of the score distribution are marked as dots or small circles beyond the whiskers, which further displays any skewness of the distribution due to extreme values (e.g., box plot for group a in Figure 1.1).

In this case, the distribution of an outcome, pretest math scores, for the four groups (*a*-*d*) could be visualized in order to begin to understand differences between them (and any potential issues with the data, such as miscoded values or potential outliers).

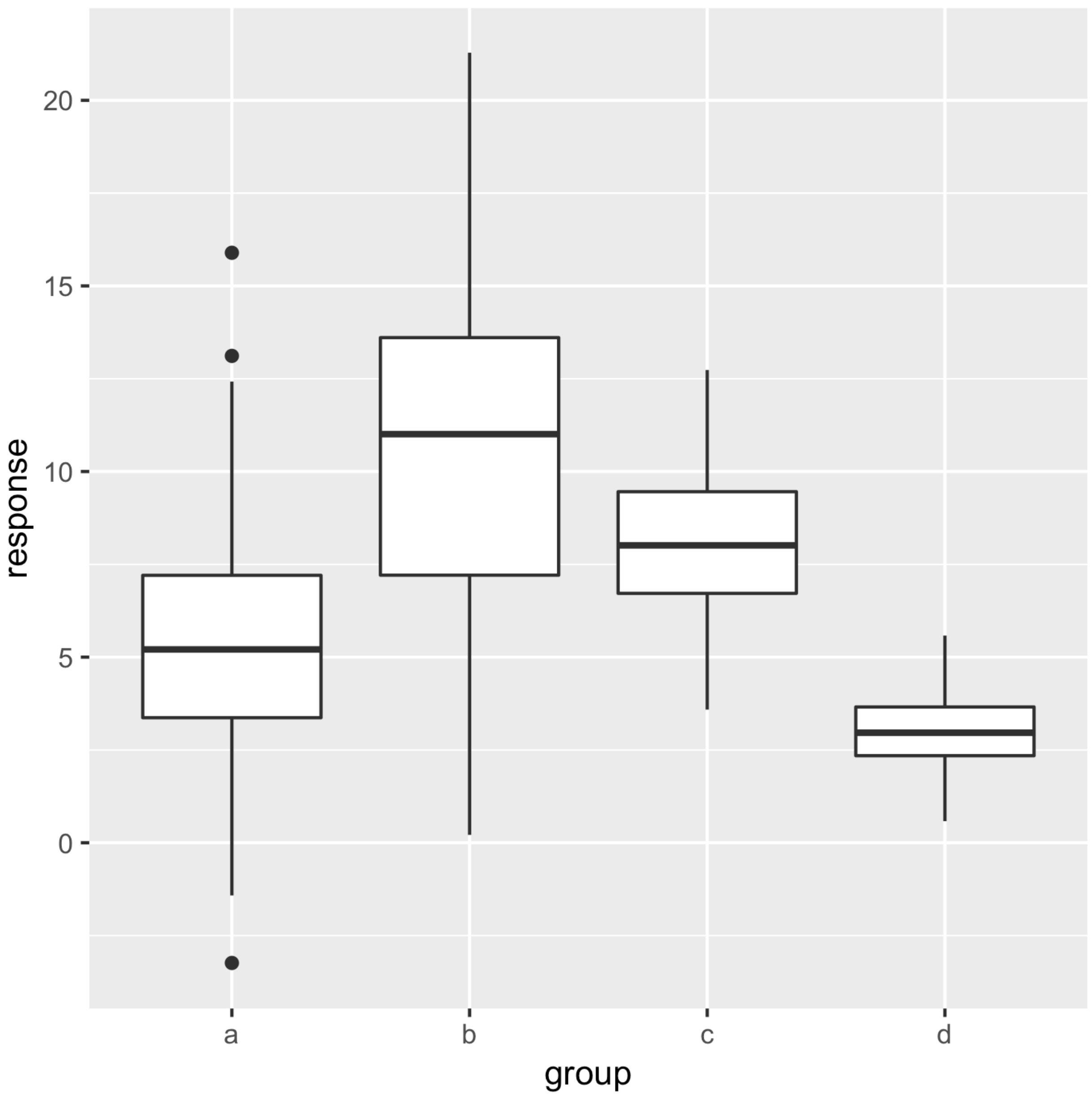

Figure 1.1. *Box plots of pretest math scores for Groups a - d.*

Box plot with raw data points. Sometimes, we may wish to display the box plot as a summary of the data along with the raw data. We can do this by adding the individual pretest math scores as a separate layer (Figure 1.2). Here, we add jittered data points, because there is no variation in the *x*-axis locations. Jittering simply adds random noise (here, only in the vertical direction) to avoid the over-plotting of points.

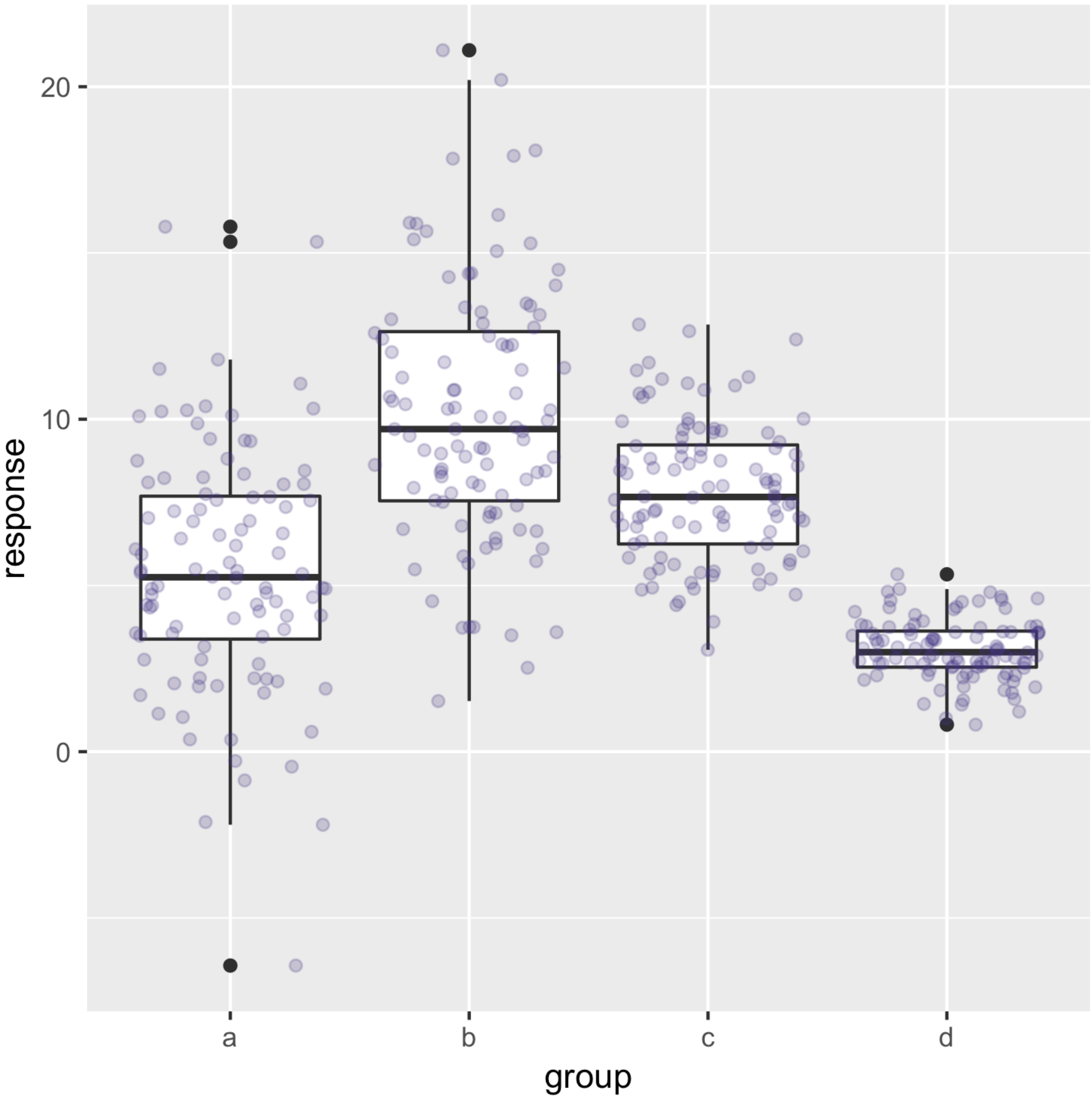

Figure 1.2. *Box plots of pretest match scores with individual data points for Groups a - d.*

Histogram. In a histogram, the *x*-axis usually represents all possible scores in a data set, and the *y*-axis represents the frequencies of the scores in the data set (Figure 1.3). This data visualization consists of three most important characteristics about a distribution of scores, including the place(s) where scores tend to congregate along the score continuum (central tendency or “average” score), the extent to which the scores are spread out (variability), and the symmetry of the score distribution. A histogram provides an intuitive display of the shape of the score distribution, which can be then described as *normal, skewed,* or *rectangular*, etc. Unlike in a box plot, outliers are not predefined, but because the full set of scores and their frequencies are displayed, one can visualize the scores that are far away from the central tendency and the rest of the scores and determine what constitutes as an outlier. In a histogram, the mode(s) of a distribution is arguably the most clearly displayed feature of a score distribution, which is the X-axis score corresponding to the highest peak of a distribution. A histogram is useful in visually describing a distribution, particularly for understanding the shape of a distribution including central tendency, the extent to which scores are spread out, and the symmetry of the distribution. Here, we present histograms faceted by each of the four groups of the 4th-grade students. Similar to the above examples, the histograms can provide a sense of how spread the data is – and whether the data is “behaving” as expected (i.e., in the histograms below, there do not appear to be miscoded values or potential outliers).

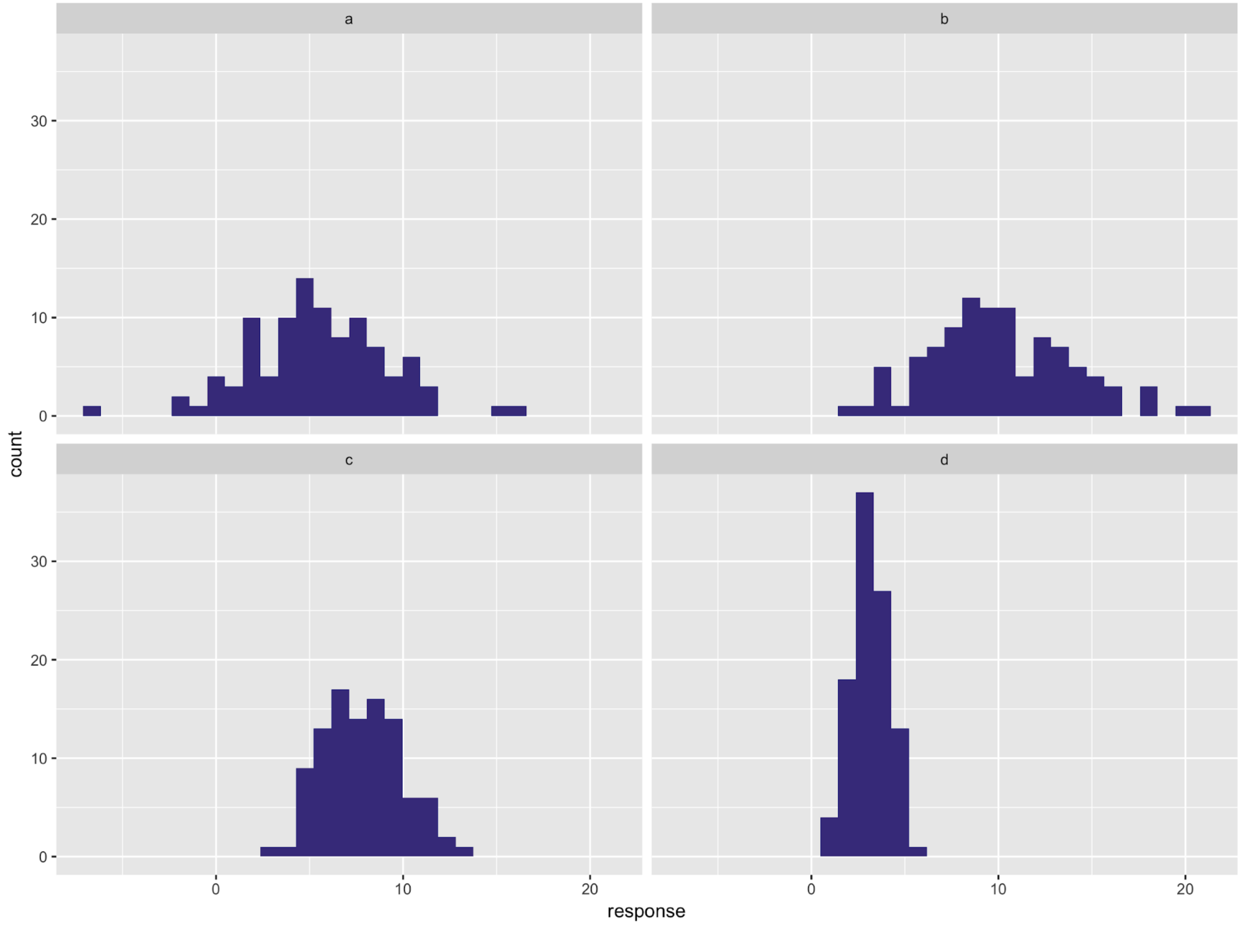

Figure 1.3. *Histograms of pretest math scores for Groups a - d*.

Density plot. A density plot is the distribution of the underlying probability of the data on a random variable represented with a continuous curve (See Figure 1.4 for the density plots for the pretest math scores of groups a - d). This curve—the underlying probabilities—is estimated from the observed sample data of a group (e.g., group a). Kernel density estimation is one commonly used estimation procedure, which uses a finite sample to make inferences about the population, also called a data smoothing procedure. Kernel density estimation does not assume normality, but among the most well-known kernels is the Gaussian kernel. A density plot is sometimes referred to as the kernel plot.

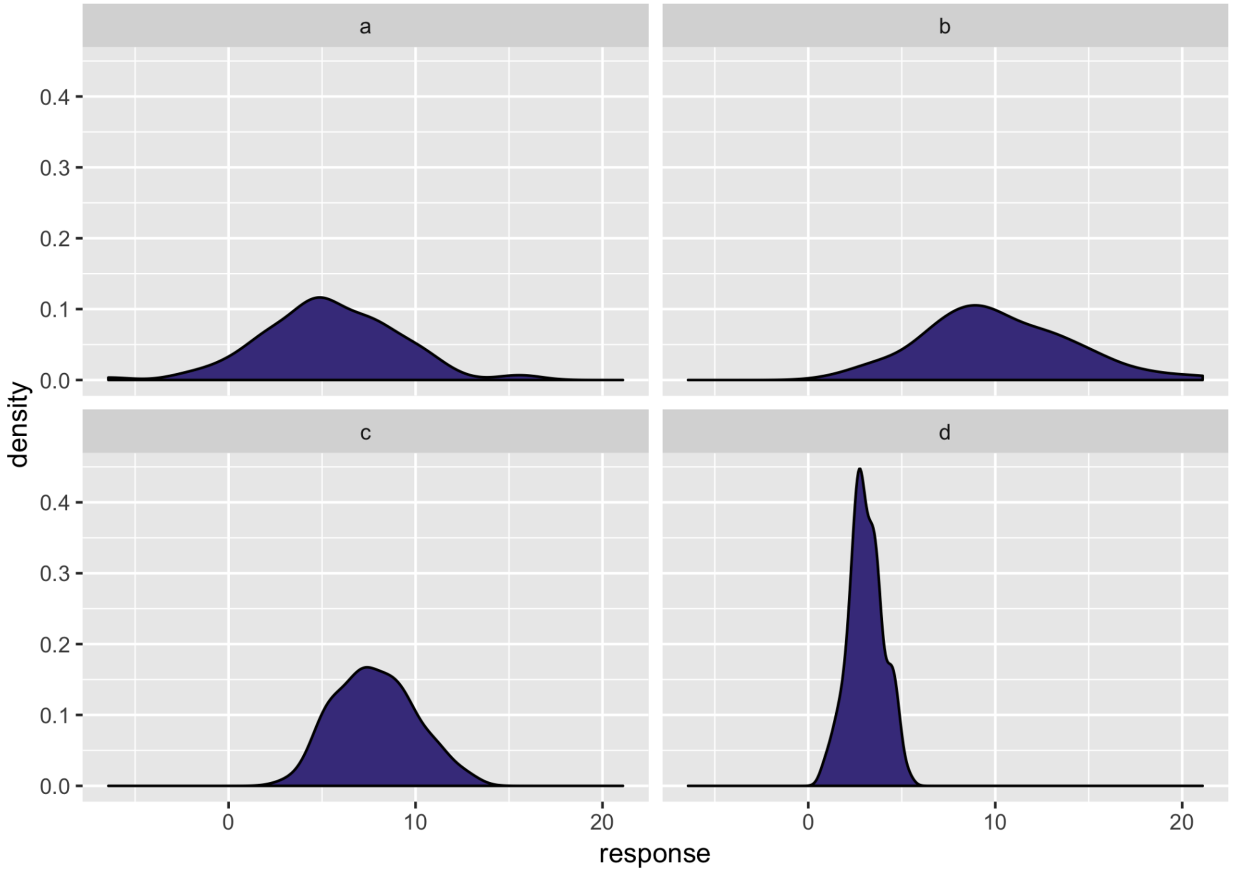

Figure 1.4. *Density plots of pretest math scores for Groups a - d.*

Column (mean) plus error bar (CI). Compared to a box plot and a histogram, a much more concise visualization of a distribution is to display its central tendency (typically the mean) with a bar (or column), and its variability (typically confidence interval or standard deviation) with an error bar. It belongs to the bar or column chart type, but we note it as an individual bar with an error bar can represent a distribution, when the focus is the mean with considerations of the variability (see Figure 1.5 for mean pretest math scores by group). A label of value represented by the bar and a reference line can both assist the understanding of the focused information of such charts. While these graphs provide a diminished view of the variability in the data, they provide a simple but clear portrayal of how the means differ, which may be useful in the context of presenting the results of hypothesis tests about mean differences between the four groups.

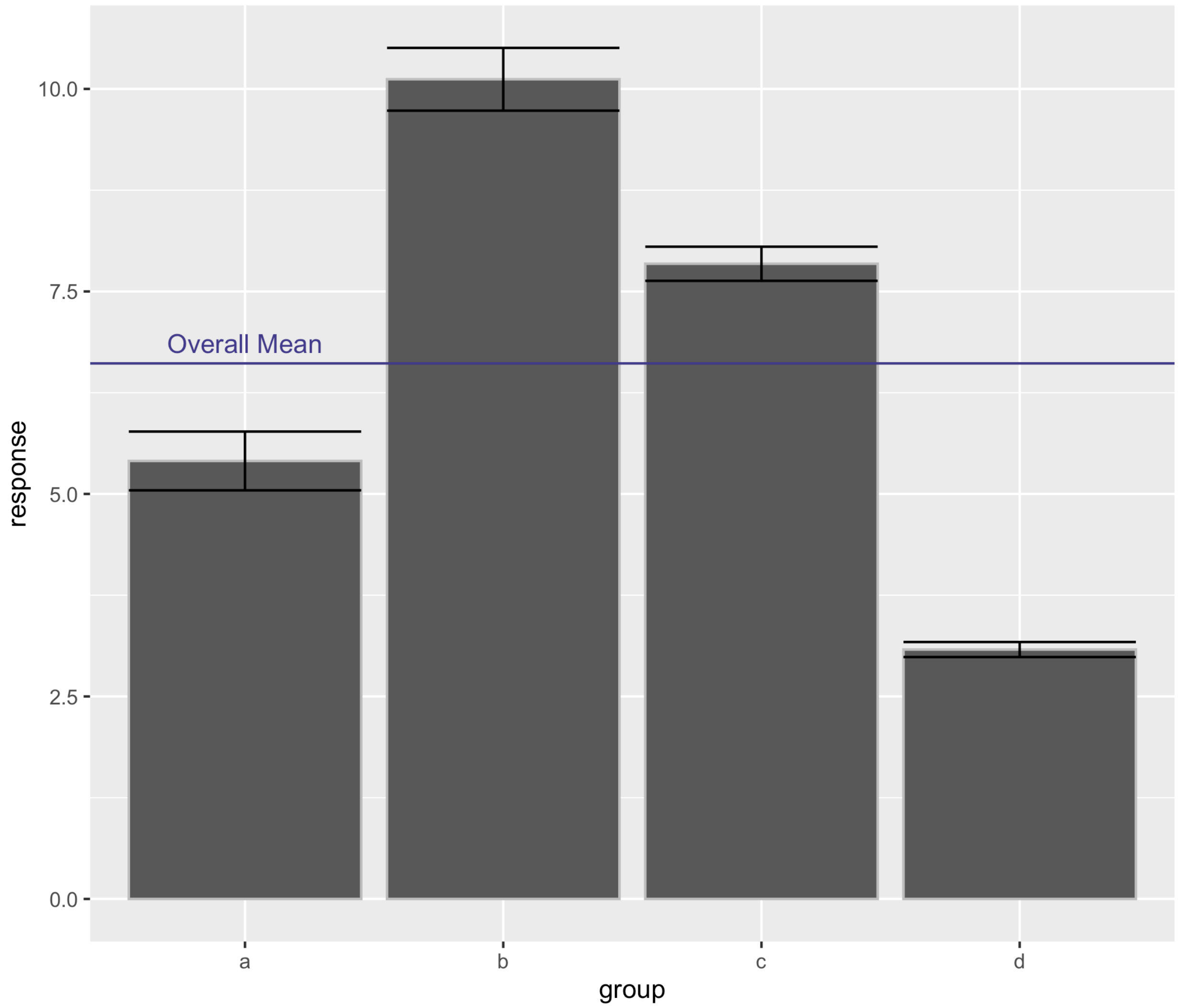

Figure 1.5. *Mean pretest math scores by group with the overall mean as a reference.*

2. Comparing Two or More Groups

The data visualization techniques of one number and one distribution discussed in the previous section are certainly applicable for each of the individual representations in a comparison. Note that what is discussed in this section centers on the comparison techniques—how to do a good job in comparing visually.

Clustered column or bar chart. Clustered (or side-by-side) bar or column charts are a very popular way to make the comparison of numbers (without error bars) or distributions (with error bars). It typically displays scores that are categorized in two ways—by condition (Control vs. Intervention) and by group (a vs. b vs. c vs. d). The only additional feature of this bar chart is that the bars are clustered by one of the two categorical variables to facilitate a particular comparison. For instance, one can cluster (or make side-by-side) the bars of the same group tightly together and place an obvious space between all pairs of two bars (see Figure 2.1) in order to understand that within groups, how are intervention participants performing compared to those in the control condition. In this clustered bar chart, one can easily observe, e.g., in group *b* students in the intervention condition outperforming those in the control condition.

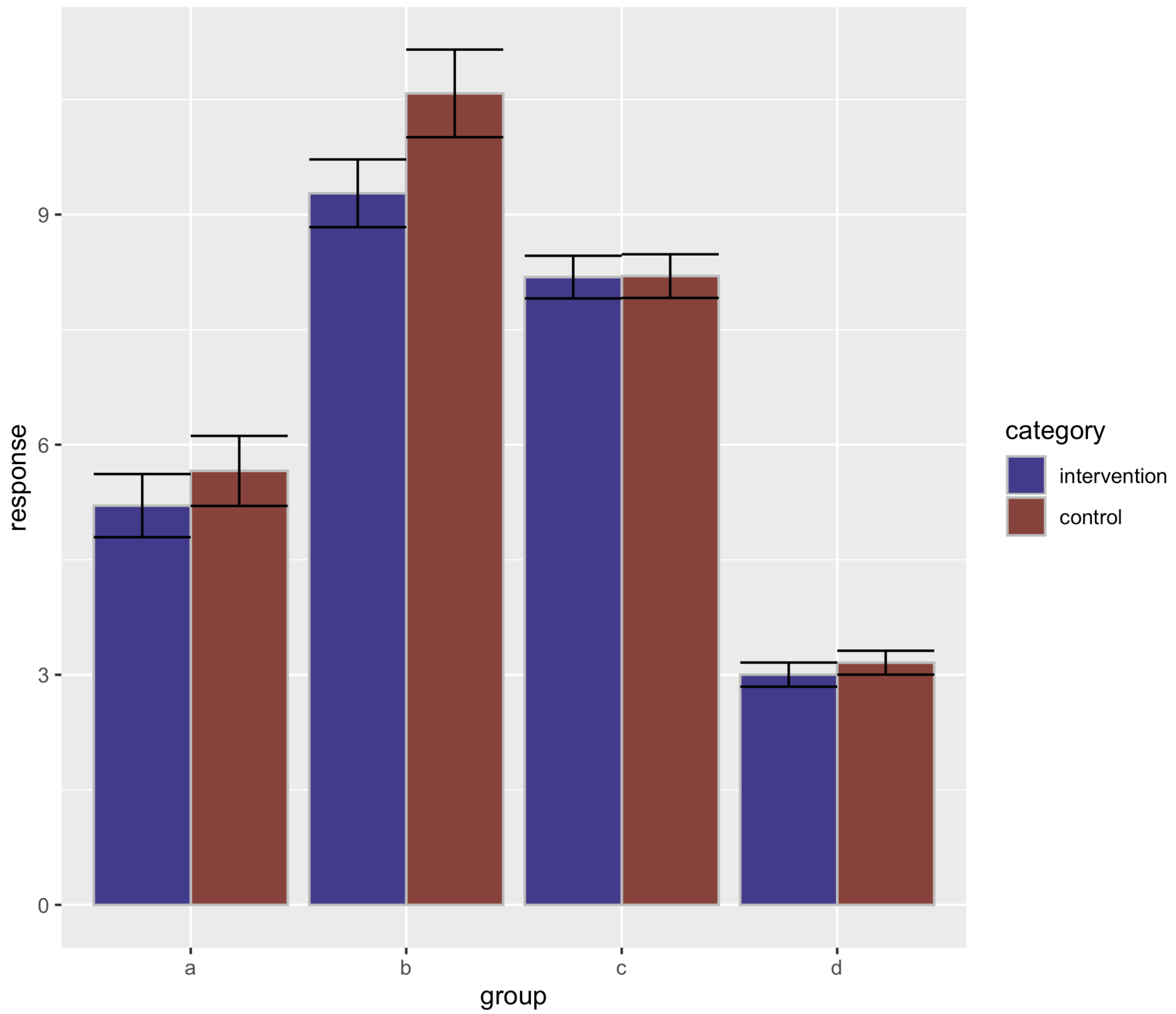

Figure 2.1. *Cluster bar chart demonstrating performance differences between Intervention and Control participants in each group of groups a - d.*

Alternative, if the comparison is more focused on between groups within each condition, then a bar chart where the four group bars are put side-by-side for Intervention and clearly separated from the Control cluster would better serve the comparison. This chart, however, because it has multiple bars (i.e., groups) in one cluster, may pose difficulty for comparing two nonadjacent bars. A small amend to make in this scenario would be to manipulate the order in which the groups appear so that the point one is trying to make easily observed from the chart (Figure 2.2).

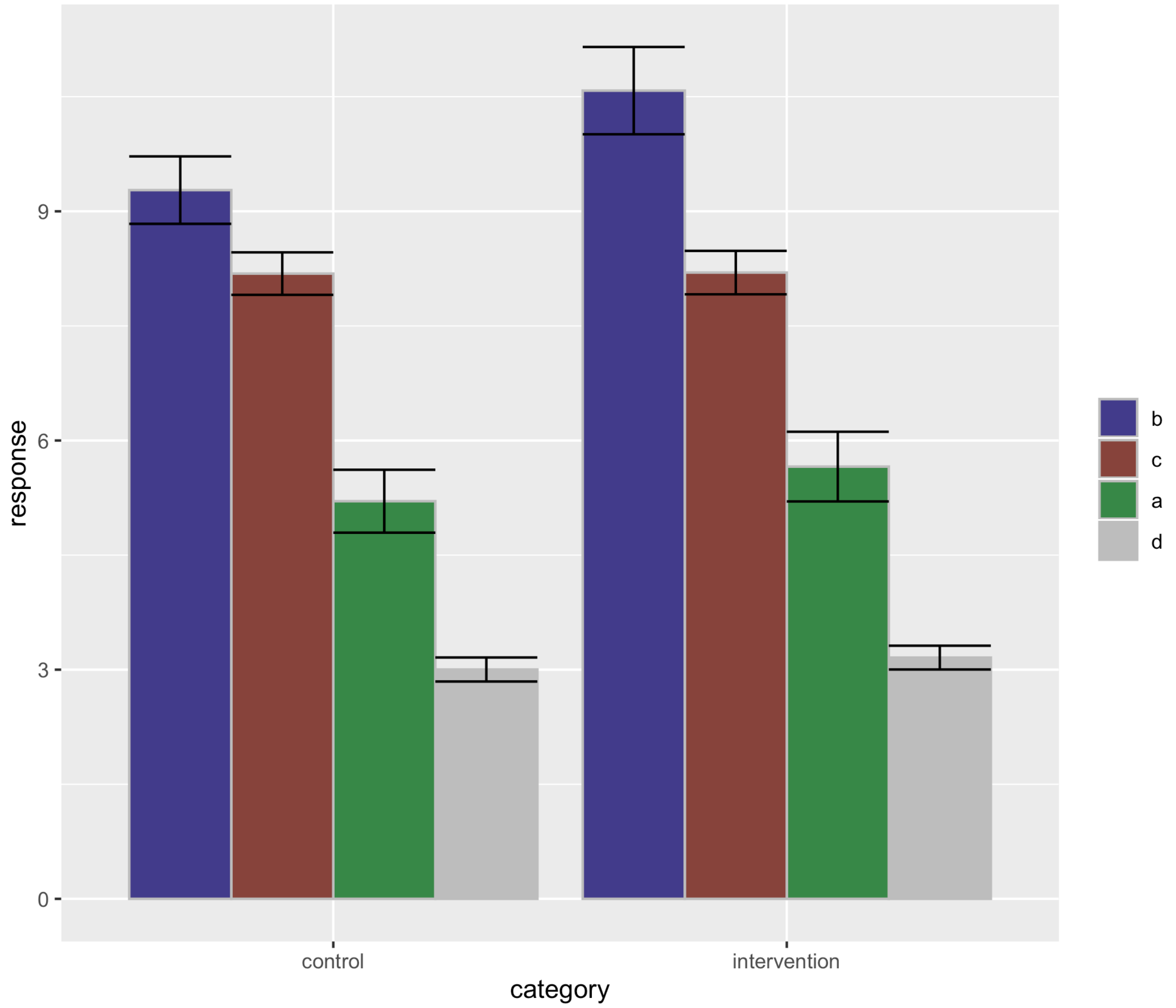

Figure 2.2. *Cluster bar chart demonstrating performance differences among the four groups in the Intervention condition and in the Control condition.*

Small multiples. In order to make comparisons of multiple categories, but the graph gets complicated and makes it unclear what patterns are shown, one solution is to segment the graph into small multiples (e.g., Figures 1.3 & 1.4). Small multiples are an array of charts that share the same axes and scales and are arranged side by side. This way, the user is not relying on the graph legends or the separating spaces to distinguish the categories, but the aligned areas that are scaled to the same.

3. Visualizing Bivariate and Multivariate Relationships

The mutual relationships between variables are crucial when data analysis and representation simultaneously involve two or more variables. Particularly, it is informative to show whether scores on one variable change as scores on another variable increase or decrease and how strong the change-against-change relationship is. Visualizations of bivariate and multivariate relationships are usually drawn prior to working out a bivariate and multivariate analysis to aid the researcher to expect and interpret their analysis results.

Scatterplot. A scatterplot (or *scattergram*) is a two-dimensional coordinate chart that displays the relationship between two quantitative variables, with an observation (or a case) in a data set plotted as a dot in the chart and its X and Y coordinates showing the corresponding values on the two variables (Figure 3.1). There are three key aspects of data shown in a scatterplot, including the strength of the correlation between the two targeted variables, the direction of the correlation, and shape of the correlation, in addition to small aspects, such as outlier cases on either or both of the variables.

Like histograms and box-and-whisker plots, scatterplots are used less frequently in publications or presentations. Scatterplots are graphically complex and may contribute to confusion especially for non-academic audiences (Knaflic, 2015; Rensink & Baldridge, 2010; Cleveland & McGill, 1984), for which reason, perhaps, a scatterplot is not of the greatest value for presenting data. However, scatterplots are arguably the most important graphs for data analysis; it is certainly a necessary first step of your correlation and regression analyses. One other important use of a scatterplot is to examine a number of derived quantities of two targeted variables, such as regression residuals. In the context of our example, here, with the whole sample (i.e., the grouping—a, b, c, and d—or the conditions—Intervention and Control—are not considered), the posttest math scores and the pretest math scores are positively correlated.

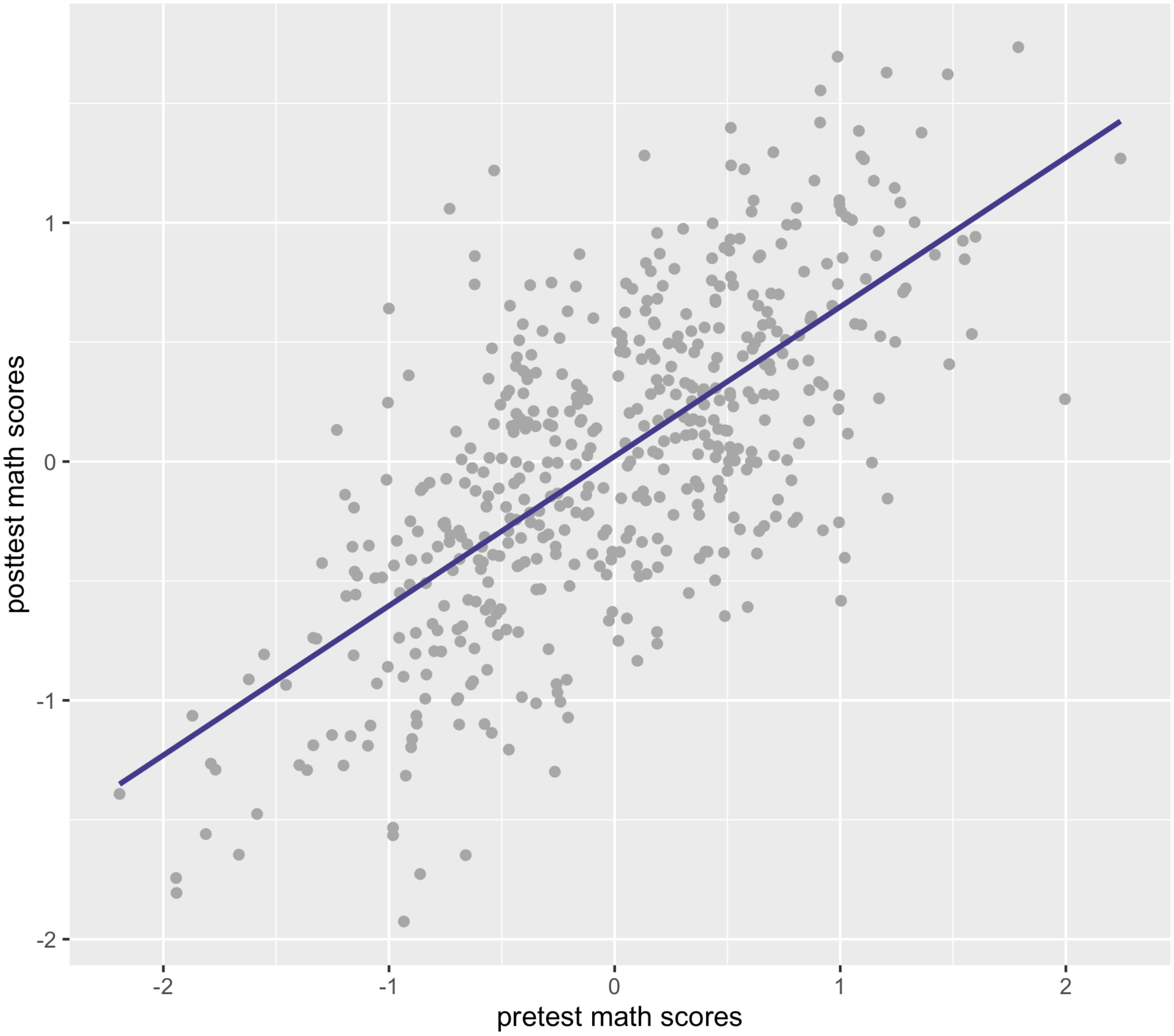

Figure 3.1. *Scatterplot showing the bivariate correlation between pretest math scores and posttest math scores*.

Scatterplot matrix. When one needs to look at several scatterplots, for multivariate analysis purposes such as multiple regression, a scatter plot matrix is a very useful tool. A scatter plot matrix is a table (or a portion of a table) of scatter plots, whose scales are not necessarily the same. Each plot is small so that many plots can be fit on a page and serve a similar purpose of small multiples—provide a visual of the patterns among multiple variables of interest. For instance, the lower-left off-diagonal elements of the matrix in Figure 3.2 shows the scatter plots of three variables—pretest math scores, posttest math scores, and math anxiety, while the diagonal elements show histograms of the variables.

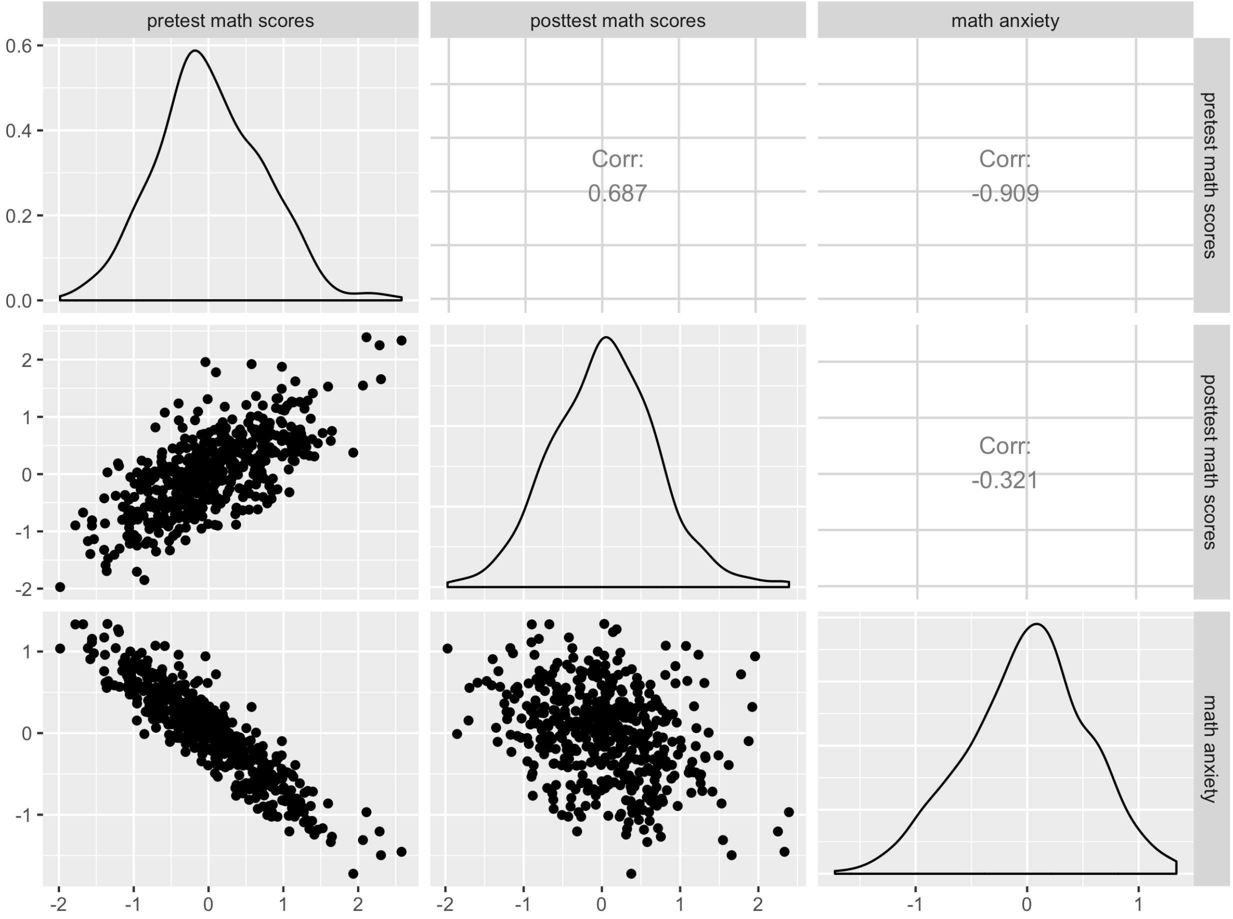

Figure 3.2. *Distributions, bivariate correlations, and Pearson correlation coefficients of pretest math scores, posttest math scores, and math anxiety.*

4. Visualizing Trends with Data over Time

There are three key elements about change over time in which researchers are often most interested—trend of the change, the possibly varying rates of change from two consecutive time points to another two consecutive time points, and exceptions in the trend of change (Few, 2009).

Line chart (with each case being represented). Line charts are employed to display how one variable (or, dependent variable) is affected by another variable (or, independent variable), and thus the trends for data on the dependent variable. In a line graph, the *y*-axis is used to represent the dependent variable of interest, which is typically a continuous variable, and the *x*-axis represents the independent variable, which can either be a discrete or continuous variable. Data points are connected by a line, which approximates the trend of change in the dependent variable. When the independent variable is time, the line graph shows the changes in the quantitative values of this targeted dependent variable over time (Figure 4.1). Line charts are arguably the most widely used graph type to visualize the change of a dependent variable over time. Due to the convention of using lines to connect points in a line chart to approximate trend, it can appear misleading if the markers for the time points of actual data collection are missing. It is highly advisable to include a marker for each point of data collection, especially for line charts where the observed time points were sparse (e.g., fewer than 10) and for those where the intervals of time points are uneven. In the context of our example, if data were collected over time (e.g., at pretest, posttest, and the 3-month follow-up), such a plot could be useful to understand the learning trajectories. Other visualizations could explore different trajectories by group (*a*-*d*).

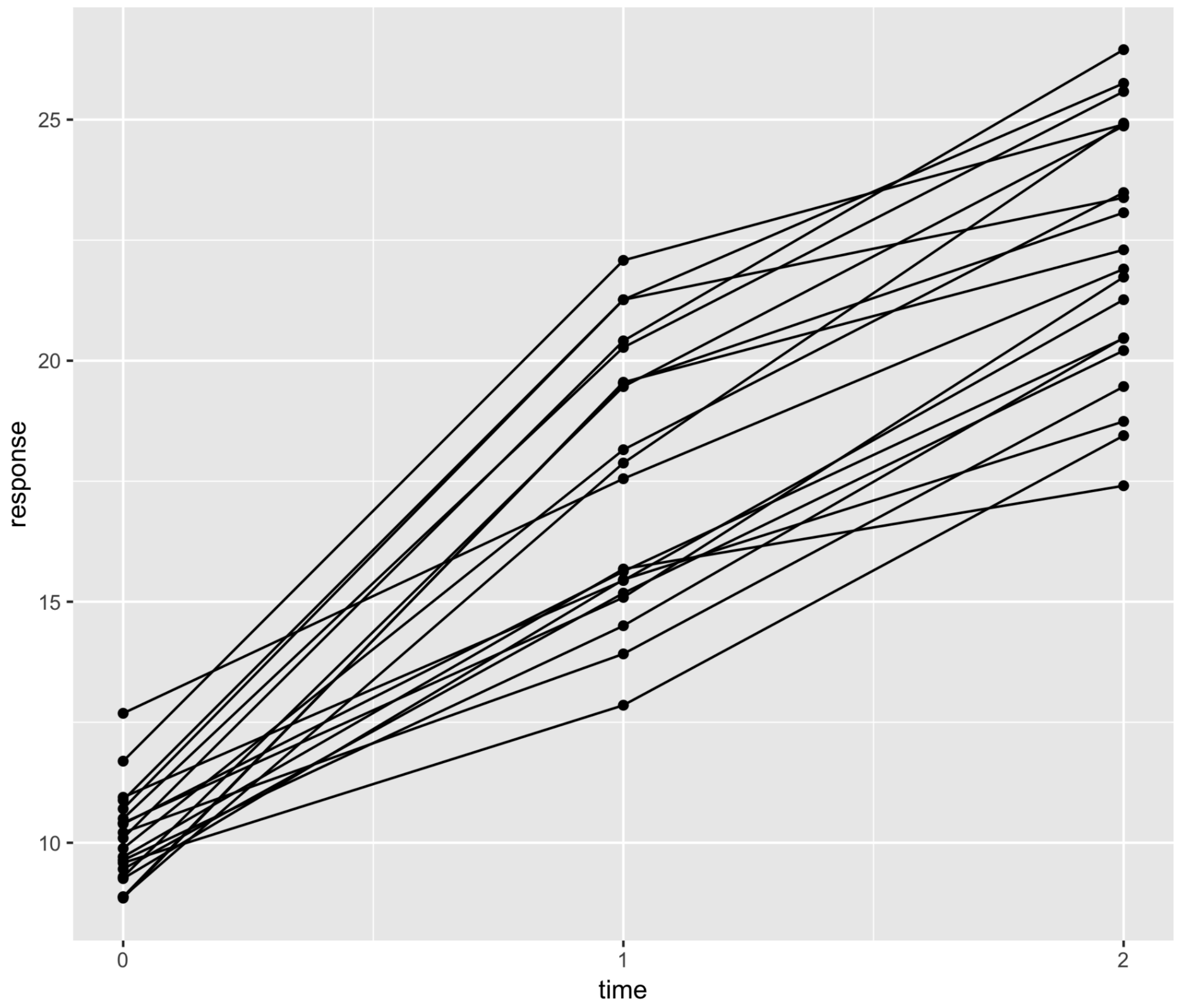

Figure 4.1. *Individual changes in response scores from over three time points—pretest, posttest, and 3-month follow-up test scores*.

Line chart (longitudinal with one line of best fit). Similar to data representations of a single distribution and those of comparing two or more distributions, line charts in the majority of its applications are used for showcasing a point estimate and a margin of error (e.g., mean and confidence interval, respectively) of a sample of individual cases. This requires a line chart to represent changes in not only the central tendency but also the variability over time. When the variability or margin of error is necessary to be represented in a line chart, it is conventional to employ an error bar for each point estimate. It is also recommended to use a band of lighter color distributed around the points that indicate the means of the dependent variable at the various time points (Evergreen, 2016). To be exact, this practice is essentially using lines to connect the upper ends of all error bars and connect the lower ends of all error bars to form a region around the line connecting all means, and to fill in a lighter color to highlight the region that approximates the variability over time. Using the same longitudinal data example presented in Figure 4.1, we present in Figure 4.2 the line surrounded by a colored band, which provides a visualization of a trend of longitudinal change with a margin of uncertainty. This shows to the audience as the mean changes longitudinally (i.e., goes up or down) how the margin of error may also change (i.e., becomes wider or narrower).

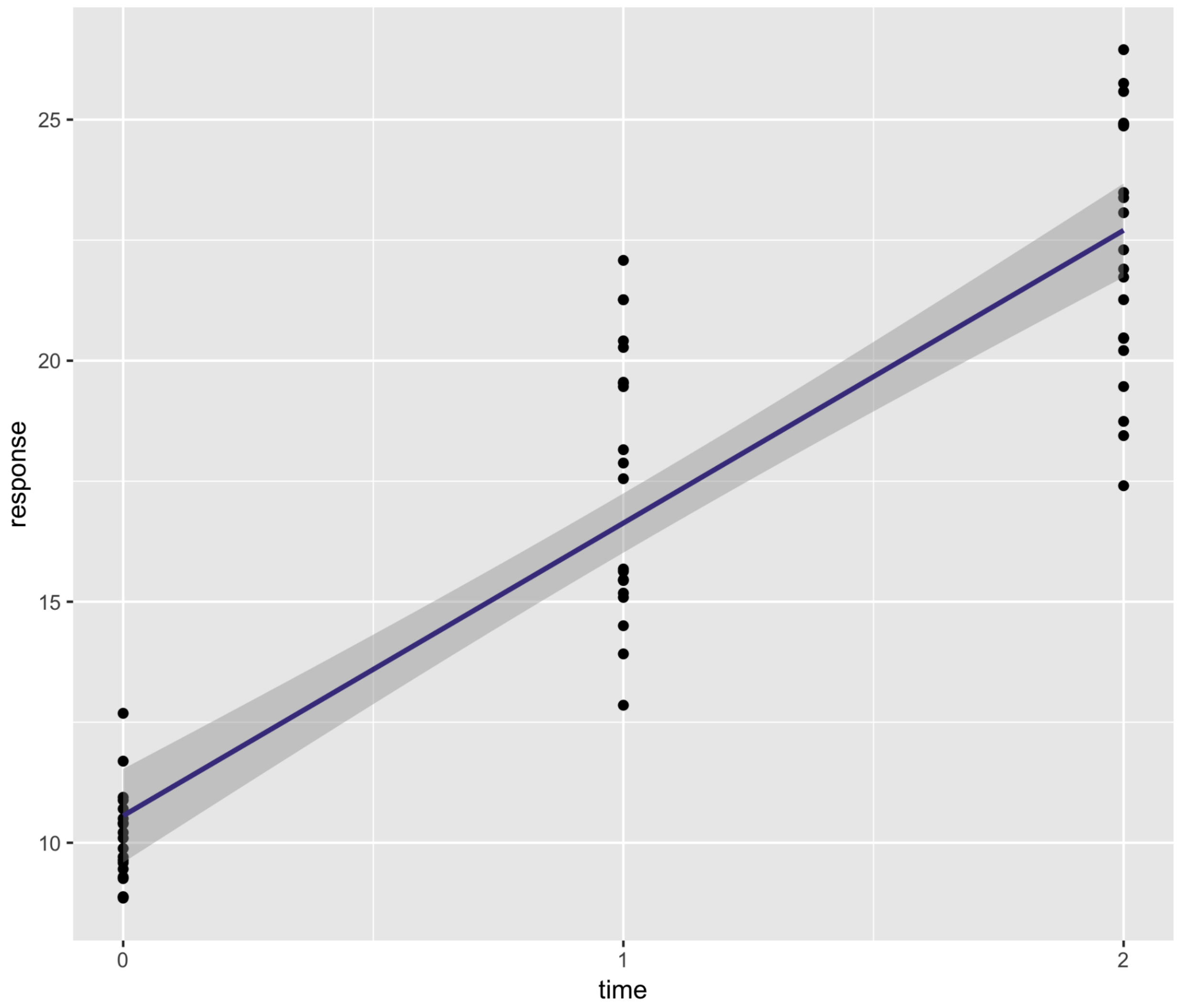

Figure 4.2. *Mean and variability of changes in response scores from over three time points—pretest, posttest, 3-month follow-up.*

5. Other Directions for Data Representation and Visualization

In our work as educational researchers, we often intend to convey a “loud” message with data representations and visualizations. Additionally, this message is meant to be communicated to a broad audience, including policymakers, academics, administrators, teachers, and families, to name a few. To convey a loud message with a single value, our options for data visualization become more focused on representations that have the potential to effectively highlight values and sections of data in familiar and appealing ways. An infographic serves these purposes well.

An information graphics, information visualization or infographic, is defined as “a visualization of data or ideas that try to convey complex information to an audience in a manner that can quickly be consumed and easily understood” (Smiciklas, 2012, p. 3). This type of data representation and visualization has been shown to be an effective communicator in various fields but is itself relatively new in the field of educational research. For instance, those working in fields of data visualization and business find that infographics have the greatest potential for audience memorability (Borkin, Bylinskii, Isola, Sunkavalli, Oliva, & Pfister, 2013; Smickiklas, 2012). In addition, in the field of medicine, researchers have found that infographics are an effective means to disseminate research findings on social media (Ibrahim, Lillemoe, Klingensmith, & Dimick, 2017). Moving forward, this form of data representation and visualization can, and should, be used to communicate educational research findings in a way that is accessible, engaging, memorable, and aesthetically pleasing to a variety of stakeholders (e.g., families, teachers, administrators, policymakers).

Additional examples of statistical graphs that have the potential to communicate “loud” messages in familiar and appealing ways are *icon arrays* and *pie charts and bar charts that emphasize only one section.*

Icon array. An icon array represents a proportion by displaying a shape of choice (i.e., the icon) repeatedly for a number of times—usually a power of 10—to represent the denominator, of which a number of the icons are edited (usually with a different color from the rest of the icons) to represent the numerator. Together, the icons represent the corresponding proportion (Figure 5.1). Research supports the use of icon arrays in effectively overcoming low numeracy (Galesic, Garcia-Retamero, & Gigerenzer, 2009), particularly in terms of reducing denominator neglect (Garcia-Retamero, Galesic, & Gigerenzer, 2010). Such a graph could be helpful in the context of our example, for example, describing the demographics of the students.

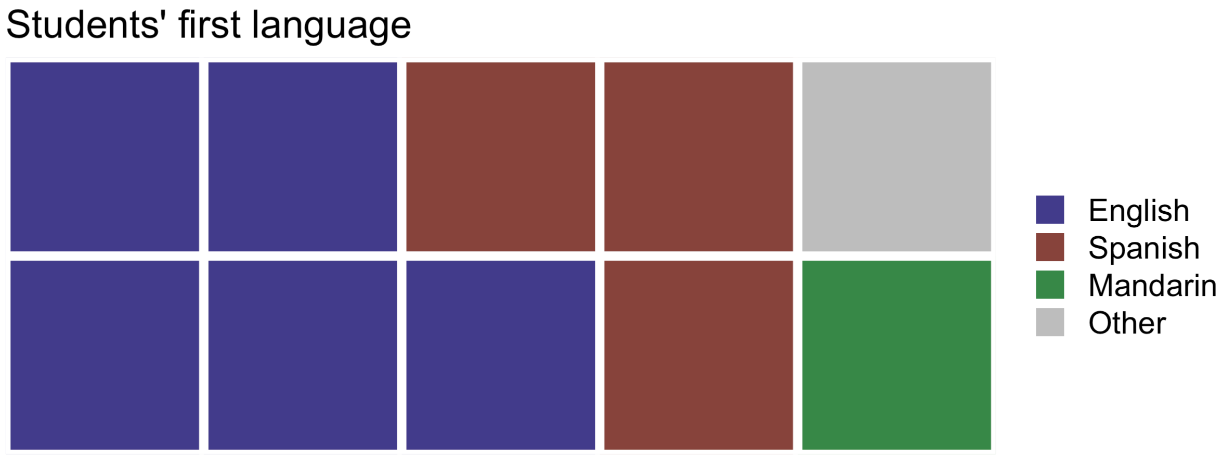

Figure 5.1. *There are 30% of the students who speak Spanish as their first language.*

Pie and Bar Charts that Emphasize One Section. Pie charts are used to illustrate proportions for categorical variables (nominal or ordinal data). An entire pie represents the entirety or 100% of the measured variable, of which sectors or slices of a pie represent various proportions that sum up to 100%. Smaller slices or having many slices in a pie chart, however, create distraction rather than provide clear information (Ben-Tovim, Asnash, & Hoter, 2005). When the target is one slice of the whole pie chart (or one proportion), it is advisable to edit uniquely and provide a text label only for the targeted slice to highlight the proportion and format the rest of the pie in a unified manner (see an example in Figure 5.2). Such a graph could be helpful for simply communicating how students responded to a particular item on, e.g., the math test at pretest.

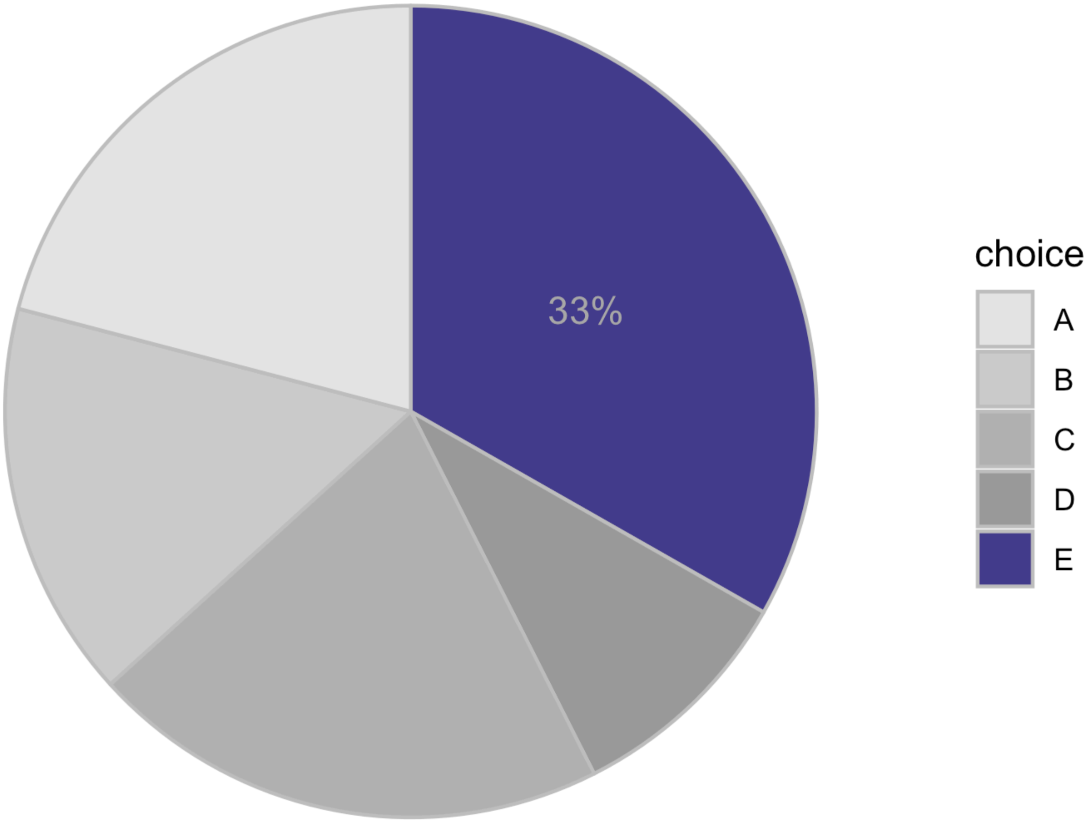

Figure 5.2. *The most popular option is E, as about 33% of the sample chose E as their answer, to Item X on the math test at pretest.*

A bar graph represents data in categories as a series of vertical bars, with the *x*-axis representing the categories contained in the data and the *y*-axis representing the frequency of the measured variable. It is different from a column chart in that column charts display the data horizontally. Bar or column charts are much more widely used than pie charts in a variety of publications, despite its lack of natural anchors to facilitate estimation of proportions (Spence, 2005). A good way to use a bar or column graph for a single number—frequency, to be exact—would be to edit the corresponding bar differently from the rest of the bars (e.g., with a different color) and only label it with text. The rest of the bars should be formatted in a unified manner (see an example in Figure 5.3). As above, in the context of our example, such a visualization could be helpful for communicating how students respond to a particular item on, e.g., the math test at posttest.

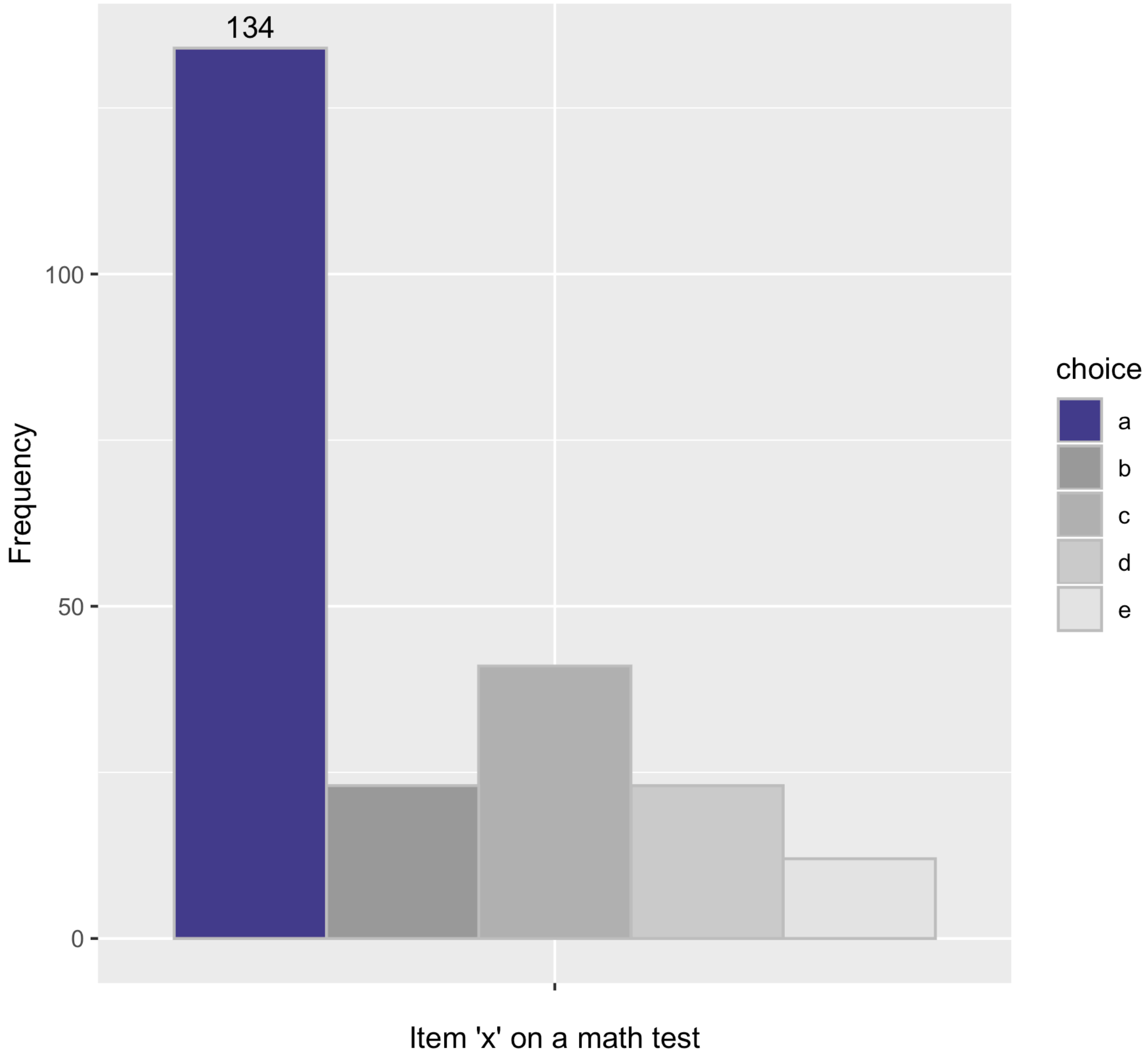

Figure 5.3. *An overwhelming number of 134 students selected Option A—the most prevalent—at the posttest.*

Conclusion

As we have sought to describe in this article, when making decisions about using data representations and visualizations, one must consider that the data visualization *makes a point* to *an audience that promotes the underlying statistical argument using data and accurately represents the data as it is*. Looking in the future, the next foray into this field of data representation and visualization is making decisions between the use of traditional statistical graphics and infographics (Dur, 2014). While traditional statistical graphics make up the bulk of this entry, the introduction of infographics into the field has forced many to consider the tradeoffs between using each visualization. While infographics may not seem ineffective in terms of communicating the underlying statistical messages, to some, they seem accessible and good at communicating a main point. Gelman and Ulwin (2013) summarize this debate asking the analysts to consider how visualizations can be *both* effective regarding communicating a point effectively while being true to the data, a point we hope that our summaries of theory and past research and examples demonstrate: “Our main practical suggestion [is] that, in the internet age, we should not have to choose between attractive graphs and informational graphs.”

References

Borkin, M., Vo, A., Bylinskii, Z. Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2306-15.

Cleveland, W. S., & McGill, R. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association*, *79*(388), 807-822.

Dur, B. I. U. (2014). Data visualization and infographics in visual communication design education at the age of information. *Journal of Arts and Humanities*, *3*(5), 39-50.

Evergreen, S. D. (2016). *Effective data visualization: The right chart for the right data*. Sage Publications.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Oakland, CA: Analytics Press.

Friendly, M. (2008). A brief history of data visualization. In *Handbook of data visualization* (pp. 15-56). Springer, Berlin, Heidelberg.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology, 28*(2), 210-16.

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making, 30*(6), 672-84.

Gelman, A., & Unwin, A. (2013). Infovis and statistical graphics: Different goals, different looks. *Journal of Computational and Graphical Statistics, 22*(1), 2-28.

Gravetter, F., & Wallnau, L. (2013). *Statistics for the Behavioral Sciences* (9th ed.). Belmont, CA: Wadsworth Publishing.

Healy, K. (2018). *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press.

Ibrahim, A., Lillemoe, K., Klingensmith, M., & Dimick, J. (2017). Visual abstracts to disseminate research on social media: A prospective, case-control crossover study. *Annals of Surgery, 266*(6), 46-8.

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons.

Rensink, R. A., & Baldridge, G. (2010, June). The perception of correlation in scatterplots. In *Computer Graphics Forum* (Vol. 29, No. 3, pp. 1203-1210). Oxford, UK: Blackwell Publishing Ltd.

Smiciklas, M. (2012). *The Power of Infographics: Using pictures to communicate and connect with your audience*. Indianapolis, IN: Que Publishing.

Spence, I. (2005). No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics, 30*(4), 353-68.

Tabachnick, B., & Fidell, L. (2013). *Using Multivariate Statistics* (6th ed.). Boston, MA: Pearson.

Tufte, E. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary Reviews: Computational Sciences, 3*(2), 180-5.

Wickham, H. (2013). Graphical criticism: some historical notes. *Journal of Computational and Graphical Statistics, 22*(1), 38-44.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wilkinson, L. (2005). *The Grammar of Graphics* (2nd ed.). Canada: Springer.

R Code for Each Data Representation

Note that all of the functions below require the ggplot2 and dplyr libraries to be loaded first with the following commands;

```
library(ggplot2)

library(dplyr)
```

If not installed, they must be installed (only once) with the following command:

```
install.packages(c(“ggplot2”, “dplyr”))
```

The different types of graphs were created with different, simulated data sets. The code used to simulate those (and to assign them to the same object as is used for the plots below, i.e., `d`, `dTime`, `dd`, `df`, and `dff`) is available here: https://osf.io/3phcz/

Note that one feature of *ggplot2* is that while there is a general framework for creating visualizations (and specific geometric objects, i.e., those for a boxplot) but not general code to create plots. As a result, the code below is as much (or even more) code to create these specific graphs (with their specific colors, themes, and so on) as we created them as it is code to create these *types of graphs*.

Type of Graph	**Figure in Article**	**Code for R to Recreate Figure**

Column/Bar Chart Error Bars	1.5	`d1 %>%`
		`  group_by(group) %>%`
		`  summarize(response_mean = mean(response)) %>%`

Type of Graph	**Figure in Article**	**Code for R to Recreate Figure**
		ggplot(aes(x = group, y = response_mean)) + geom_col()
Box Plot	1.1	d1 %>% ggplot(aes(x = group, y = response)) + geom_boxplot()
Clustered Column/Bar Graph	2.1	d2 %>% ggplot(aes(x = group, y = response, fill = category)) + stat_summary(fun.y = mean, geom = "bar", color = "gray", position = position_dodge()) + stat_summary(fun.data = mean_se, geom = "errorbar", position = position_dodge()) + scale_x_discrete(labels = toupper(letters[1:4])) + scale_fill_manual(values = c("slateblue4", "#8b473c"))
Column/Bar (mean) With Error Bars (CI)	1.5	d1 %>% ggplot(aes(x = group, y = response)) + stat_summary(fun.y = mean, geom = "bar", color = "gray") + stat_summary(fun.data = mean_se, geom = "errorbar", color = "black") + scale_x_discrete(labels = toupper(letters[1:4])) + geom_hline(yintercept = mean(d$response), color = "slateblue4", linetype = 2) + annotate("text", y = mean(d$response) + .25, x = 1, label = "Overall Mean", color = "slateblue4")
Density plot	1.4	d1 %>% ggplot(aes(x = response)) + geom_density(color = "slateblue4", fill = "slateblue4") + facet_wrap(~group)

Type of Graph	**Figure in Article**	**Code for R to Recreate Figure**

Histogram	1.3	`d1 %>%`
		`  ggplot(aes(x = response)) +`
		`  geom_histogram() +`
		`  facet_wrap(~group)`

Line Chart (longitudinal with each case being represented)	4.1	`d2 %>%`
		`  ggplot(aes(x = period, y = Y, group = id)) +`
		`  geom_point() +`
		`  geom_line() +`
		`  scale_x_continuous(breaks = c(0, 1, 2)) +`
		`  xlab("time") +`
		`  ylab("response")`

Line Chart (longitudinal with one line of best fit)	4.2	`d2 %>%`
		`  ggplot(aes(x = period, y = Y)) +`
		`  geom_point() +`
		`  geom_smooth(method = "lm") +`
		`  scale_x_continuous(breaks = c(0, 1, 2)) +`
		`  xlab("time") +`
		`  ylab("response")`

Scatterplot with a Line of Best Fit (fitted linear model)	3.1	`d4 %>%`
		`  ggplot(aes(x = x1, y = x2)) +`
		`  geom_point() +`
		`  geom_smooth(method = "lm", se = FALSE) +`
		`  ylab("response1") +`
		`  xlab("response2")`

Scatterplot Matrix	3.2	`library(GGally) # must install first with install.packages(“GGally”)`
		
		`dd %>%`
		`  select(response1 = x1, response2 = x2, response3 = x3) %>%`
		`  ggpairs()`

Type of Graph	**Figure in Article**	**Code for R to Recreate Figure**

Waffle Plot (Icon Array)	5.1	`parts <- c(Science = 50, Math = 20, Humanities =`
		`30)`
		`waffle(parts / 10, rows=2,`
		`       title = "Courses taken by students") +`
		`  theme(text = element_text(size = 18))`

Column/Bar With One Section Emphasized	5.3	`d5 %>%`
		`  ggplot(aes(x = "",  y = val, fill = choice)) +`
		`geom_bar(stat="identity", width = 1, position =`
		`"dodge"+`
		`    geom_text(label = rev(c(134, rep("", 4))),`
		`position = position_dodge(1), vjust = -.5)`

Pie Chart With One Section Emphasized	5.2	`df %>%`
		`  ggplot(aes(x= "", y = val, fill = choice)) +`
		`  geom_bar(stat = "identity", width = 1) +`
		`  coord_polar("y", start = 0) +`
		`  geom_text(aes(label = c(rep("", 4), "33%")),`
		`position = position_stack(vjust = .5)) +`
		`  labs(x = NULL, y = NULL, fill = NULL) +`
		`  theme_classic() +`
		`  theme(axis.line = element_blank(),`
		`        axis.text = element_blank(),`
		`        axis.ticks = element_blank(),`
		`        plot.title = element_text(hjust = 0.5))`

Note. We have removed elements of the theming with respect to the colors used in plots as well as axis labels in order to provide the basis for representations that are more generally useful (and not useful only for recreating those included in this article).