ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Testing Statistical Charts: What Makes a Good Graph?

Susan Vanderplas,[1] Dianne Cook,[2] and Heike Hofmann[3]

[1]Department of Statistics, University of Nebraska, Lincoln, Nebraska 68583, USA; email: susan.vanderplas@unl.edu

[2]Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia

[3]Department of Statistics, Iowa State University, Ames, Iowa 50011, USA

Annu. Rev. Stat. Appl. 2020. 7:61–88

First published as a Review in Advance on January 3, 2020

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041252

Keywords

graphics, visualization, user testing, visual inference, perception, chart design

Abstract

It has been approximately 100 years since the very first formal experimental evaluations of statistical charts were conducted. In that time, technological changes have impacted both our charts and our testing methods, resulting in a dizzying array of charts, many different taxonomies to classify graphics, and several different philosophical approaches to testing the efficacy of charts and graphs experimentally. Once rare, charts and graphical displays are now everywhere—but do they help us understand? In this article we review the history of graphical testing across disciplines, discuss different direct approaches to testing graphics, and contrast direct tests with visual inference, which requires that the viewer determine both the question and the answer. Examining the past 100 years of graphical testing, we summarize best practices for creating effective graphics and discuss what the future holds for graphics and empirical testing of interactive statistical visualizations.

1. INTRODUCTION

Any survey of literature on statistical charts and graphs will be complicated by the fragmentation of the literature, which occurs because any discipline that uses charts to summarize information generally also has individuals who research and assess the utility of these graphics. A quick survey reveals publications in journals from computer science, psychology, marketing and business, economics, ergonomics and human factors, statistics, sociology, communication and rhetoric, engineering, instructional design, and education. A review of early surveys of graphical forms suggests that this problem has long plagued the study of graphics, as Funkhouser (1937) discusses the broad disciplines affected by statistical graphics and Kruskal (1977) addresses the difficulty of a comprehensive review of the relevant literature.

Historically, the development of graphs and charts has been linked to the development of coordinate systems (Funkhouser 1937, Beniger & Robyn 1978, Fienberg 1979) and abstract representations of data. Preceding the development of formal mathematical coordinates, however, humanity has been representing information in abstract visual form since the origins of civilization; for example, spatial information using maps (Smith 1996) is displayed with varying degrees of abstraction. **Figure 1** shows the oldest known world map, from sixth-century-BCE Babylon.

Exploring visual abstractions of data was one of the pursuits of polymaths who developed the foundations of current scientific pursuits while grappling with increasing quantities of economic and demographic data and measurements of the natural world. During the eighteenth and

Figure 1

Imago Mundi Babylonian map, which is the oldest known world map (sixth century BCE). The representation of the world is relatively abstract. The world is shown as a disc, with Babylon represented by a rectangle at the right end of the Euphrates river, which flows south to the border of the disc. Several other population centers are marked with small circles (Br. Mus. 1882). Image reproduced courtesy of the British Museum, released under a CC By-NC-SA 4.0 license.

nineteenth centuries, governments became involved, assembling data and graphical representations into the statistical atlases and government-issued reports, as created by Harms (1991), Playfair (1801), Walker (1874), and many other contemporaries. In the twentieth century, corporations began using charts and graphs to understand their inner workings—studies of the use of charts and graphs at AT&T (Chandar et al. 2012) and DuPont (Yates 1985) show efforts to standardize and formalize the use of graphics in decision-making at both companies.

As new charts were invented to represent data differently and highlight features of data (Bernstein & Cowden 1937, Yates 1985, McDonald 2014), discussions about the use of statistical graphics began to appear in the literature (Peuchet & Gilbert 1805, Brinton 1917, Karsten 1923), including the relative strengths and weaknesses of various types of charts. In most cases, the drive to produce a classification system for charts and graphs or a system of recommendations for presenting charts and graphs was based on heuristics and largely unsupported by experimentation (Kruskal 1977, MacDonald-Ross 1977). Many of these ad hoc classification systems could not accommodate the large numbers of new plot types being developed.

Calls for experimental validation of the perception and utility of statistical charts were heeded, though at first the experiments were fraught with methodological issues (Croxton & Stryker 1927). Much of the early experimentation regarding the accuracy of graphical forms was based in psychophysics research (Teghtsoonian 1965) on the perception of size and shape. Eventually experiments became more naturalistic: Cognitive psychologists and statisticians began testing different types of graphics, identifying types of perceptual errors associated with different plots (Cleveland & McGill 1985, Spence 1990). In most cases, this testing was limited to simply reading information from the charts, using accuracy or response time measurement. More recently, other methods for examining statistical charts have been developed, including the lineup protocol (Wickham et al. 2010). Even with these developments, the aim of most experimental research in statistical graphics focuses on the initial perception and graph comprehension. Very little work has been done to understand the effect of charts and graphs on higher cognitive processes such as learning or analysis (Green & Fisher 2011).

1.1. Design of Statistical Graphics

Charts and graphs are used for many purposes (Tukey 1972, Fienberg 1979): to summarize data; for analysis, exploration, and discovery; for diagnosis of statistical relationships; to make a rhetorical argument; or even as a substitute for tables. The initial heyday of graphic design was enabled by color lithography and used charts and graphs to tell stories about nations and events (Kostelnick 2016), but in the first half of the twentieth century, graphics were regularly used for mundane purposes as well, such as supporting business decisions (Yates 1985, Chandar et al. 2012) and communicating weather forecasts. As technology has developed, allowing charts to be created quickly for exploratory purposes, the gaps between graphics for presentation, entertainment, and analysis have widened. The different purposes motivating the creation of the chart influence its form and complexity, and the intended audience and reach of the chart are also important considerations.

It is useful to consider a continuum from utilitarianism to artistry, where purely utilitarian charts are, as advocated by Tufte (1991), devoid of "chartjunk" or any decoration, and purely artistic charts trade accurate representation of the data for visually compelling renderings. The distinction between infographics and statistical graphics is primarily one of intent: The infographic, often composed of many small, simple plots interspersed with pictures and text, is designed to attract attention and tell a story. In contrast, the statistical chart is designed to effectively and accurately show the data, potentially with accompanying statistical model information—any visual enhancements should contribute to that aim (Gelman & Unwin 2013, Wickham 2013).

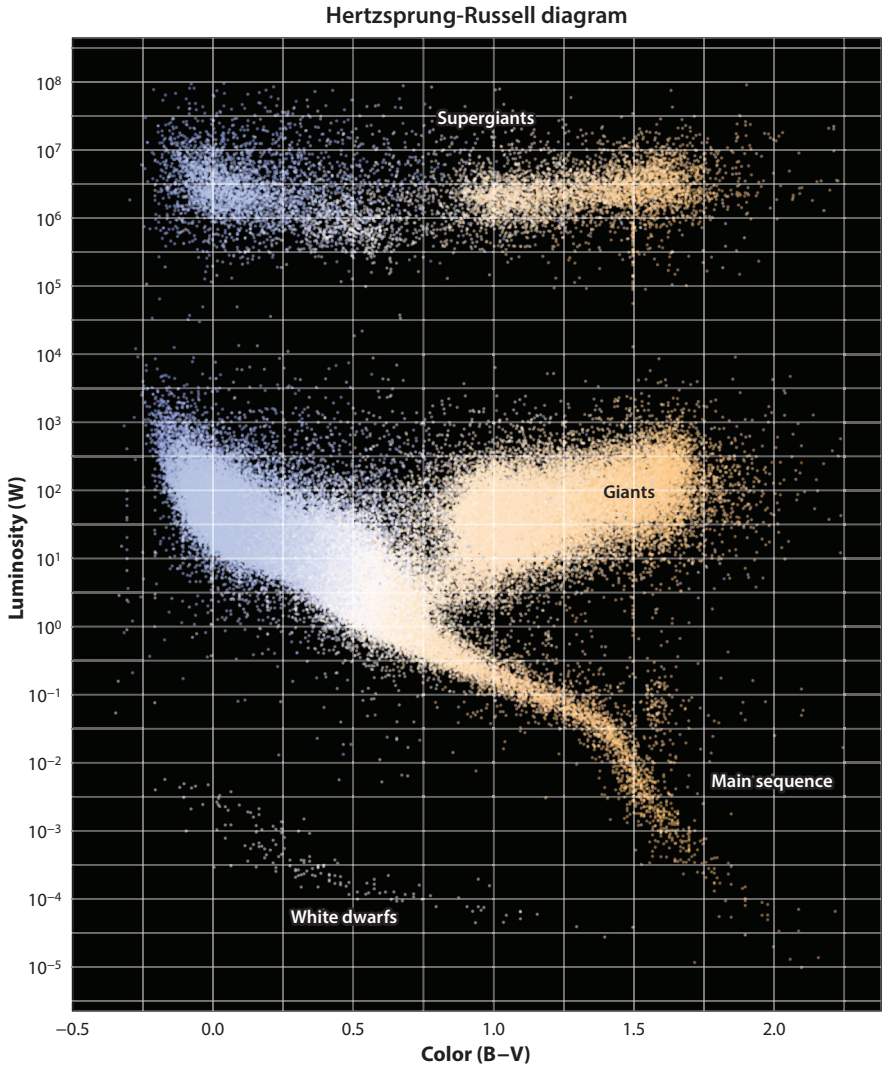


Figure 2

The Hertzsprung-Russell diagram (devised circa 1910) allowed astrophysicists to make the connection between a star's temperature, visual color, and size, enabling a better understanding of the life cycle of a star. B and V are two different light spectra (blue and visual/yellow, respectively), so the difference of the magnitude measured using two adjacent light spectra is an indicator of the star's color, which corresponds to its temperature.

Figures 2, **3**, and **4** show plots designed for different statistical purposes. The Hertzsprung–Russell diagram (**Figure 2**) allowed astrophysicists to make a convincing argument about the life cycle of a star based on its temperature, color, and size. The hurricane forecast map shown in **Figure 3** is designed to communicate urgent information and facilitate decision-making by the US National Hurricane Center; it shows the prediction for Hurricane Michael, which hit the Florida panhandle in October 2018. It shows forecasts of the hurricane's path, wind field, strength, and coastal warnings. The map is primarily utilitarian, with additional decorations such as ocean color

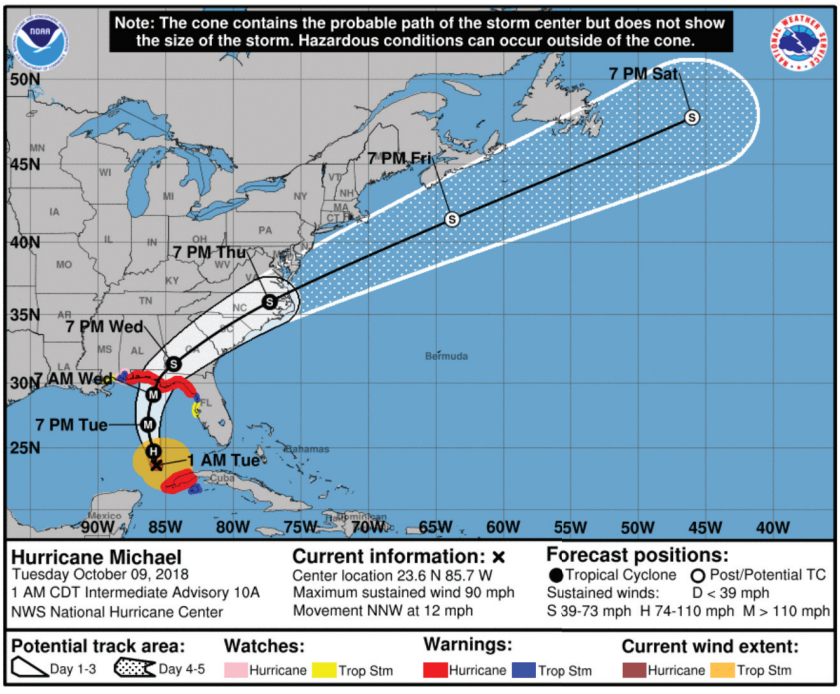


Figure 3

Hurricane forecast map issued on October 9, 2018, shortly before Hurricane Michael made landfall in the Florida panhandle. The predicted hurricane path, wind field, size, and coastal weather status allow viewers to make educated decisions about evacuation orders. Image reproduced with permission from the National Weather Service Historical Archive (**https://www.nhc.noaa.gov/archive/2018/MICHAEL_graphics.php?product=5day_cone_with_line_and_wind**).

and state boundaries for geographic context. The birthday chart in **Figure 4** shows the average number of births in the United States on each day of the year, organized by month. **Figure 4** is a static reproduction of an interactive plot used as part of a web page telling a story about birth dates. It is simple in form and almost infographic in style. By engaging with the interactive chart, a reader could easily deduce that there are relatively few babies born on major holidays, such as Memorial Day, Independence Day, and Christmas; there are also considerably more babies born in the summer months than in the winter.

1.2. Statistical Mapping Using a Grammar of Graphics

In parallel with efforts to understand the perception of charts, there have been many attempts to develop systems for classifying graphics, including those of Bertin & Berg (1983), Desnoyers (2011), and Wilkinson (1999). Systems that attempt to categorize charts based on their geometric representations generally make no effort to include all types of graphics, and they have difficulty accommodating charts that may fall into two or more categories. The classifications of graphics based on the underlying components and their relationships, as in the grammar of graphics developed by Wilkinson, are more robust; they also provide an elegant framework for comparing different types of graphical representations separate from the underlying data structure. An

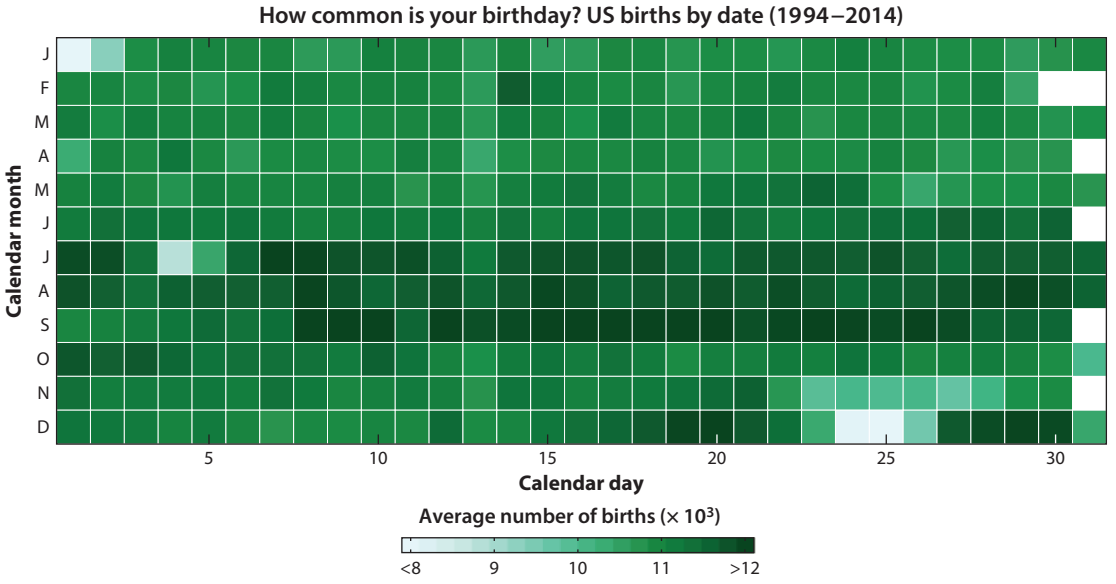


Figure 4

A static reproduction of a birthday chart, which shows the average number of births on a yearly basis in the United States, by month and day of the month. The interactive version, available at **http://thedailyviz.com/2016/09/17/how-common-is-your-birthday-dailyviz/**, is designed to be a storyteller that encourages the viewer to interact, check the estimated conception dates, and read the frequency. Data are from the Centers for Disease Control and Prevention National Center for Health Statistics (1994–2003) and the US Social Security Administration (2000–2014), as reproduced at **https://github.com/fivethirtyeight/data/tree/master/births**.

analogy to conceptualize the difference is that the former is like treating plots like creatures in a zoo, with a unique name for each, while the latter is analogous to having a phylogeny based on genetic data showing how plots are related.

Figure 5 shows the framework of the grammar of graphics, where the data are filtered, variables are mapped, transformations are specified, and then finally, transformed data are mapped to plot aesthetics and coordinate system specifications to produce an abstract visual representation of the data. Full or partial implementations of the grammar of graphics are available for most common scientific computing languages, for example, `ggplot2` in R (Wickham 2010); `plotnine`, a Python implementation of `ggplot2` (Kibirige 2017); and `Gramm` in Matlab (Morel 2018).

The grammar of graphics also enables data plots to be considered to be statistics (Majumder et al. 2013). A statistic is a functional mapping of a variable or set of variables. With tidy data, that is, data where each variable is in its own column, each observation is in its own row, and each value is in its own cell (Wickham & Grolemund 2017), the grammar of graphics creates visual statistics. Variables, as columns in the data table, are mapped to graphical elements, such as the *x*- or *y*-axis, or to color, shape, or even facet, using the grammar. The data plot can then be treated like other statistics: by imagining what the plot might look like in the absence of any structure, we can use the plot of the actual observed data to test for the likelihood of any perceived structure being significant.

Using the grammar of graphics, it is easy for experimenters to compare different types of charts using the same data, as the underlying structure of the graph remains the same. **Figure 6** shows three plots created using the same data and different geometric objects, using the `ggplot2` code to create the plots. Comparing these graphics experimentally would be reasonably simple as the grammar of graphics helps to control the extraneous variables introduced by utilizing different plot

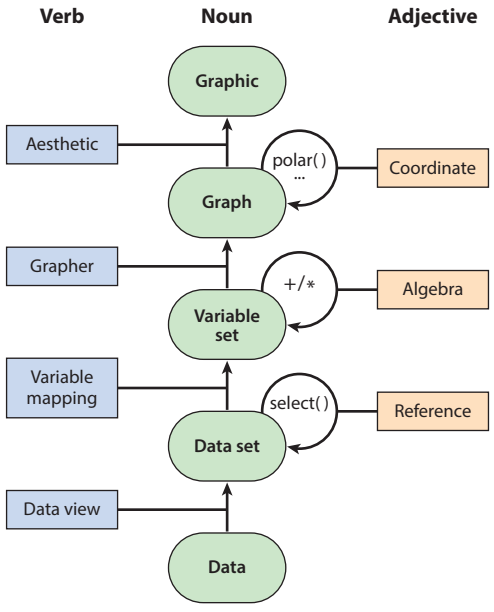


Figure 5

From data to graphic: a schematic representation of the steps required to create a graphic from a data set, which can be used to specify the variable mapping, data transformations, coordinate system, and aesthetic features independently. Adapted with permission from Wilkinson (1999, figure 2.1).

types. In addition, this approach to transformations and scales allows experimenters to easily test judgments made utilizing different axis transformations and color scales to compare perceptual accuracy (Hofmann et al. 2012, Vanderplas & Hofmann 2017).

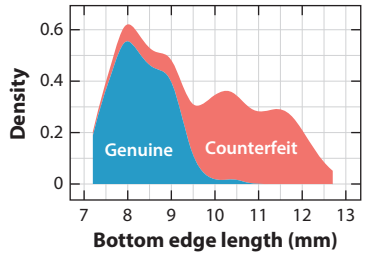
2. TESTING METHODS

In this section, we distinguish between explicitly structured graphical tests, which require the participants to answer specific questions about the graphical objects under experimentation, and implicitly structured tests, where the participant must infer the questions of interest from the provided stimuli.

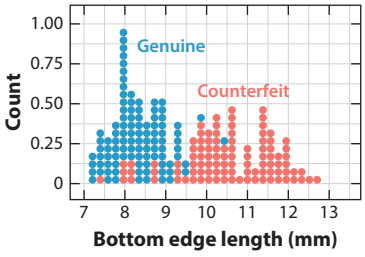
2.1. Explicitly Structured Graphical Tests

During the perceptual process, information from a visual scene is processed by the brain, with information extracted at many different levels of the cognitive process. Preattentive perceptual effects are those that do not require sustained cognitive attention; they are processed automatically within the first 500 ms of viewing a chart or graph. Components processed preattentively include color and shape, as well as some basic information about coarse relationships between individual components. After the preattentive stage, attention is necessary for subsequent processing; this directed attention scaffolds relationships between components and helps us interpret the chart or graph in context. Most of the insights we gain from charts and graphs are due to the cognitive processes that occur after attention is focused on specific aspects of the graph; as a result, most of the testing methods we discuss are focused on the attentive portion of the perceptual process.

```
# Density plot
ggplot(data = banknote,
       aes(x = Bottom,
           group = Type,
           fill = Type)) +
  geom_density(position = "stack")
```




```
# Dotplot
ggplot(data = banknote,
       aes(x = Bottom,
           group = Type,
           fill = Type)) +
  geom_dotplot(method = 'histodot',
               stackgroups = TRUE)
```




```
# Histogram
ggplot(data = banknote,
       aes(x = Bottom,
           group = Type,
           fill = Type)) +
  geom_histogram(position = "stack")
```

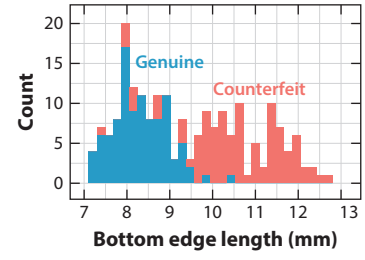


Figure 6

Three different plots of the Swiss bank note data (Flury & Riedwyl 1988), created using the grammar of graphics as implemented in the `ggplot2` code. The data consist of measurements of the dimensions of the banknotes, including the length of the bottom edge, which is shown in the plots. The main plot specification and syntax remain the same, with the form of the plot changing due to the specification of the geometric object used to represent the data.

2.1.1. Preattentive graph perception. Initial research into preattentive perception used a search task, where participants had to identify a particular object in a field of distractors, manipulating display size and varying one or more features such as color or shape; participants' search times were measured to determine the amount of effort necessary to complete the search task. Preattentively perceived features show a near-constant reaction time over increasing display size, while features that are processed attentively show an increasing reaction time with increased display size (Treisman 1980). A primary question in the discussion of preattentive graph perception is whether there are advantages in designing a graph to promote the preattentive perception of features, ideally reducing cognitive load.

We distinguish between tests that use graphical forms and more primitive tests that use basic geometric elements during the testing process. The results from more primitive experimental designs still apply to the design of graphs and charts, but the experimental design does not involve any display of actual data (**Figure 7** is an example of experimental stimuli that do not represent actual data). Preattentively processed features include shape, angle, size, and texture; however,

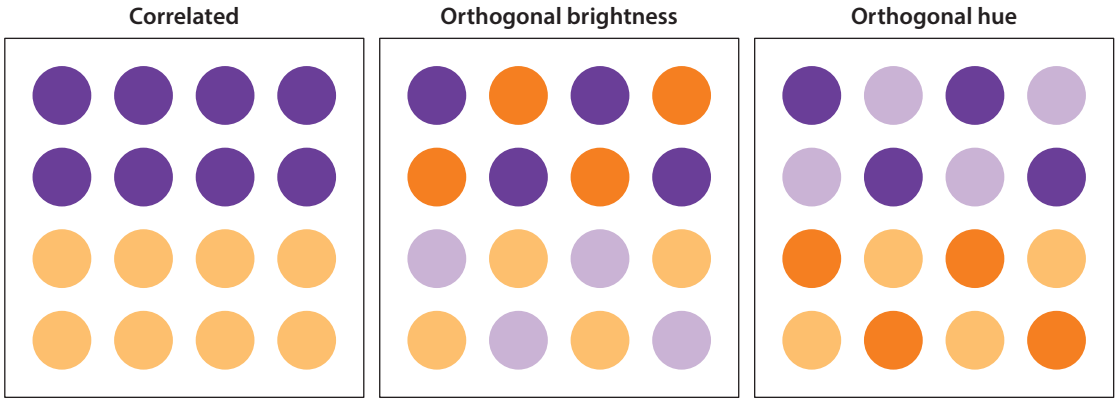


Figure 7

Callaghan (1984) showed that irrelevant variation increases response times, indicating that hue and brightness are integrated and preattentively perceived. The images shown here are examples of the types of stimuli used in the experiment. While hue and brightness can be manipulated separately, they are preattentively perceived as a single unit.

typically, combinations of preattentive features which represent separate features in the data are processed attentively, with at least one major exception. Callaghan (1984) demonstrates that hue and brightness are integrated, that is, that even though they can be separately manipulated, they are still perceived preattentively as a single unit, using arrays of tiles similar to those shown in **Figure 7**.

Extending this work, Healey et al. (1996) use the same segmentation paradigm, applied to more complex charts utilizing actual data, with the goal of exploring region segmentation using preattentive cues. Healey's experiments use multivariate displays, leveraging the preattentive grouping of similar objects to separately represent features using color, height, and texture (Healey & Enns 1999). While these displays may not follow best graphical practice in other respects, they do show the utility of designing with the preattentive perceptual process in mind.

2.1.2. Attention mediated testing methods. Creators of a chart or graph typically operate under the (hopefully safe) assumption that readers will spend more than 300 ms considering its contents; as a result, attention mediated testing methods allow a more realistic mechanism for testing overall performance of different graphical forms. Cleveland & McGill (1987) discuss the different approaches to research methodology in this area, and while they do not include all of the experimental approaches we discuss, the article makes an important distinction between informal and formal graphical exploration. In the informal approach, changes are made to the graph, and the iterative versions are compared to determine what information is easily accessible; in the formal approach, an experiment is designed, and participants are tested in a controlled manner. This section describes several different experimental approaches that can be used to answer the general question of "How effective is this graph at communicating useful information?"

2.1.2.1. Direct observation: numerical estimation, speed, and error rates. One of the simplest ways to test the utility of a graph is to verify that information can be accurately read from it. Graphs are, after all, generally recognized as having more utility than tables for presenting information in an accessible and useful way; if that information cannot be read back out in a relatively accurate manner, the graph's utility is suspect. Early experiments, such as those of Eells (1926), Croxton & Stryker (1927), and Croxton (1932), used accuracy alongside speed and other considerations for

plot evaluation. Later studies (Peterson & Schramm 1954, Cleveland & McGill 1984, Broersma & Molenaar 1985, Dunn 1988, Tan 1994, Amer 2005) were conducted with similar methodology; in essence, the participants are provided with a chart and asked to estimate some quantity or answer a predefined question using the information provided in the chart. The accuracy of various types of charts, as measured by participants' responses to the questions, is then used to determine which charts are superior. It is important to ensure that the specific charts and questions used are aligned; studies are commonly critiqued on the basis that the charts or the questions were not appropriate for the task—indeed, the first studies of pie versus bar charts were heavily criticized on these grounds (von Huhn 1927).

A similar type of study avoids the pitfalls associated with numerical estimation by showing participants multiple charts in sequence and asking them to evaluate the differences in the charts with respect to the dimension of interest. Shah & Carpenter (1995) required participants to examine the relationship between three variables as shown in several different types of line charts. One set of charts used in the experiment is recreated in **Figure 8**; in the first plot, it is much easier to determine that there is no relationship between pupil size and depression for one level of thyroid activity compared with the plot.

In some studies, response time is the main quantity of interest. Participants are provided with a certain chart or stimulus, and the amount of time spent considering the stimulus before answering is used together with accuracy of the answer to assess the difficulty of the task, under the premise that more difficult or mentally demanding tasks will require more time before a response is

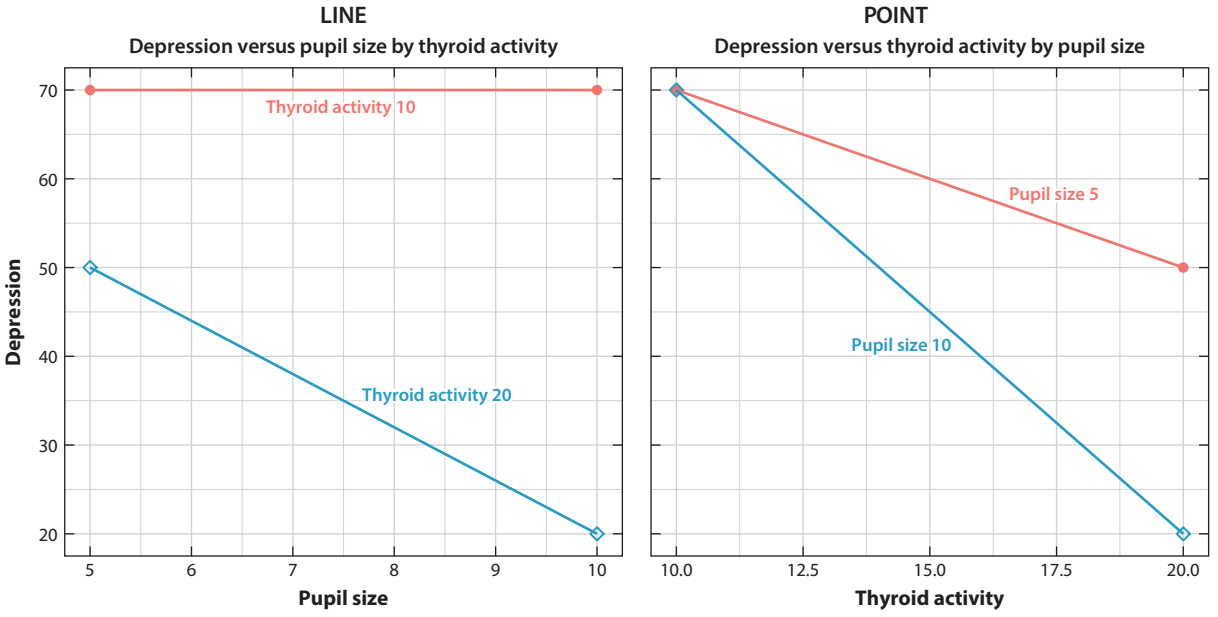


Figure 8

Sample plots from Shah & Carpenter (1995), showing data where one independent variable (pupil size) has no effect on the *y* variable (depression) for one value of the independent variable (thyroid activity). The two representations of this data, referred to as line and point, are not equally effective for communicating the joint relationship of the two independent variables. Line and point are terms that refer to the angle from which the planar representation of the 3D data is viewed. To make the comparison in the line representation, the viewer needs to compare the slopes of the lines, whereas to make the comparison in the point representation, the viewer needs to explicitly compare the values of each data point.

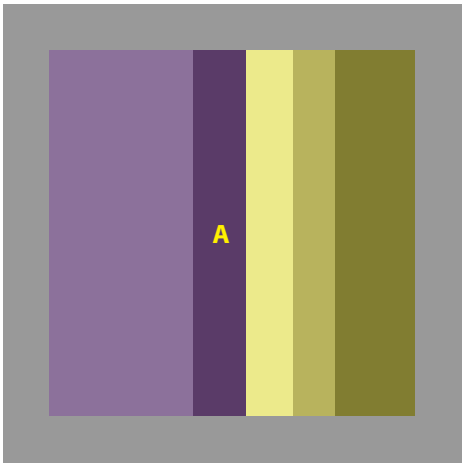


Figure 9

A framed spine plot similar to those shown in the *Statistical Atlas* (Walker 1874), recreated using modern methods. The chart shows the proportion of church accommodations in Oregon, with the four largest denominations in the state shown in the center and the unaccommodated population shown in the gray region framing the plot. Participants were asked to estimate the size of the area labeled A relative to the entire chart; the presence of the frame significantly decreases the accuracy of these estimates (Vanderplas et al. 2019).

generated. Carswell & Wickens (1987) use the error rate and response time to determine whether triangular object displays are superior to bar graphs to represent the inputs and outputs of a dynamic system (see their figure 2 for an example of an object display); Legge et al. (1989) use efficiency, a function of time and accuracy, to assess different types of charts. More modern studies may collect both response time and numerical estimation data online, using services such as Amazon Mechanical Turk to conduct usability studies in statistical graphics (Heer & Bostock 2010). We have even used direct estimation in combination with online testing to examine very old charts from Walker's (1874) *Statistical Atlas* (Vanderplas et al. 2019), determining that framed mosaic, spine, and pie charts are suboptimal graphical representations. **Figure 9** shows an example of the direct estimation task used in the experiment; participants entered an estimate of the proportion of the chart represented by the area labeled A. Some of the other experimental paradigms, such as psychophysics and implicit tests, can also be used in combination with Amazon Mechanical Turk and other online testing services.

There are limits to what one can test using direct estimation: It is generally preferable to test only very straightforward assessments of the content of a chart or graph, to fit within a simple experimental paradigm. In addition, open-ended estimation tasks elicit certain well-known biases such as the tendency to round to multiples of 5 or 10 (Baird et al. 1970). Long-term interaction with a complex graph or chart showing multiple layers of data is generally not ideal within this paradigm, which requires a fixed set of numerical assessments that do not accurately represent how we explore a new, complex graphic. This complexity explains why there are so few studies of rich, complex graphics that may require domain expertise in addition to the ability to read information from a visual display. To approach situations with more complicated graphics, or charts that are known to induce perceptual biases in the participants, it is often more useful to utilize other experimental paradigms that facilitate examination of specific parts of the perceptual process—such

as the use of eye tracking to measure attention and motivation, or the use of verbal descriptions to assess more complicated graphs. The next sections cover some of these more nuanced approaches to explicit testing of graphs.

2.1.2.2. Psychophysics and signal detection theory. Some studies of graphs utilize psychophysics methodology to assess data visualizations. Initially, of course, a significant portion of the research in statistical graphics came from the fields of psychophysics and cognitive psychology (Teghtsoonian 1965, Lewandowsky & Spence 1989, Spence 1990), but in most cases this was not accompanied by the use of the methods of psychophysics for experimental testing of charts and graphs. Psychophysical experimental design is focused on whether an effect is detectable and whether the magnitude of the effect can be accurately estimated. Common methods, such as the method of constant stimuli and the method of adjustment, involve repeatedly presenting a participant with charts and asking them to evaluate the chart on the basis of a particular question of interest. In the method of adjustment, this is done with the control of the participant, who adjusts the stimuli interactively until the effect is just barely noticeable; in the method of constant stimuli, the effect size changes randomly from trial to trial to reduce continuity effects. In graphical testing, Hughes (2001) used the method of constant stimuli to assess the ability to detect a difference in height in 2D or 3D bar charts; Vanderplas & Hofmann (2015) used the method of adjustment to experimentally determine the size of the line-width illusion's distortion of variance perception. Psychophysics methods also seem to be relatively common in studies of map perception, particularly when the goal is to estimate the amount of exaggeration or other corrective distortion necessary for realistic perception of the map (for an overview of relevant cartography studies, see Brandes 1976).

2.1.2.3. Thinking aloud. Another approach to testing graphics is to examine the cognitive processes that occur as a graph is read. Lacking mind-reading devices, the next best option is to ask participants to talk through their thoughts as they read and use a graph in a realistic setting (concurrent think-aloud, or CTA), or as they recall a graph after the fact (retrospective think-aloud, or RTA) (Guan et al. 2006). The think-aloud process allows experimenters to examine the use of complex graphics in the wild, or at least in situations that are less artificial than the paradigms allowed by numerical estimation and psychophysics methods. Think-aloud studies allow researchers to attempt to measure insight (North 2006) and reasoning (Dunbar 1995) in complex situations such as experimental design, decision-making (Normand & Bailey 2006), or the process of weather forecasting (Trafton et al. 2000, Kirschenbaum 2003). In forensics, think-aloud studies are known as white box studies, because it is possible to understand what a forensic examiner is thinking and why they make a specific conclusion about the evidence (Ulery et al. 2011). Studies using think-aloud protocols have also examined the process of exploratory data analysis, finding that unexpected results are more likely to be represented in informal terms initially, but that with familiarity, language shifts to formal explanations (Trickett et al. 2000a,b). The think-aloud protocol is also conducive to use with interactive graphics, combined with logging or video recording software that can record the state of the graphics device in parallel with the user's monologue. Studies have also combined think-aloud protocols and eye-tracking studies with the goal of validating CTA (Cooke 2010) and RTA protocols (Guan et al. 2006) for use with statistical graphics and general usability testing. While the data that result from the think-aloud protocol are typically more qualitative and less quantitative than results produced using other methods, they provide significant additional insight into the underlying cognitive processes affecting visualization, which cannot be obtained through other means.

2.1.2.4. Eye tracking. Where think-aloud protocols allow insight into the cognitive process, eye tracking facilitates insight into the process of visual attention, providing data on the approximate spatial location of visual focus. The attention-fixation process occurs too quickly to be accurately verbally communicated, but eye-tracking equipment allows experimenters to identify the portions of the chart that require attention and sustained cognitive effort or that attract interest from participants, inferring from gaze and fixation the cognitive processes occurring during graph comprehension. Eye tracking allows researchers to determine that viewers spend relatively little time examining the axes in scatter plots, but significant amounts of time examining the axes in parallel coordinates plots (Netzel et al. 2017), suggesting that the process of reading these two chart types is fundamentally different. Another study leveraged eye tracking to identify features that provide useful information during the graph reading process for several different types of charts (Goldberg & Helfman 2010). Eye tracking is a powerful tool when combined with good experimental design: Fabrikant et al. (2010) examined fixations when users were shown meteorological charts before and after being provided with introductory training about meteorology, finding that after training, users' conclusions were more accurate, response time increased, and fixations were directed to more useful areas of the maps. The use of eye tracking with different source populations also allows researchers to understand how dyslexia (Kim et al. 2014) and graph literacy (Woller-Carter et al. 2012, Okan et al. 2016) affect the graph comprehension process, providing better design guidelines for specific target audiences.

2.1.2.5. Combination experiments. Many studies use a combination of the explicitly structured graphical tests discussed here. Think-aloud studies are relatively easy to integrate with eye-tracking studies, and it is not difficult to add in direct observations or psychophysical evaluation methods as part of the trial design. Psychophysics and direct observation studies are limited by the questions that are asked, eye-tracking results only provide information on visual focus, and think-aloud results are generally qualitative, but when these techniques are combined, they provide a much more complete picture of the cognitive processes that underlie graphical perception. Ryu et al. (2003) used a combination of eye tracking, cognitive tests, direct observation, and think-aloud protocols to examine the integration of information across multiple types of charts, determining that integrating information from multiple parallel coordinates plots is slow, difficult, and inaccurate compared with information integration when a scatter plot or map is presented with a parallel coordinates plot. Zgraggen et al. (2018) also used a combination of think-aloud, eye tracking, direct observation, and interactive graphics to examine the impact of exploratory data analysis on multiple testing problems. Many of the attention mediated, explicitly structured testing methods can also be combined with implicitly structured tests, which we discuss in the next section, to produce a more comprehensive view of the process of graphical perception.

2.2. Implicit Graphical Tests Using Visual Inference

Explicit graphical tests, as we have referred to them, are tests in which the user is directed to assess a specific feature of a plot or answer a specific question. That is, the tested hypothesis is explicitly stated, providing the user with cues to the intended purpose and function of the plot and/or the relevant features of the data shown within the plot. In contrast, in an implicit graphical test, the user must identify both the purpose and function of the plot and use that information to evaluate the plot as shown. Typically, these tests are structured as visual inference problems, as introduced by Buja et al. (2009), though other formulations of implicit tests exist as well (Hasanhodzic et al. 2010). Explicit tests are typically conducted on plots that have been created to showcase specific structure in the data in order to present results; in contrast, implicit tests are designed to inform

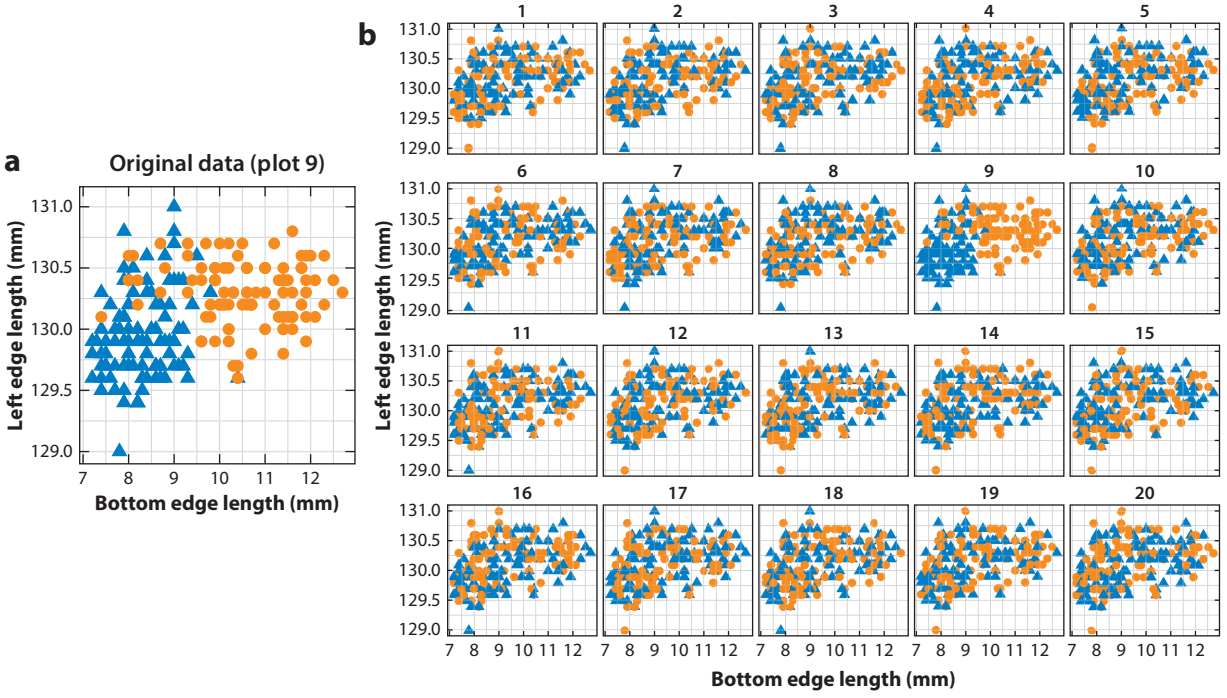


Figure 10

The original data (panel *a*) are embedded into a set of 19 generated under the null hypothesis (panel *b*). A visual hypothesis test is conducted by asking multiple users to select one (or more) plots that are different; a *p*-value can then be calculated from the answers. In this particular lineup, plot 9 contains the original data; the other plots are generated by randomizing the vector of color/shape values as in a randomization test.

exploratory data analysis (as advocated by Tukey 1965) and the iterative model diagnostic process. During data exploration (EDA), statisticians typically examine the data using many different plots, considering different aspects of the data and selecting only those results that are visually interesting for further exploration. This raises the question of whether the visually interesting features in those plots show actual signal and do not arise simply by chance. This question is a direct consequence of our general use of graphics and our definition of "interesting." Generally speaking, something is interesting if it does not follow our expectations in some way—in other words, we conduct an implicit hypothesis test when drawing and looking at charts. Interesting charts are those deemed significant according to this implicit hypothesis test. Making this hypothesis test explicit allows for a more formal evaluation of the significance of a visual finding. In terms of a classical hypothesis test, a plot of the data is taking the role of a (visual) test statistic. Null plots are created from data in accordance with the null hypothesis (e.g., permuted data, if the null hypothesis assumes that there is no nonrandom relationship between the variables). If the data plot is visibly different from the null plots (i.e., it can be picked out through visual inspection by an independent evaluator), this counts as evidence against the null hypothesis, and once enough evidence is accumulated, we would conclude that the null hypothesis has to be rejected and accept that the feature in the data plot is not random. Formally, a statistical lineup of m visual statistics consists of $m - 1$ plots $T(y_0)$ simulated from the model specified by the null hypothesis, and the test statistic $T(y)$ produced by plotting the actual data, which may arise from the alternative hypothesis. **Figure 10** demonstrates this process, with the 19 null plots generated by randomizing the values of y. The statistical lineup

is then evaluated by K independent observers, with the resultant p-value calculated according to the null hypothesis that each of the m visual statistics are equally likely to be selected (Majumder et al. 2013). The `nullabor` package (Wickham et al. 2018) contains tools for graphical inference, including p-value calculations and power calculations for visual inference.

Implicit graphical tests approach the problem of spurious plot relationships at the level of the data, leveraging the human visual system to conduct a suite of visual tests for features such as outliers, clusters, and linear and nonlinear relationships. The advantage to implicit testing is that lineups do not require a specification of a feature of interest in the testing framework, i.e., we do not have to ask questions such as "which group has a higher proportion of responses." Much of the historical research of comparing different types of charts has been criticized because the specific question phrasing does not provide readily generalizable results; the lineup protocol removes this obstacle by charging the user with the task of identifying the most different looking plot and thereby selecting the feature with the visually most salient difference compared with the other plots. This allows an evaluation of competing plot designs without the complications of potentially steering participants toward a particular outcome by phrasing questions. This leads to a completely data-driven result: If bar charts are indeed better suited for a task than pie charts, the target plot will be selected more frequently when the lineup is presented with bar charts than when the lineup is presented with pie charts. The real power of the lineup protocol is that when combined with the grammar of graphics, we can hold the underlying data and summary statistics constant, isolating the effect of different plot types, coordinate transformations, and aesthetic mappings on our ability to detect effects in the data.

Statistical lineups have been used experimentally to examine single plot types in many contexts: residual plots in hierarchical linear models (Loy & Hofmann 2015), perceived clustering in high-dimensional data (Roy Chowdhury et al. 2015), and spatial clustering in geographical research (Widen et al. 2016). Loy et al. (2016) used statistical lineups to evaluate different types of Q-Q plots for assessing violations of normality, determining that the visual tests were generally more powerful than common numerical tests when assessing violations of normality. Beecham et al. (2017) used lineups to assess the effect of spatial autocorrelation when represented using different grid structures, finding that lineups can be used for choropleth maps if the null plots are generated under models with reasonable spatial autocorrelation models. Other studies have also found that the approach to null model generation is critical. It can be difficult to specify the null data-generating model in a way that adequately mimics the data plots, which suggests that visually, we are able to identify many more features than those typically tested using standard quantitative hypothesis tests. This implicit testing of many different hypotheses does make null distribution specification challenging, but it also highlights the power of visual cognition to detect subtle differences in data.

Typical lineups contain a single target plot, but this is not a requirement. Vanderplas & Hofmann (2017) used two targets, each generated from a competing data model: clustering or linear association. The null plots in this experiment were composed of a mixture of points generated from each of the two data models, ensuring that the two targets were both slightly more extreme than the null plots, but that the null plots and the target plots shared some features. Using this approach, the authors tested the effect of different aesthetics on the selection of each of the two targets, examining the strength of aesthetics such as color, shape, trend lines, error bands, and 95% ellipses for highlighting clustering or linear trends in the data. By providing viewers with a choice between data generated by competing models, the two-target lineup approach provides a way to directly examine the visual strength of each model compared with the null and comparing the models directly. For instance, when comparing a model generating clustered data to a model generating data with a linear association between y and x, this protocol establishes that

color and shape aesthetics slightly increase the likelihood that the cluster target plot is selected, while a trend line aesthetic slightly increases the probability that the linear relationship target plot is selected. More broadly, Vanderplas & Hofmann (2017) show that the aesthetics used in a plot can significantly impact the perceived relationship between variables.

In the ten years since the introduction of the lineup protocol, many studies have leveraged the grammar of graphics to ensure that the underlying data mapping remains the same while manipulating the geometric representation of the data and overall visual appearance of the plot. Combining the conclusions from both implicit lineup tests and the explicit tests described in the previous section, what can we say about best graphical practices, beyond "pie charts are awful?"

3. CURRENT BEST GRAPHICAL PRACTICE

All of the user testing in the world cannot identify the best possible graphic—we can instead only experimentally assess which graphical designs are better for a specific purpose. This can lead to a rather fragmented approach when describing best practice, and so in order to avoid this, we examine graphical practice using the principle of "first, do no harm" from the Hippocratic oath.

3.1. Cognitive Principles

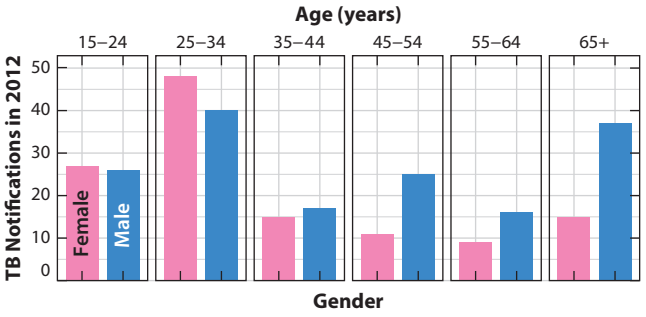
A useful starting point is to apply gestalt principles of visual perception (Wagemans et al. 2012a,b), such as proximity, similarity, common region, common fate, continuity, and closure, to data plots. These principles are useful because good graphics take advantage of the human visual system's ability to process large amounts of visual information with relatively little effort. Understanding the principles that underlie this processing allows us to create charts that require less cognitive effort to read, freeing us to think about the content rather than the form of the chart.

3.1.1. Proximity. This principle is that objects or shapes that are close to one another appear to form groups. For plot design, proximity is used to place items to compare close together, and less important comparisons further apart. **Figure 11** illustrates this principle when using faceting in a plot, using data on tuberculosis (TB) incidence in Australia in 2012. When plots are faceted by age, it means gender is easier to compare. We learn that more females than males are detected to have TB in their early twenties, but in the aging population, males are more commonly detected with TB. Conversely, when plots are faceted by gender, the distribution of age is easier to examine. We learn that the age distribution of TB incidence for females skews heavily toward younger women. For males, TB incidence is more uniform across ages—there are high counts at young, old, and middle-age categories. Effectively utilizing proximity in organizing plots makes a huge difference in ease of information communication.

3.1.2. Similarity. The gestalt principle of similarity suggests that we group things that have similar appearances and exclude objects with different appearances. In charts, this principle is often leveraged by coloring points or bars according to a categorical variable or by using points of different shapes to represent different categories. Vanderplas & Hofmann (2017) showed that the addition of color and shape to a scatter plot increases the likelihood that individuals will perceive clustered groups of points. In **Figure 11**, the coloring of bars allows us to easily see that the similarly colored rectangles represent the same group of people, even though the bars are separated by facets and other groups.

3.1.3. Common region. The gestalt principle of common region suggests that elements contained within a common region belong together. Common region helps us to easily read small

```
# genderinage
ggplot(tb_au_12,
       aes(x = gender,
           y = count,
           fill = gender)) +
 geom_bar(stat = "identity") +
  facet_grid(~age)
```




```
# ageingender
ggplot(tb_au_12,
       aes(x = age,
           y = count,
           fill = age)) +
 geom_bar(stat = "identity") +
  facet_grid(~gender)
```

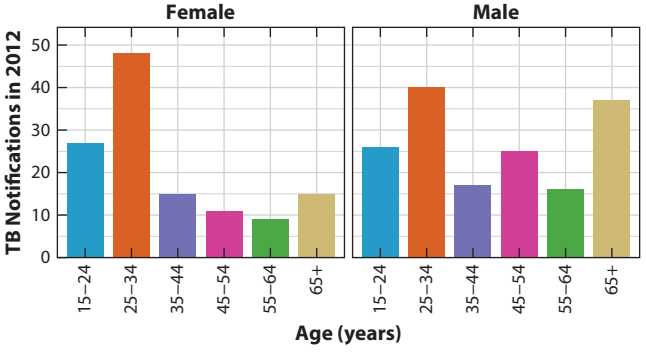


Figure 11

Two different arrangements of the same data to illustrate how the proximity of elements makes a question easier to answer. Both plots show TB notification counts by demographics in Australia for 2012. It is much easier to make cross-gender comparisons using the first chart, which places bars for each gender in the same age category next to each other. In the second chart, cross-gender comparisons are difficult because the distance between comparative bars is much greater, requiring more mental work. Abbreviation: TB, tuberculosis.

multiple plots because the graphical elements of each small plot are grouped into a single entity that can be examined on its own. In addition, confidence bands and bounding ellipses also activate this gestalt principle by grouping points within the boundaries together (Vanderplas & Hofmann 2017), highlighting the presence of outliers that do not belong to the main group.

3.1.4. Common fate. The gestalt principle of common fate describes the tendency to group objects that are moving together in the same direction and at the same speed together. Common fate is certainly active in animated plots that use fading or transitions over time, but even in static plots, continuity can be activated when multiple time series plots are shown together. **Figure 12** shows an example. Four examples are displayed as overlaid time series (*top*), and another four are shown as scatter plots (*bottom*). The scatter plots show the values of the two separate time series plotted against each other, with observations from the same time point represented as a single point on the scatter plot. In the time series, plot 4 is likely perceived as having stronger association than plot 3, due to the lines moving roughly in a similar pattern. Strong negative association is not easily detected from overlaid line plots, but it is easily seen in a scatter plot (Tomasetti 2015). If negative association is suspected, either using a scatter plot or inverting one series is suggested.

Overlaying a few time series on a common scale is one way to activate the heuristic of common fate, but this does not scale well to larger numbers of simultaneous measurements. Javed et al. (2010) provide experimental evidence that small multiples are better than other alternatives when there are many simultaneous time series to display and the series cover a large visual span.

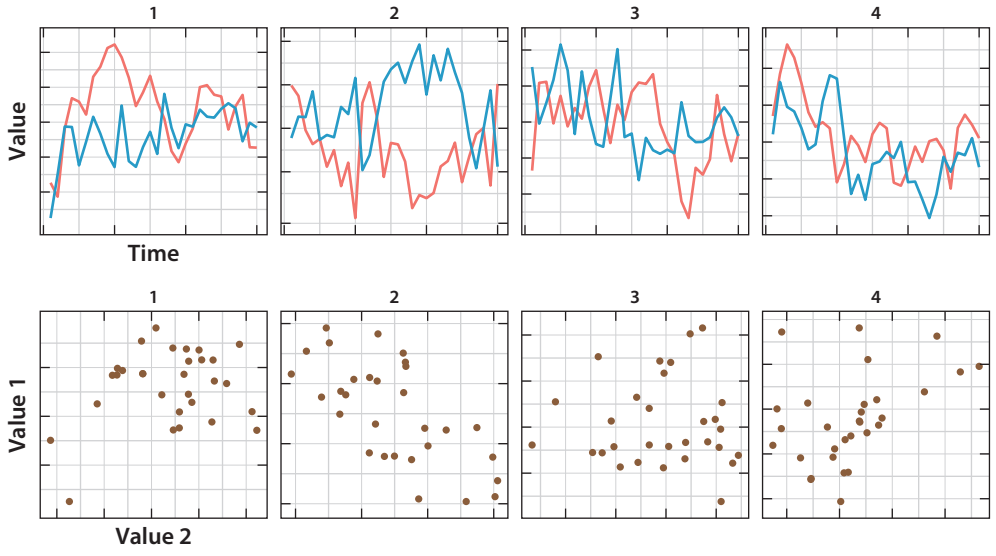


Figure 12

Four examples of pairs of time series displayed as line plots (*top*) and scatter plots (*bottom*). The gestalt principle of common fate may dominate the perception of correlation in the overlaid line plots. In these four examples of line plots, viewers will likely say that plot 4 has the most similar series, but it is actually plot 2, with strong negative correlation between the two series. A scatter plot clearly shows the negative association.

3.1.5. Working memory. Another cognitive limitation that affects plot comprehension is the limit on working memory. Typically, working memory is limited to approximately seven (plus or minus two) items, or chunks. In practice, this means that categorical scales with more than seven categories decrease readability, increase comprehension time, and require significant attentional resources, because it is not possible to hold the legend mapping in working memory.

3.1.6. Change blindness. Not all information is stored in memory (working or long-term) as it is represented. Simons & Levin (1997) suggested that the vivid details available to us instantaneously when examining the world fade quickly and are replaced with broad semantic descriptions or less meaningful cognitive representations of a scene. As a result of this compression, it is difficult to identify changes between similar scenes. This phenomenon, known as change blindness, affects both static and interactive plots. In static plots, it can be difficult to compare between different small multiples or facets because the contents of the plots are not reliably represented in working memory when switching attention between them. In animated plots, it is important to use transition effects to connect successive frames of the animation: This reduces change blindness and also activates the gestalt principle of common fate, allowing us to quickly identify groups of objects that are transitioning in the same direction.

3.1.7. Ease of comparisons. Much of the psychophysics research on statistical charts examines the accuracy of comparisons and quantitative evaluations made during the process of understanding a plot. This research can be distilled into a hierarchy of comparisons (based primarily on Cleveland & McGill 1984, 1985; Shah & Miyake 2005; Lewandowsky & Spence 1989), ranked by their accuracy and difficulty as follows (roughly equivalent tasks are listed together):

1. Position (common scale)
2. Position (nonaligned scale)
3. Length, direction, angle, slope
4. Area
5. Volume, density, curvature
6. Shading, color saturation, color hue
7. Discriminable shape
8. Indiscriminable shape

Examples of these charts are shown in **Figure 13**.

This ranking of cognitive tasks provides some consistent guidance for chart design: If the same data can be represented in a way that allows the user to make a comparison more accurately (based on the hierarchy), then that design is preferable. Thus, this hierarchy indicates that in most cases, it is better to use a stacked bar chart than the equivalent polar-coordinate pie chart, because a stacked bar chart requires evaluation of length, while a pie chart requires area comparisons. If information can be shown on an *x*- or *y*-axis rather than using color (saturation, hue, or shading), it will be easier to make numerical comparisons—we can generally order information based on color, but estimation of numerical quantities is much less precise using color than position. While the hierarchy of graphical comparisons provides some guidance, there are other design choices that can be informed by experimental research in a less systematic way.

3.2. Chart Design

Vanderplas & Hofmann (2017) provide compelling evidence that the aesthetics used in a chart can significantly affect how plots of the same data are read, and they explain these differences relative to the gestalt heuristics activated by each combination of aesthetics. The use of redundant aesthetics that activate the same gestalt principles (such as color and shape in a scatter plot, which both activate similarity) results in higher identification of corresponding data features. In addition, dual encoding increases the accessibility of a chart to individuals who have impaired color vision or perceptual processing (e.g., dyslexia, dysgraphia). This experimental evidence directly contradicts the guidelines popularized by Tufte (1991), which suggest the elimination of any feature that is not dedicated to representing the core data, including redundant encoding and other unnecessary graphical elements.

3.2.1. Color. While historically there were constraints on the use of color in graphics due to technological limitations and the economics of printing, these restrictions have evaporated with the advent of computer-generated graphics, relatively inexpensive color printing, and an increasing tendency to share charts and graphs digitally instead of in print. Color can encode categorical and continuous variables and, when used effectively, provides a nearly effortless way to group plot elements using a medium that does not require conscious attention. Unfortunately, while there are many ways to use color correctly in a plot, there are generally even more ways to use it incorrectly.

Color scales should be chosen to best match the data values and plot type: If the goal is to show magnitude, a univariate color scheme is typically preferable, while a double-ended color scale is typically more effective when showing data that differ in sign and magnitude. Where possible, color scales should use a minimal number of hues, varying intensity or lightness of the color to show magnitude, and transitioning through neutral colors (white, light yellow) when utilizing a gradient. **Figure 14** shows an example of perceptually suboptimal, good, and color-blind–accessible diverging color schemes. Cognitive load can also be reduced by selecting colors

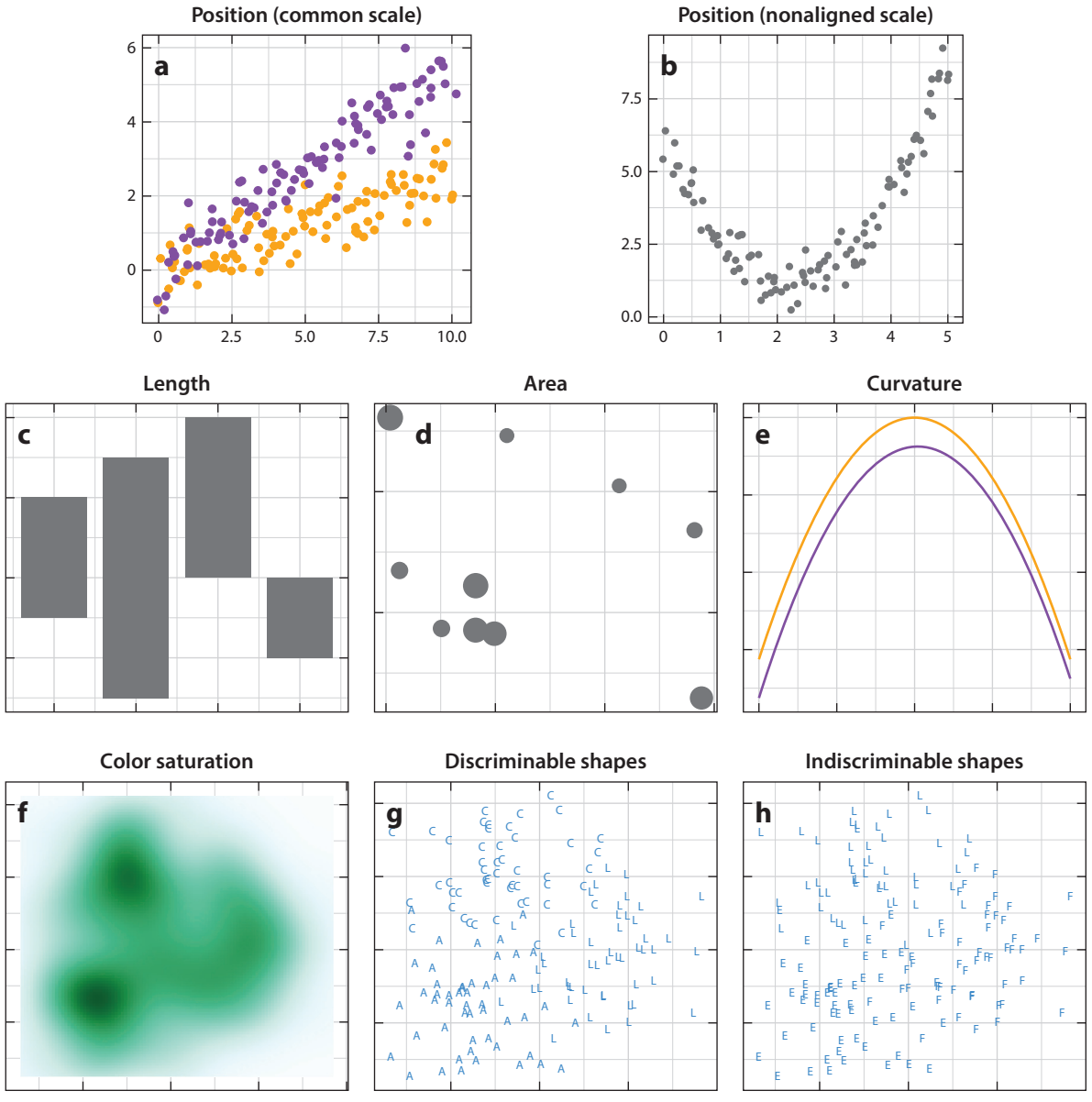


Figure 13

Examples of visual comparisons ranked by difficulty. (*a*) Two sets of points are plotted on a common scale; it is easy to compare the relative trajectories of the two lines. (*b*) This plot shows another set of points that could be compared with the points in panel *a*; however, because the scales are not common between the two, accurate comparisons are more difficult. (*c*) The rectangles must be compared by length, because they do not start or end at the same point. (*d*) Points are sized based on another variable, requiring comparisons of area; these comparisons are more difficult to make with numerical accuracy than judgments based on length or position. (*e*) Two curves are shown that are not completely parallel, requiring us to make a comparison of curvature; in addition, note that the difference between the two curves is hard to perceive accurately (Vanderplas & Hofmann 2015). (*f*) This plot shows two-dimensional density using the fill of the tiles; it is possible to make ordered comparisons here but much more difficult to estimate numerical values from the plot even when a legend is provided. (*g*,*h*) These show the same data as panel *f*, but using discriminable shapes and indiscriminable shapes, respectively. Discriminable shapes have different features, while indiscriminable shapes tend to have the same features and require more cognitive effort to separate the clusters.

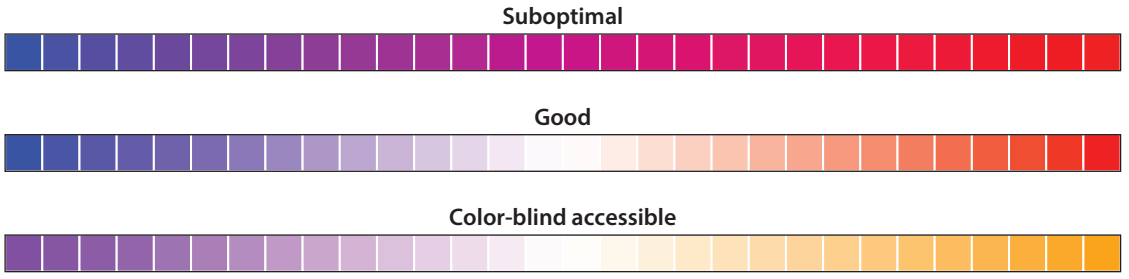


Figure 14

Examples of diverging color schemes that are perceptually suboptimal (no transition through a neutral color), good, and color-blind–accessible. The purple–orange gradient through white is distinguishable by individuals with any type of color deficiency affecting one of the three types of cones in the retina.

with cultural associations that match the data display, such as the use of blue for men and red (or pink) for women, or the use of blue for cold temperatures and red/orange for warm temperatures.

It is also important to consider the human perceptual system, which does not perceive hues uniformly: We can distinguish more shades of green than any other hue, and fewer shades of yellow, so green univariate color schemes will provide finer discriminability than other colors because the human perceptual system evolved to work in the natural world, where shades of green are plentiful. **Figure 15** shows the International Commission on Illumination (CIE) 1931 color space, which maps the wavelength of a color to a physiologically based perceptual space; a significant portion of the color space is dedicated to greens and blues, while much smaller regions are dedicated to violet, red, orange, and yellow colors. This unevenness in mapping color is one reason that the multi-hued rainbow color scheme is suboptimal—the distance between points in a given color space may not be the same as the distance between points in perceptual space (Light & Bartlein 2004, Wakita & Shimamura 2005, Borland & Ii 2007). As a result of the uneven mapping between color space and perceptual space, multi-hued color schemes are not recommended.

While color is an extremely useful aesthetic for the majority of the population, between 4 and 8% of males (and a much smaller proportion of females) have some sort of color perception deficiency (color-blindness) (Wakita & Shimamura 2005), which reduces the space of distinguishable colors. Color-blindness is common enough that it is reasonable to expect that any given chart used in a presentation or publication will be read by someone with a color perception deficiency. The use of dual encoding allows color-blind individuals to more readily read graphics that utilize color, and as hue and lightness can be varied separately, it is possible to use dual encoding without adding another aesthetic. If accessibility is the goal, the default colors used in `ggplot2` (Wickham 2016) should be avoided, as the saturation is held constant in this color scheme and only the hue of the colors is varied. A perceptually more successful approach to color manipulation can be found in the R package `colorspace` (Zeileis et al. 2009). This package is also based on the HCL (hue, chroma, luminance) space of colors, but it makes use of both luminance and hue when setting up color schemes and allows an assessment of colors on these three dimensions. There are several preexisting color schemes that may be more discriminable to color-blind individuals, such as those provided by Lumley (2013) or Brewer (2019). Similarly, when color accessibility of a plot is a primary consideration, the use of the default gray background in `ggplot2` and other plotting systems should be replaced with white to maximize the contrast between background and plotted features. The default gray background for plots, which by now has become `ggplot2`'s signature look, dates back to a recommendation from Carr (1994) to make grid lines in plots readable without dominating the data. Obviously, light gray grid lines on a white background serve the same purpose.

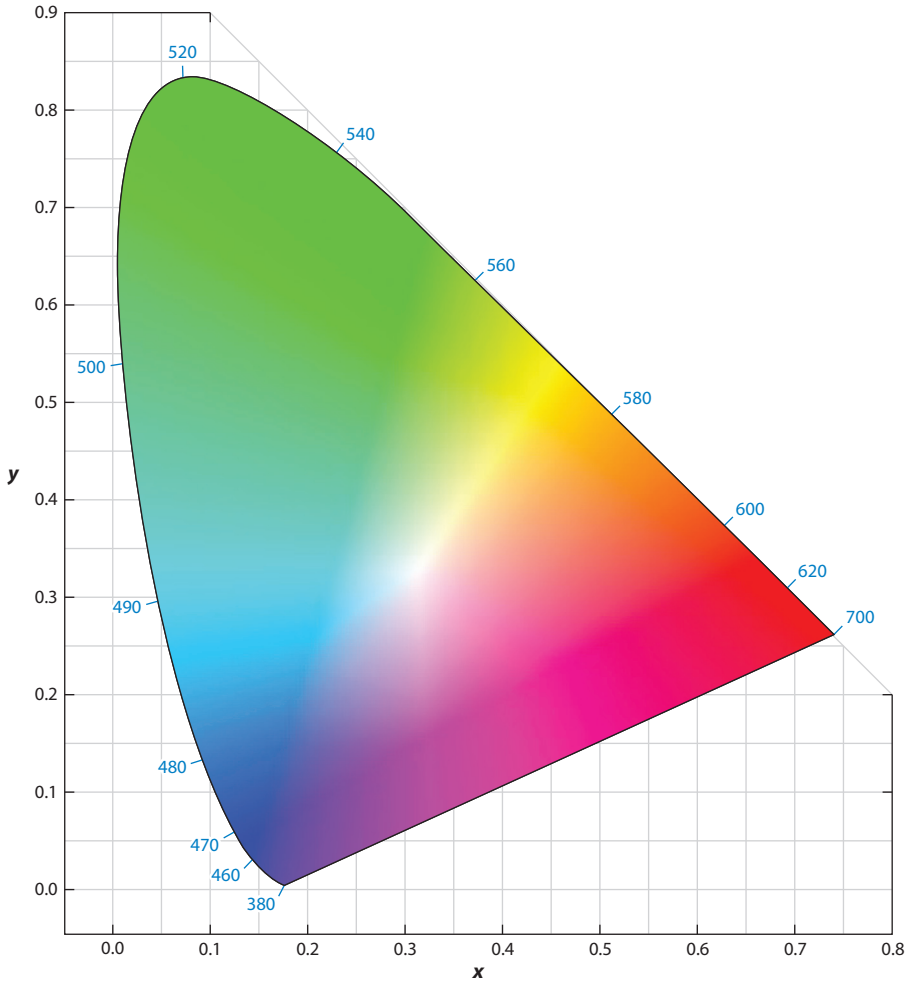


Figure 15

The CIE 1931 color space chromaticity diagram. The CIE 1931 color spaces defined the color space based on physiologically perceived colors in human color vision. The outer boundary of the curve marks the color spectrum, with wavelengths in nm. Note that the distance between successive markings on this boundary is not constant, indicating that the perceptual distance between colors does not match the physical distance in wavelength. More broadly, there is much less area devoted to red, orange, yellow, and violet compared with blue and green; as a result, a rainbow color scheme with equal spacing by wavelength would perceptually overemphasize the tails of the range of values represented by the scheme. Abbreviation: CIE, International Commission on Illumination.

4. OPEN QUESTIONS AND FUTURE RESEARCH

One of the areas in graphics showing the most recent growth and creating the most excitement in the community is interactive graphics. There is a huge abundance of applications and graphics claiming to be interactive—yet then (Swayne & Klinke 1999) and now, there is not much agreement over what "interactive graphics" actually means. While any communication between user and device can be considered an interaction, interactive graphics should be defined as a user-driven direct manipulation of plots and plotting elements with an immediate reaction (Becker

et al. 1987, Eick & Wills 1995, Unwin 1999). One of the foundations of implementing interactive graphics is the formalization of the data pipeline necessary to support user interactions with the plot (Buja et al. 1988, Lawrence et al. 2009, Wickham et al. 2009, Xie et al. 2014). The R package `shiny` (Chang et al. 2019) is built on this idea; thousands of interactive dashboards and web applications have been created using that platform.

Formal testing for interactive graphics is difficult without an established and generally accepted grammar of interactive graphics. However, there are several promising approaches into this direction, such as the Python framework Vega-Lite (Satyanarayan et al. 2017) and R packages, e.g., `ggvis` (Chang & Wickham 2016), `plotly` (Sievert 2018), and most recently, `ggvega` (Yang et al. 2019). Least well-defined in these grammars is usually the aspect of linked brushing and highlighting in plots (Becker & Cleveland 1987), a technique crucial to interactive graphics for defining and exploring subsets of the data. An additional problem for interactive graphics is reproducibility of a user's work. By definition, interactive graphics enable a flow, rather than a single static result. Very recent advances such as `trackr` (Becker et al. 2019) not only record the user's interactions with the data but also try to infer the user's intent by collecting metadata and automatically analyzing the structure of the data and the code. This approach might be used in testing to evaluate both different users' approaches and the tools used.

Another open question is the acceptance of results—much historical research on best practices exists, but how much of it is being put into practice? One does not have to look far to find astonishingly bad graphics in astonishingly good outlets. We will not point fingers here but will defer to online communities such as **https://www.reddit.com/r/dataisugly/** that have passionate discussions on the worst misuses of graphics. What can we do about this? It is up to statisticians to teach more graphics and better graphics in the classroom in the hope of slowly changing the climate. One incentive in the adoption of best practices might be that where there is a cost–benefit, such as when assessing business performance or a marketing campaign, there is scope for providing a measure on which to gauge effective visual communication. There is also significant room to improve plots in academic publications in order to better communicate research results (Unwin 2019).

In the 100 years of empirical evaluation of the perception and utility of statistical graphics, we have assembled a working knowledge of how to best create graphs that are easily read and understood. Once rare, charts are now everywhere we look—on the news, in papers and magazines, and online in interactive form. Going forward, we must do a better job of translating the academic research into practice, making it easier for academics and nonacademics alike to create useful, well-designed graphics.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Heike Hofmann and Susan Vanderplas are partially supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California–Irvine, and the University of Virginia.

LITERATURE CITED

Amer T. 2005. Bias due to visual illusion in the graphical presentation of accounting information. *J. Inf. Syst.* 19:1–18

Baird JC, Lewis C, Romer D. 1970. Relative frequencies of numerical responses in ratio estimation. *Percept. Psychophys.* 8:358–62

Becker G, Moore SE, Lawrence M. 2019. trackr: a framework for enhancing discoverability and reproducibility of data visualizations and other artifacts in R. *J. Comput. Graph. Stat.* 28:644–58

Becker RA, Cleveland WS. 1987. Brushing scatterplots. *Technometrics* 29:127–42

Becker RA, Cleveland WS, Wilks AR. 1987. Dynamic graphics for data analysis. *Stat. Sci.* 2:355–83

Beecham R, Dykes J, Meulemans W, Slingsby A, Turkay C, Wood J. 2017. Map LineUps: effects of spatial structure on graphical inference. *IEEE Trans. Vis. Comput. Graph.* 23:391–400

Beniger JR, Robyn DL. 1978. Quantitative graphics in statistics: a brief history. *Am. Stat.* 32:1–11

Bernstein EM, Cowden DJ. 1937. Graphic presentation of trend data. *South. Econ. J.* 3:443–51

Bertin J, Berg WJ. 1983. *Semiology of Graphics: Diagrams, Networks, Maps*, Vol. 1. Madison, WI: Univ. Wis. Press

Borland D, Ii RMT. 2007. Rainbow color map (still) considered harmful. *IEEE Comput. Graph. Appl.* 27:14–17

Br. Mus. 1882. The map of the world. *British Museum Collection Online*. **https://www.britishmuseum.org/research/collection_online/collection_object_details.aspx?assetId=404485001&objectId=362000&partId=1**

Brandes D. 1976. The present state of perceptual research in cartography. *Cartogr. J.* 13:172–76

Brewer CA. 2019. ColorBrewer2.0: color advice for cartography. *Software diagnostic tool for maps*. **http://colorbrewer2.org**

Brinton WC. 1917. *Graphic Methods for Presenting Facts*. New York: Eng. Mag. Co.

Broersma H, Molenaar IW. 1985. Graphical perception of distributional aspects of data. *Comput. Stat. Q.* 2:53–72

Buja A, Asimov D, Hurley C, McDonald JA. 1988. Statistical inference for exploratory data analysis and model diagnostics. In *Dynamic Graphics for Statistics*, ed. WS Cleveland, ME McGill, pp. 277–308. Belmont, CA: Wadsworth

Buja A, Cook D, Hofmann H, Lawrence M, Lee EK, et al. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. R. Soc. Lond. A* 367:4361–83

Callaghan TC. 1984. Dimensional interaction of hue and brightness in preattentive field segregation. *Percept. Psychophys.* 36:25–34

Carr DB. 1994. Using gray in plots. *Stat. Comput. Graph. Newsl.* 5:19–23

Carswell CM, Wickens CD. 1987. Information integration and the object display: an interaction of task demands and display superiority. *Ergonomics* 30:511–27

Chandar N, Collier D, Miranti P. 2012. Graph standardization and management accounting at AT&T during the 1920s. *Account. Hist.* 17:35–62

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. 2019. `shiny`: web application framework for R. *R package*, version 1.3.2. **https://shiny.rstudio.com**

Chang W, Wickham H. 2016. `ggvis`: interactive grammar of graphics. *R package*, version 0.4.3. **https://ggvis.rstudio.com/**

Cleveland WS, McGill R. 1984. Graphical perception: theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc.* 79:531–54

Cleveland WS, McGill R. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229:828–33

Cleveland WS, McGill R. 1987. Graphical perception: the visual decoding of quantitative information on graphical displays of data. *J. R. Stat. Soc. A* 150:192–229

Cooke L. 2010. Assessing concurrent think-aloud protocol as a usability test method: a technical communication approach. *IEEE Trans. Prof. Commun.* 53:202–15

Croxton FE. 1932. Graphic comparisons by bars, squares, circles, and cubes. *J. Am. Stat. Assoc.* 27:54–60

Croxton FE, Stryker RE. 1927. Bar charts versus circle diagrams. *J. Am. Stat. Assoc.* 22:473–82

Desnoyers L. 2011. Toward a taxonomy of visuals in science communication. *Tech. Commun.* 58:119–34

Dunbar K. 1995. How scientists really reason: scientific reasoning in real-world laboratories. *Nat. Insight* 18:365–95

Dunn R. 1988. Framed rectangle charts or statistical maps with shading: an experiment in graphical perception. *Am. Stat.* 42:123

Eells WC. 1926. The relative merits of circles and bars for representing component parts. *J. Am. Stat. Assoc.* 21:119–32

Eick SG, Wills GJ. 1995. High interaction graphics. *Eur. J. Oper. Res.* 81:445–59

Fabrikant SI, Hespanha SR, Hegarty M. 2010. Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making. *Ann. Assoc. Am. Geogr.* 100:13–29

Fienberg SE. 1979. Graphical methods in statistics. *Am. Stat.* 33:165–78

Flury B, Riedwyl H. 1988. *Multivariate Statistics: A Practical Approach*. London: Chapman & Hall

Funkhouser HG. 1937. Historical development of the graphical representation of statistical data. *Osiris* 3:269–404

Gelman A, Unwin A. 2013. Infovis and statistical graphics: different goals, different looks. *J. Comput. Graph. Stat.* 22:2–28

Goldberg JH, Helfman JI. 2010. Comparing information graphics: a critical look at eye tracking. In *Proceedings of the Third BELIV'10 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pp. 71–78. New York: ACM

Green TM, Fisher B. 2011. The personal equation of complex individual cognition during visual interface interaction. In *Human Aspects of Visualization*, ed. A Ebert, A Dix, ND Gershon, M Pohl, pp. 38–57. Berlin: Springer

Guan Z, Lee S, Cuddihy E, Ramey J. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '06*, ed. R Grinter, T Rodden, P Aoki, E Cutrell, R Jeffries, G Olson, pp. 1253–62. New York: ACM

Harms H. 1991. August Friedrich Wilhelm Crome (1753-1833). *Cartogr. Helv.* 33:33–38

Hasanhodzic J, Lo AW, Viola E. 2010. Is it real, or is it randomized?: A financial Turing test. arXiv:1002.4592 [q-fin.GN]

Healey CG, Booth KS, Enns JT. 1996. High-speed visual estimation using preattentive processing. *ACM Trans. Comput. Hum. Interact.* 3:107–35

Healey CG, Enns JT. 1999. Large datasets at a glance: combining textures and colors in scientific visualization. *IEEE Trans. Vis. Comput. Graph.* 5:145–67

Heer J, Bostock M. 2010. Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–12. New York: ACM

Hofmann H, Follett L, Majumder M, Cook D. 2012. Graphical tests for power comparison of competing designs. *IEEE Trans. Vis. Comput. Graph.* 18:2441–48

Hughes BM. 2001. Just noticeable differences in 2D and 3D bar charts: a psychophysical analysis of chart readability. *Percept. Motor Skills* 92:495–503

Javed W, McDonnel B, Elmqvist N. 2010. Graphical perception of multiple time series. *IEEE Trans. Visualization Comput. Graph.* 16:927–34

Karsten K. 1923. *Charts and Graphs: An Introduction to Graphic Methods in the Control and Analysis of Statistics.* Upper Saddle River, NJ: Prentice-Hall

Kibirige H. 2017. `plotnine`: a grammar of graphics for Python. *Graphical software*. **https://plotnine.readthedocs.io/en/stable/**

Kim S, Lombardino LJ, Cowles W, Altmann LJ. 2014. Investigating graph comprehension in students with dyslexia: an eye tracking study. *Res. Dev. Disabil.* 35:1609–22

Kirschenbaum SS. 2003. Comparative cognitive task analysis: the cognition of weather forecasting. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 47:473–77

Kostelnick C. 2016. The re-emergence of emotional appeals in interactive data visualization. *Tech. Commun.* 63:116–35

Kruskal W. 1977. Visions of maps and graphs. In *Proceedings of the International Symposium on Computer-Assisted Cartography (Auto-Carto II)*, ed. J Kavaliunas, pp. 27–36. Washington, DC: US Dep. Commer.
Lawrence M, Wickham H, Cook D, Hofmann H, Swayne DF. 2009. Extending the GGobi pipeline from R: rapid prototyping of interactive visualizations. *Comput. Stat.* 24:195–205
Legge GE, Gu Y, Luebker A. 1989. Efficiency of graphical perception. *Percept. Psychophys.* 46:365–74
Lewandowsky S, Spence I. 1989. The perception of statistical graphs. *Sociol. Methods Res.* 18:200–42
Light A, Bartlein PJ. 2004. The end of the rainbow? Color schemes for improved data graphics. *Eos Trans. Am. Geophys. Union* 85:385–91
Loy A, Follett L, Hofmann H. 2016. Variations of Q-Q plots: the power of our eyes! *Am. Stat.* 70:202–14
Loy A, Hofmann H. 2015. Are you normal? The problem of confounded residual structures in hierarchical linear models. *J. Comput. Graph. Stat.* 24:1191–209
Lumley T. 2013. `dichromat`: color schemes for dichromats. *R package*, version 2.0-0. **https://cran.r-project.org/web/packages/dichromat/index.html**
MacDonald-Ross M. 1977. How numbers are shown: a review of research on the presentation of quantitative data in texts. *AV Commun. Rev.* 25:359–409
Majumder M, Hofmann H, Cook D. 2013. Validation of visual statistical inference, applied to linear models. *J. Am. Stat. Assoc.* 108:942–56
McDonald L. 2014. Florence Nightingale, statistics and the Crimean War. *J. R. Stat. Soc. A* 177:569–86
Morel P. 2018. Gramm: grammar of graphics plotting in Matlab. *J. Open Source Softw.* 3:568
Netzel R, Vuong J, Engelke U, O'Donoghue S, Weiskopf D, Heinrich J. 2017. Comparative eye-tracking evaluation of scatterplots and parallel coordinates. *Vis. Inform.* 1:118–31
Normand MP, Bailey JS. 2006. The effects of celeration lines on visual data analysis. *Behav. Modif.* 30:295–314
North C. 2006. Toward measuring visualization insight. *IEEE Comput. Graph. Appl.* 26:6–9
Okan Y, Galesic M, Garcia-Retamero R. 2016. How people with low and high graph literacy process health graphs: evidence from eye-tracking. *J. Behav. Decis. Making* 29:271–94
Peterson LV, Schramm W. 1954. How accurately are different kinds of graphs read? *Audio Vis. Commun. Rev.* 2:178–89
Peuchet J, Gilbert C. 1805. *Statistique élémentaire de la France*. Paris: Chez Gilbert Cie.
Playfair W. 1801. *The Statistical Breviary: Shewing the Resources of Every State and Kingdom in Europe*. London: J. Wallis
Roy Chowdhury N, Cook D, Hofmann H, Majumder M, Lee EK, Toth AL. 2015. Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Comput. Stat.* 30:293–316
Ryu YS, Yost B, Convertino G, Chen J, North C. 2003. Exploring cognitive strategies for integrating multiple-view visualizations. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 47:591–95
Satyanarayan A, Moritz D, Wongsuphasawat K, Heer J. 2017. Vega-Lite: a grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graph.* 23:341–50
Shah P, Carpenter PA. 1995. Conceptual limitations in comprehending line graphs. *J. Exp. Psychol. Gen.* 124:43–61
Shah P, Miyake A. 2005. *The Cambridge Handbook of Visuospatial Thinking*. Cambridge, UK: Cambridge Univ. Press
Sievert C. 2018. `plotly`: create interactive web graphics via 'plotly.js'. *R package*, version 4.9.0. **https://cran.r-project.org/web/packages/plotly/index.html**
Simons DJ, Levin DT. 1997. Change blindness. *Trends Cogn. Sci.* 1:261–67
Smith CD. 1996. Imago Mundi's logo the Babylonian map of the world. *Imago Mundi* 48:209–11
Spence I. 1990. Visual psychophysics of simple graphical elements. *J. Exp. Psychol. Hum. Percept. Perform.* 16:683–92
Swayne D, Klinke S. 1999. Introduction to the special issue on interactive graphical data analysis: What is interaction? *Comput. Stat.* 14:1–6
Tan JK. 1994. Human processing of two-dimensional graphics: information-volume concepts and effects in graph-task fit anchoring frameworks. *Int. J. Hum. Comput. Interact.* 6:414–56
Teghtsoonian M. 1965. The judgment of size. *Am. J. Psychol.* 78:392–402

Tomasetti N. 2015. *Comparing the power of plot designs to reveal correlation.* Honors Thesis, Fac. Bus. Econ., Monash Univ., Melbourne, Aust. **https://github.com/dicook/lineplots_v_scatterplot**

Trafton GJ, Kirschenbaum SS, Tsui TL, Miyamoto RT, Ballas JA, Raymond PD. 2000. Turning pictures into numbers: extracting and generating information from complex visualizations. *Int. J. Hum. Comput. Stud.* 53:827–50

Treisman AM. 1980. A feature-integration theory of attention. *Cogn. Psychol.* 12:97–136

Trickett SB, Fu WT, Schunn CD, Trafton JG. 2000a. From dipsy-doodles to streaming motions: changes in representation in the analysis of visual scientific data. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 22, ed. LR Gleitman, AK Joshi, pp. 959–64. Ann Arbor, MI: Cogn. Sci. Soc.

Trickett SB, Trafton JG, Schunn CD. 2000b. Blobs, dipsy-doodles and other funky things: Frame-work anomalies in exploratory data analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 22, ed. LR Gleitman, AK Joshi, pp. 965–70. Ann Arbor, MI: Cogn. Sci. Soc.

Tufte E. 1991. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press. 2nd ed.

Tukey JW. 1965. The technical tools of statistics. *Am. Stat.* 19:23–28

Tukey JW. 1972. Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T Bancroft, pp. 293–316. Ames, IA: Iowa State Univ. Press

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. 2011. Accuracy and reliability of forensic latent fingerprint decisions. *PNAS* 108:7733–38

Unwin A. 1999. Requirements for interactive graphics software for exploratory data analysis. *Comput. Stat.* 14:7–22

Unwin A. 2019. Why is data visualization important? What is important in data visualization? *Harv. Data Sci. Rev.* In press

Vanderplas S, Goluch R, Hofmann H. 2019. Framed! Reproducing and revisiting 150-year-old charts. *J. Comput. Graph. Stat.* 28:620–34

Vanderplas S, Hofmann H. 2015. Signs of the sine illusion–why we need to care. *J. Comput. Graph. Stat.* 24:1170–90

Vanderplas S, Hofmann H. 2017. Clusters beat trend!? Testing feature hierarchy in statistical graphics. *J. Comput. Graph. Stat.* 26:231–42

von Huhn R. 1927. Further studies in the graphic use of circles and bars. I. A discussion of the Eells' experiment. *J. Am. Stat. Assoc.* 22:31–39

Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, et al. 2012a. A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol. Bull.* 138:1172–217

Wagemans J, Feldman J, Gepshtein S, Kimchi R, Pomerantz JR, et al. 2012b. A century of gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138:1218–52

Wakita K, Shimamura K. 2005. SmartColor: disambiguation framework for the colorblind. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 158–65. New York: ACM

Walker FA. 1874. *Statistical Atlas of the United States, Based on the Results of the Ninth Census, 1870, with Contributions from Many Eminent Men of Science, and Several Departments of the Government*. New York: Bien

Wickham H. 2010. A layered grammar of graphics. *J. Comput. Graph. Stat.* 19:3–28

Wickham H. 2013. Graphical criticism: some historical notes. *J. Comput. Graph. Stat.* 22:38–44

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer

Wickham H, Chowdhury NR, Cook D, Hofmann H. 2018. `nullabor`: tools for graphical inference. *R package*, version 0.3.5. **https://cran.r-project.org/web/packages/nullabor/index.html**

Wickham H, Cook D, Hofmann H, Buja A. 2010. Graphical inference for infovis. *IEEE Trans. Vis. Comput. Graph.* 16:973–79

Wickham H, Grolemund G. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol, CA: O'Reilly. 1st ed.

Wickham H, Lawrence M, Cook D, Buja A, Hofmann H, Swayne DF. 2009. The plumbing of interactive graphics. *Comput. Stat.* 24:207–15

Widen HM, Elsner JB, Pau S, Uejio CK. 2016. Graphical inference in geographical research. *Geogr. Anal.* 48:115–31

Wilkinson L. 1999. *The Grammar of Graphics*. New York: Springer
Woller-Carter MM, Okan Y, Cokely ET, Garcia-Retamero R. 2012. Communicating and distorting risks with graphs: an eye-tracking study. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 56:1723–27
Xie Y, Hofmann H, Cheng X. 2014. Reactive programming for interactive graphics. *Stat. Sci.* 29:201–13
Yang W, Jeppson H, Lyttle IJ. 2019. `ggvega`: translator from 'ggplot2' to 'Vega-Lite'. *R package*, version 0.0.0.9001. **https://github.com/vegawidget/ggvega**
Yates J. 1985. Graphs as a managerial tool: a case study of Du Pont's use of graphs in the early twentieth century. *J. Bus. Commun.* 22:5–33
Zeileis A, Hornik K, Murrell P. 2009. Escaping RGBland: selecting colors for statistical graphics. *Comput. Stat. Data Anal.* 53:3259–70
Zgraggen E, Zhao Z, Zeleznik R, Kraska T. 2018. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing*, pap. no. 479. New York: ACM

Annual Review of Statistics and Its Application

Volume 7, 2020

Contents

Statistical Significance
D.R. Cox 1

Calibrating the Scientific Ecosystem Through Meta-Research
Tom E. Hardwicke, Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P.A. Ioannidis 11

The Role of Statistical Evidence in Civil Cases
Joseph L. Gastwirth 39

Testing Statistical Charts: What Makes a Good Graph?
Susan Vanderplas, Dianne Cook, and Heike Hofmann 61

Statistical Methods for Extreme Event Attribution in Climate Science
Philippe Naveau, Alexis Hannart, and Aurélien Ribes 89

DNA Mixtures in Forensic Investigations: The Statistical State of the Art
Julia Mortera 111

Modern Algorithms for Matching in Observational Studies
Paul R. Rosenbaum 143

Randomized Experiments in Education, with Implications for Multilevel Causal Inference
Stephen W. Raudenbush and Daniel Schwartz 177

A Survey of Tuning Parameter Selection for High-Dimensional Regression
Yunan Wu and Lan Wang 209

Algebraic Statistics in Practice: Applications to Networks
Marta Casanellas, Sonja Petrović, and Caroline Uhler 227

Bayesian Additive Regression Trees: A Review and Look Forward
Jennifer Hill, Antonio Linero, and Jared Murray 251

Q-Learning: Theory and Applications
Jesse Clifton and Eric Laber 279

Representation Learning: A Statistical Perspective
Jianwen Xie, Ruiqi Gao, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu 303

Robust Small Area Estimation: An Overview
Jiming Jiang and J. Sunil Rao 337

Nonparametric Spectral Analysis of Multivariate Time Series
Rainer von Sachs 361

Convergence Diagnostics for Markov Chain Monte Carlo
Vivekananda Roy 387

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics