The perception of statistical graphics

by

Susan Ruth VanderPlas

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Heike Hofmann, Major Professor

Di Cook

Sarah Nusser

Max Morris

Stephen Gilbert

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 0. INTRODUCTION

There has been quite a bit of research on statistical graphics and visualization, generally focused on new types of graphics, new software to create graphics, interactivity, and usability studies. Our ability to interpret and use statistical graphics hinges on the interface between the graph itself and the brain that perceives and interprets it, and there is substantially less research on the interplay between graph, eye, brain, and mind than is sufficient to understand the nature of these relationships.

The goal of the work presented here is to further explore the interplay between a static graph, the translation of that graph from paper to mental representation (the journey from eye to brain), and the mental processes that operate on that graph once it is transferred into memory (mind). Understanding the perception of statistical graphics should allow researchers to create more effective graphs which produce fewer distortions and viewer errors while reducing the cognitive load necessary to understand the information presented in the graph.

Chapter 1 contains a review of past literature which is relevant to the encoding and memory of statistical graphics, encompassing studies from psychology, psychophysics, meterology, business, and statistics. Chapter 2 disucsses an optical illusion which is frequently present in even simple statistical graphics, but is incredibly difficult to resolve without altering the graph itself. This illusion (and its relative obscurity in the literature on graph perception) serves as a reminder that our graphs are only useful if they are designed with the perceptual system in mind. Chapter 4 contains an outline of a study examining the relationship between spatial reasoning abilities and graph perception; this study is currently in progress and should provide useful information about the skills necessary to read statistical graphics. Finally, Chapter 5 discusses a study which seeks to understand visually salient features of a plot, to aid researchers in creating graphics which are constructed to efficiently convey information visually.

0.1 Literature Review

The literature review encompasses many different areas of interdisciplinary research; to ensure that there is a reference for some of the more niche vocabulary in this dissertation, it begins with the physiology of the eye and includes a summary of some basic neuropsychology as well. From there, I discuss some of the psychology and psychophysics research that concerns visual perception, memory, and attention. This research roughly falls in the domain of 'cognitive psychology', but includes studies from fields as diverse as meterology and business. After establishing the underlying mechanisms of perception, I then cover some of the research which focuses specifically on statistical graphics, and discuss some of the methodology used in these experiments. I expect that this section will continue to expand during the process of writing the other chapters of this dissertation; as the literature is fragmented across many different subject areas and spans 75 years, it is difficult to ensure that this is a comprehensive summary of the field as it stands today.

0.2 The Sine Illusion

Chapter 2 documents an illusion known as the 'line-width illusion' or the 'sine illusion' which occurs with some frequency in statistical graphics. It discusses the illusion, experimental evidence that the illusion exists even in simple graphics, and the implications of the illusion for statisticians. The presence of this illusion (and the solution for reducing the illusion's effects) serve as an introduction and case study for whether we can actually trust the things we see; the remainder of this dissertation aims to quantify variables which may affect how we answer the simple question "What am I seeing?" in relation to statistical graphics.

0.3 Visual Reasoning

The use of statistical graphics is encouraged because graphics summarize statistical models and results in a form that is easy for most people to understand. The lineup protocol provides a convenient way to compare different types of graphics which display the same data, but results from comparative studies(Hofmann et al., 2012) have demonstrated that individuals are highly

variable in their ability to identify the target plot in a lineup. This study aims to explore the association between spatial reasoning, pattern recognition, and the ability to identify a target plot in a lineup successfully. In order to assess spatial reasoning, several tests from the Factor-Referenced Cognitive Test battery (Ekstrom et al., 1976) which are associated with tasks similar to those utilized to identify a target plot in a lineup were assembled, along with lineups tested in (Hofmann et al., 2012) and a visual search task similar to the lineup task. This study explores the similarities and differences between spatial reasoning tasks and lineup tasks.

0.4 Hierarchy of Graphical Features

Cleveland and McGill (1984) created a hierarchy of graphical tasks which have informed graph design for the past 30 years, ranking numerical estimations of graphical features by accuracy. Similarly, Healey and Enns (1999) found that there is a hierarchy to preattentive perception of graphical features, in that color and texture can interfere under certain circumstances. Preattentive perception is not particularly applicable to statistical graphics, as most viewers spend more than a second looking at the image, but it would be useful to understand what features in a graph will dominate a viewer's perception: if linear trend information contradicts coloring of points, which feature will a viewer take away from the image? This experiment uses the lineup protocol to examine the features which are most salient to viewers, using pairwise comparisons of shape, color, linear trend, and outliers.

Taken together, these experiments should lay a foundation for exploring the perception of statistical graphics. There has been considerable research into the accuracy of numerical judgments viewers make from graphs, and these studies are useful, but it is more effective to understand how errors in these judgments occur so that the root cause of the error can be addressed directly. Understanding how visual reasoning relates to the ability to make judgments from graphs allows us to tailor graphics to particular target audiences. In addition, understanding the hierarchy of salient features in statistical graphics allows us to clearly communicate the important message from data or statistical models by constructing graphics which are designed specifically for the perceptual system.

CHAPTER 1. LITERATURE REVIEW

Statisticians produce graphics for a multitude of reasons: to understand the structure of raw data, to check model assumptions, to present model predictions in an informative fashion; all of these goals are facilitated by appropriate visualizations. Some of these goals are best served by quick-and-dirty representations of the data, while highly polished graphics may be more useful for other goals. Regardless, it is important to convey the data appropriately, which means we must understand how graphics are perceived on at least a basic level. Tukey focused on graphics as a tool for exploratory analysis, because pictures can often display the data in a more coherent fashion than a table or spreadsheet; his book, Tukey (1977) details many different types of graphics and the appropriate situations to utilize them. This chapter focuses on the interaction between graphics and the human visual system, with the goal of understanding how to most efficiently and effectively convey information using statistical graphics. We will first consider the physiology and psychology of the perceptual experience in general, and then address the issue of graph-specific perception.

1.1 The Human Visual System

In order to design graphics for the human perceptual system, we must understand, at a basic level, the makeup of the perceptual system. There are multiple levels of perception that must correctly function in order to perceive visual stimuli successfully, but a somewhat simplistic higher-level analogy would be that we must understand both the hardware and software of the human visual system to create effective graphics.

The "hardware", in this analogy, consists of the neurons that make up the eyes, optic nerve, and the brain itself. The higher-level functions (object recognition, working memory,

etc.) comprise the "software" component. In addition, much like computer software, there are different programs running simultaneously; these programs may interact with each other, run sequentially, or run in parallel. The following sections provide an overview of the grey-matter (hardware) components of the visual system as well as the higher-level cognitive heuristics (software) that order the raw input and construct our visual environment.

1.1.1 Hardware

The physiology of perception is complex; what follows is a brief overview of the physiology of perception, focusing on the areas most important to the perception of statistical graphics. This physiological information is important in understanding the difference between the sensation (i.e. the retinal image) and the perception (the corresponding mental representation), which is an important distinction in understanding how statistical graphics are perceived. This overview will entirely ignore the finer details of the organization of the brain: feature detector cells, specific processing units for certain types of visual stimuli, and most of the horrifying or amusing experiments and incidents that led to the understanding of the brain. A more thorough presentation of these aspects of perception can be found in Goldstein (2009b).

The Eye The eye is a complex apparatus, but for our purposes, the most important component of the eye is the retina, which contains the sensory cells responsible for transforming light waves into electrical information in the form of neural signals. These sensory cells are specialized neurons, known as rods and cones, which perceive light intensity (brightness) and wavelength (color), respectively. One section of the retina, known as the fovea, contains only cones; the rest of the retina contains a mixture of rods and cones. Figure 1.1 depicts the structure of the eye with a closeup of the retina.

Another important region of the retina is the blindspot, the area where the optic nerve exits the eye to connect the retina to the brain. There are no rods or cones in this region of the retina, and any vision in the region of space that maps onto this point is a result of two mechanisms: binocular vision (the other eye fills in the missing information) and your brain "filling in" what it believes should be there.

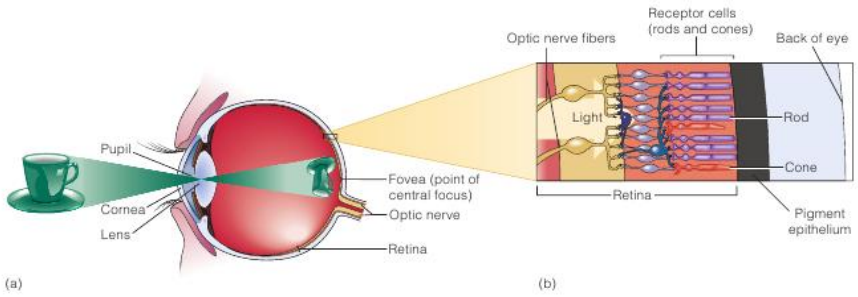


Figure 1.1: The human eye, with closeup of receptor cells in the retina (image from Goldstein 2009b, chap 3.1).

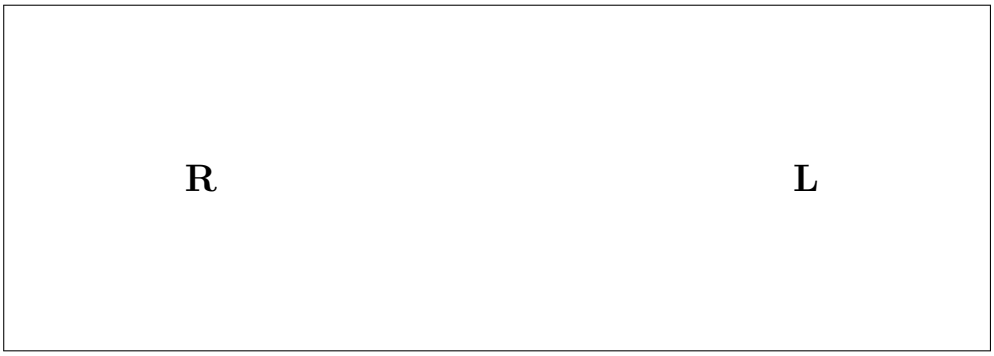


Figure 1.2: Illustration of the blind spot in the retina. Close one eye, and focus the other on the appropriate letter (R if the right eye is still open, L if the left eye is still open). Place the paper approximately 1 foot from your eye, and move the paper toward or away from your face until you notice the other letter disappear. Your brain "fills in" the other letter with the background.

Figure 1.3 shows the responsiveness of rods and each of the three types of cones to wavelengths of light in the visual spectrum. This image suggests that we have relatively good visual discrimination of the yellow-green portion of the color spectrum, but relatively poor discrimination of colors in the red and blue portions of the color spectrum.

As a result, rainbow-style color schemes are seldom appropriate for conveying numerical values, because the correspondance between the perceived information and the displayed information is not accurately maintained by the visual system (Treinish and Rogowitz, 2009). In addition, if any of the cones are missing or damaged as a result of genetic mutations, color perception is impaired, resulting in a smaller range of distinguishable colors. This set of impairments is known coloquially as color-blindness, and occurs in an estimated 5% of the population

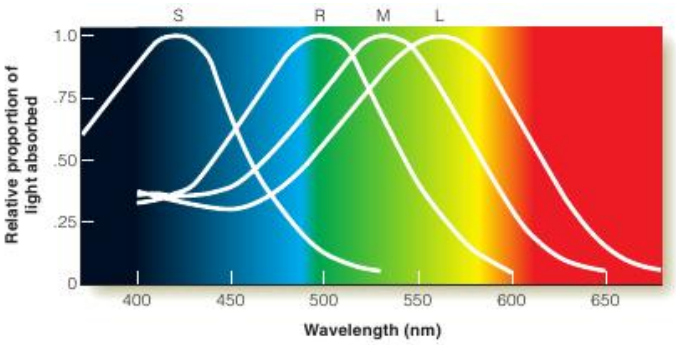
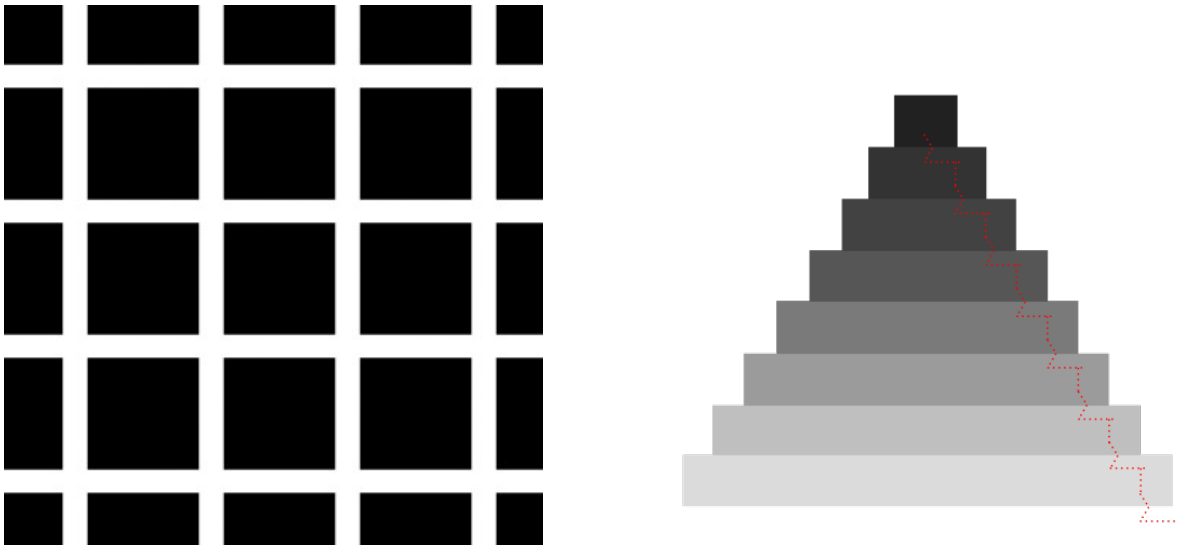


Figure 1.3: Absorption spectra of rods and short, medium, and long wave cones. (image from Goldstein 2009b, chap 3.3).

(approximately 10% of males, and less than 1% of females). Color blindness is discussed in more depth in section 1.1.3.2.

The Brain Once light hits the retina and causes a signal in the receptor cells, the information travels along the optic nerve and into the brain. Multiple neighboring rods are connected to the same neuron, where each cone is connected to a single neuron. The combined wiring of rod cells is responsible for the Hermann grid illusion and the Mach bands seen in figure 1.4. Both of these illusions are a product of lateral inhibition, which is a result of the wiring of rod cells in the retina. Essentially, neurons can only fire at a specific rate, so when neighboring cells are all stimulated simultaneously, the combined neuron cannot fire fast enough to pass on all of the signals, causing "inhibition". The specifics of this response and its relationship with the wiring of the receptor cells are somewhat complex; a more thorough explanation can be found in Goldstein (2009b), chapter 3.4.

Once neural impulses have left the retina through the optic nerve, they travel to the visual cortex by way of several specialized structures within the brain that process lower-level signals. Receptor cells in the visual cortex respond to specific angles, spatial locations, colors, and intensities, and arrays of these special 'feature detector cells' process the information into a form that higher-level processes can utilize. These higher-level processes are what we have previously called 'software': they are not directly related to the physical brain, but they do process information heuristically to produce higher-level reasoning and conclusions. In the next section, we explore some of the higher-level processes responsible for visual perception.



(a) Hermann Grid Illusion

(b) Mach Bands.

Figure 1.4: Optical Illusions resulting from lateral inhibition. The Hermann Grid illusion causes dark circles to appear at the intersection of white lines; the Mach bands illusion causes the borders of adjacent rectangles to appear more strongly defined.

1.1.2 Software

Many of the processes for visual perception run simultaneously; in absence of a strict temporal ordering, we will start with the more basic tasks of visual perception and proceed towards higher-level processes. We will begin with attention.

1.1.2.1 Attention and Perception

In many tasks, it is necessary to pay attention to many parallel input streams simultaneously; this is particularly true for complex tasks like driving a car. These tasks demand divided attention; the brain must process many different sources of information in parallel. By contrast, most image recognition tasks require selective attention, that is, focusing on specific objects and ignoring everything else. The brain accomplishes this attention through several mechanisms.

Selective attention is accomplished by focusing the fovea (the area with the highest visual acuity) on the object. For instance, if the object is a page of text, each word will pass through the fovea, producing a focused stream of visual input. This stream of input consists of saccades

(jumps between points of focus) and pauses in which the visual information is relayed to the brain. Figure 1.5 shows the saccades (lines) and pauses (circles) resulting when someone scans a paragraph of text. These saccades and pauses are utilized in eye-tracking technology to determine which parts of an image the observer is focusing on (and by extension, which information is being encoded by the brain). Further discussion of eye tracking is provided in section 1.3.2.1.

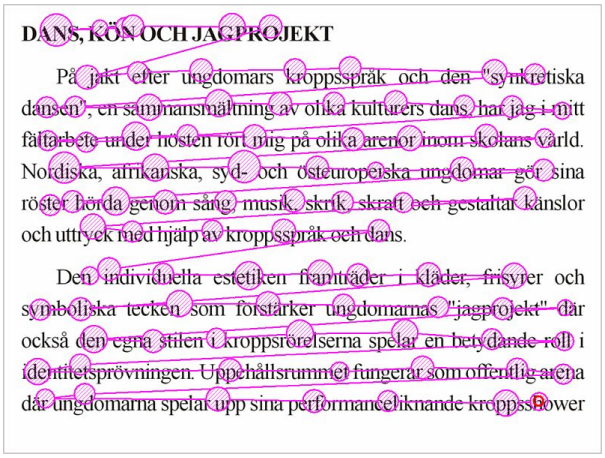


Figure 1.5: A plot of saccades made while reading text. Saccades, shown by the lines, indicate "jumps", while pauses are shown by circles, with size proportional to the time spent focusing on that area.

Selective attention is generally necessary for perception to occur, though there is some information that is encoded automatically. Experiments such as the fairly famous "gorilla" film[1] demonstrate that even when there is attention focused on a task, information extraneous to that task is not always encoded, that is, even when participants focused on counting the number of passes between players in the basketball game, they did not notice the gorilla walking through the middle of the court. It is important to understand which parts of a visual stimulus are the focus of a given perceptual task, because most of the information encoded by the brain is a result of selective attention. Eye-tracking can be an important tool useful to understand these perceptual processes, but participants are often able to report which parts of a stimulus contributed to their decision as well.

Within the brain, attention is important because it allows different regions of the brain which

[1] http://www.theinvisiblegorilla.com/videos.html

process color, shape, and position to integrate these perceptions into a multifaceted mental representation of the object (Goldstein, 2009b). This process, known as binding, is essential to coherently encode a scene into working memory. Feature integration theory (Treisman and Gelade, 1980) suggests that these separate streams of information are initially encoded in the preattentive stage of object perception; focusing on the object triggers the binding of these separate streams into a single coherent stream of information. Many single features, such as color, length, and texture are preattentive, because they can be pinpointed in an image without focused attention (and thus can be located faster), but specific combinations of color and shape require attention (because the features must be bound together) and are thus more difficult to search. Preattentive features are generally processed in parallel (that is, the entire scene is processed nearly simultaneously), while features requiring attention are processed serially. Examples of features processed serially and in parallel are shown in figure 1.6, taken from Chapter 6 of Helander et al. (1997). The importance of preattentive processing to statistical graphics is discussed in Section 1.2.1, which includes a demonstration of preattentive search in figure 1.16.

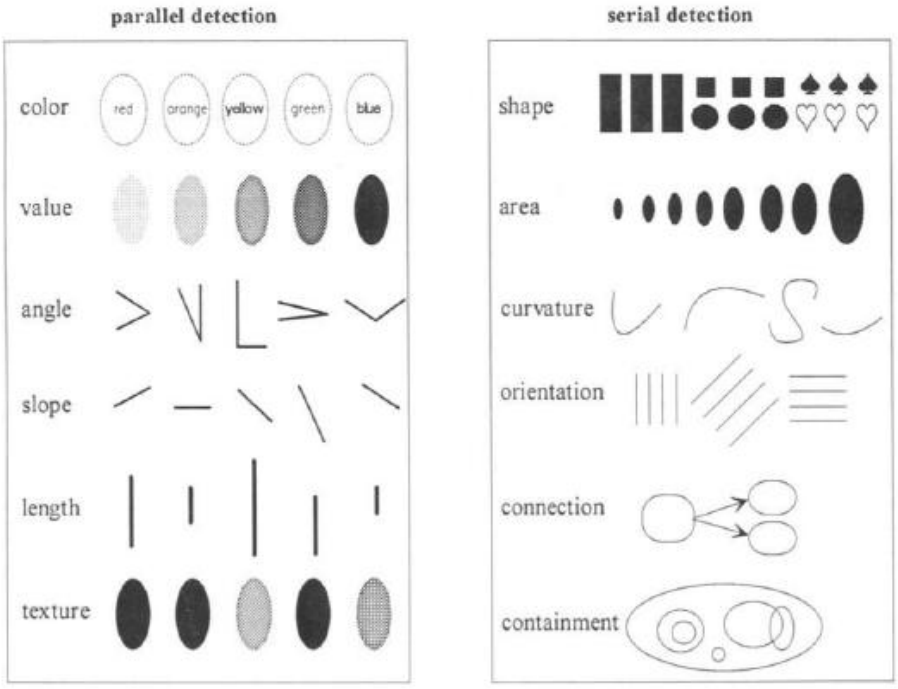


Figure 1.6: Examples of features detected serially or in parallel (Chapter 6, Helander et al. 1997)

Feature integration as a result of attention enables the brain to process a figure holistically and integrate all of the separate aspects of the object into a single perceptual experience. This processing is important for the most basic visual processes we take for granted, including object perception.

1.1.2.2 Object Perception

The most basic task of the visual system is to perceive objects in the world around us. This is an inherently difficult task, however, because the retina is a flat, two-dimensional surface responsible for conveying a three-dimensional visual scene. This dimensional reduction means that there are multiple three-dimensional stimuli that can produce the same visual image on the retina. This is known as the inverse projection problem - an infinite number of three-dimensional objects produce the same two-dimensional image. Less relevant to statistical graphics, but still complicating the object perception process, a single object can be viewed from a multitude of angles, in many different situations which may affect the retinal image (lighting, partial obstruction, etc). In addition, we recognize objects even when they are partially obscured or viewed from an angle we have not previously seen. These problems mean that the brain must utilize many different heuristics to increase the accuracy of the perceived world relative to an ambiguous stimulus.

The most commonly cited set of heuristics for object perception (and the set most relevant to statistical graphics) are known as the Gestalt Laws of Perceptual Organization (Goldstein 2009b, Chapter 5.2). These laws are related to the idea "the whole is greater than the sum of the parts", that is, that the components of a visual stimulus, when combined, create something that is more meaningful than the separate components considered individually. The Gestalt laws are as follows (Goldstein, 2009b):

- **Pragnanz - the law of good figure**. (Also referred to as the law of closure) Every stimulus pattern is seen so that the resulting structure is as simple as possible.
- **Proximity**. Things that are close in space appear to be grouped.

- **Similarity**. Similar items appear to be grouped together. The law of similarity is usually subordinate to the law of proximity.
- **Good Continuation**. Points that can be connected to form straight lines or smooth curves seem to belong together, and lines seem to follow the smoothest path.
- **Common Fate**. Things moving in the same direction are part of a single group.
- **Familiarity**. Things are more likely to form groups if the groups are familiar.
- **Common Region**. Things that are in the same region (container) appear to be grouped together
- **Uniform Connectedness**. A connected region of objects is perceived as a single unit.
- **Synchrony**. Events occurring at the same time will be perceived as belonging together.

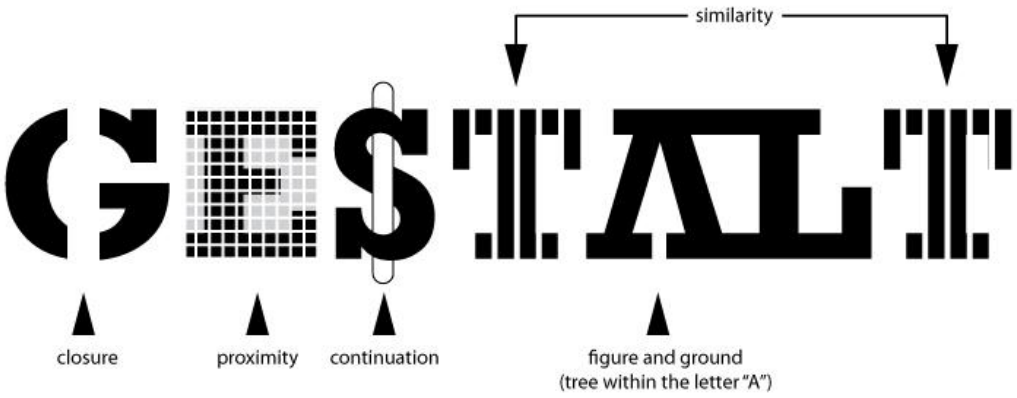


Figure 1.7: The gestalt laws of perception [2]

Figure 1.7 shows examples of many of the gestalt laws, which when combined help to order our perceptual experience. We have discussed how we attend to visual stimuli and how they are recognized, but for the perceptual experience to be meaningful, previous experiences must be stored into memory in some coherent way. The next section discusses how visual scenes are stored into long-term memory.

[2] From http://yusylvia.files.wordpress.com/2010/03/gestalt_illustration-01.jpg

1.1.2.3 Visual Memory

We have discussed how visual stimuli are perceived and how objects are recognized; we now must examine how visual stimuli are encoded into memory. Most researchers believe that visual perceptions are encoded in an analog fashion, so that the memory of an image is closely related to the perception of that same image (Matlin, 2005). Other theories suggest that visual perceptions are encoded semantically, that is, the description of a visual scene would be encoded, rather than a mental "image" of that scene. Both theories are likely at least partially correct, but the analog encoding of visual images is more relevant to statistical graphics because the accuracy of the stored image has the potential to affect recall of the contents of that image (and thus what people remember about a particular graphic). Experimental evidence for analog encoding includes the mental rotation task, where participants must determine whether or not a figure is a rotation of a target figure, as shown in figure 1.8. Shepard and Metzler (1988) showed that reaction time was proportional to the angle of rotation of the stimuli, which suggests that participants were mentally rotating the figure as they would rotate a three-dimensional figure in space.

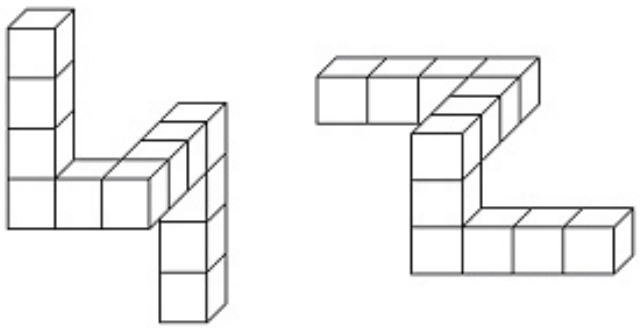


Figure 1.8: Rotation task (Shepard and Metzler, 1988). Are the two images the same?

In addition, Kosslyn et al. (1978) showed that mental representation of distances in a figure are accurate and that the time to encode those distances is proportional to the distances in the actual figure. These studies suggest that the memory of an image (statistical graphic or otherwise) is a reasonably accurate facsimile of the original image (though the accuracy of the mental representation is of course likely to be moderated by attention and recall ability).

Another facet of visual memory that will be important to understanding perception and

memory of statistical graphics is that the "gist" of an image is stored along with the image. In these cases, recall ability is more consistent with the semantic encoding of images; that is, when shown an ambiguous figure and immediately asked for a description, participants could not give an alternate interpretation of the figure after the experiment was complete. In the case of figure 1.9, participants who initially said the figure was a duck did not report having viewed a picture of a rabbit after the experiment, even though the image is consistent with either interpretation. This suggests that in some cases, verbal encoding of a figure (i.e. describing it as a duck) disrupts the mental representation of the picture. This is common in other types of memories as well: when the gist of a passage is stored, the actual content of the passage is no longer accessible. In other words, we would expect that if someone had to interpret a graph, they would remember the interpretation much more strongly than the actual graph, even if that interpretation was incorrect or incomplete.

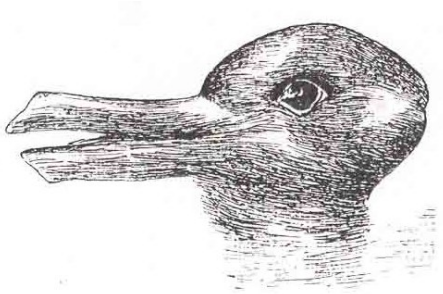


Figure 1.9: An ambiguous image that could be either a rabbit or a duck. When participants were asked to identify the image initially, they could not provide an alternate interpretation of the figure later.

The "software" of the visual system is of course more complex than the few modules listed here, but understanding attention, object perception, and how images are stored for later retrieval in the brain will make designing statistical graphics for the visual system easier and will also help with evaluating graphics based on the capabilities of the human visual system.

We have discussed the neural hardware of the visual system and some of the higher-level software that contributes to our ability to create and understand meaning in the world around us. Occasionally, our highly tuned perceptual system fails in unusual ways due to the heuristics and algorithms that were optimized for operation in a three-dimensional world where the main

tasks were hunting, gathering, and avoiding predators. The next section will examine three interesting results of this tuning that are important to the design of statistical graphics as we transition from the psychophysics and cognitive psychology literature to statistics and human-computer interaction literature.

1.1.3 Bugs and Peculiarities of the Visual System

1.1.3.1 Logarithmic perception

One of the earliest psychophysics researchers, Ernst Weber, discovered that the difference threshold, the smallest detectable difference between two sensory stimuli, increased proportionately with the magnitude of the stimulus. This statement, known now as Weber's Law, holds true for a large range of intensities of a number of senses. Numerically, Weber's Law is stated as

$$\frac{\triangle S}{S} = K \tag{1.1}$$

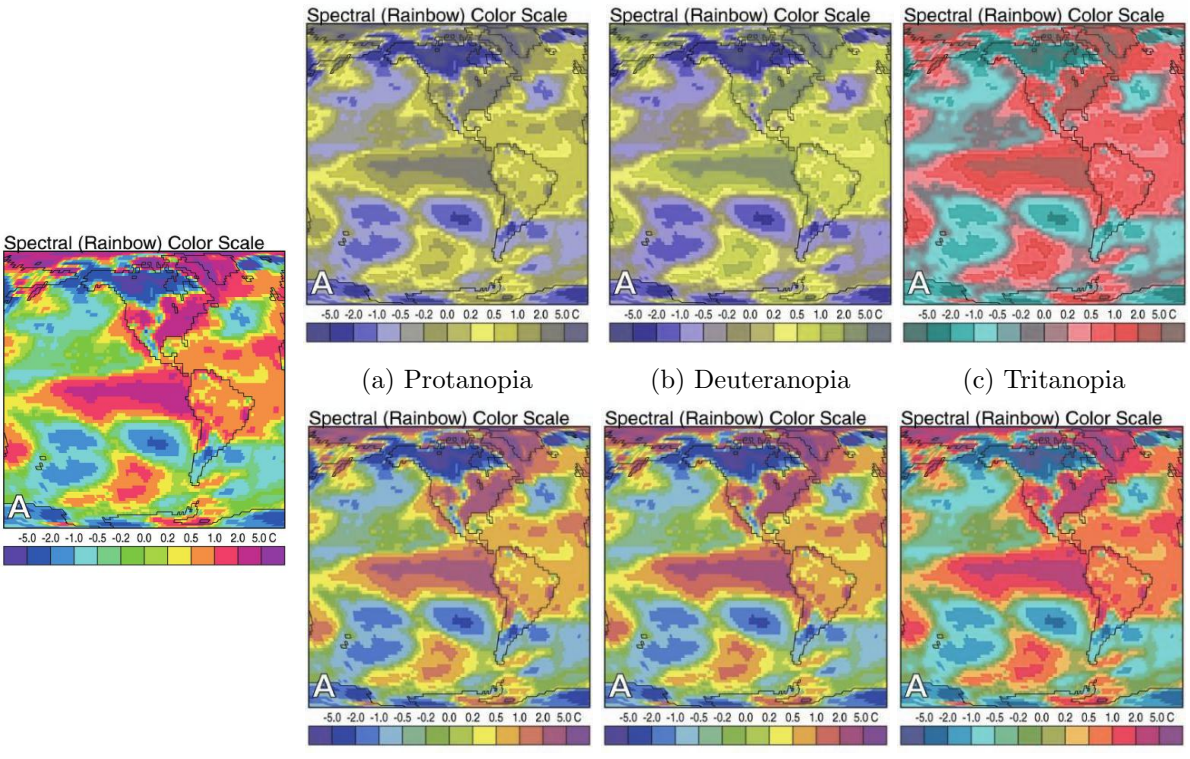
where K is a constant called the Weber fraction, S is the value of the standard stimulus, and $\triangle S$ is the difference between the standard stimulus and the test stimulus. So if a participant is given a 100-g weight and a 102-g weight and can just barely tell the difference between the two, then $K = 0.02$ and we would assume that the the difference between a 200-g weight and a 204-g weight would be just barely detectable as well (Chapter 1, Goldstein 2009b). While this example concerns the ability to distinguish weight, the same law holds for the ability to distinguish sounds of different intensities as well as intensity of colors. The tendency of the brain to perceive stimuli in a logarithmic fashion is true across many perceptual domains. In fact, when kindergarden children are asked to place numbers 1-10 along a number line, they place 3 in about the middle, just as one would expect from a logarithmic perspective. This ability disappears with mathematical education, but persists in those who are not given a formal education in mathematics, indicating that our brains are naturally wired to perceive numbers logarithmically (Varshney and Sun, 2013). Some sensory domains utilize logarithmic scales to measure stimuli; scales such as sound intensity, earthquake intensity, and frequency along the electromatic spectrum are logarithmic. Information theory suggests that logarithmic

scaling provides optimal compression of information to minimize relative errors in perception while accounting for limits in our neural bandwidth. Sun et al. (2012); Varshney and Sun (2013) showed that a bayesian model for perception would result in a model that mimics the logarithmic relationship in Weber's Law. This suggests that the logarithmic nature of human perception is a result of an heuristic that increases processing power by reducing the neural bandwidth necessary to process information through quantization of continuous information and compression of discrete information. From a statistical graphics point of view, then, log-transformed scales should be used instead of linear scales for continuous color scales and for contour plots with a fixed number of contours, as this provides more information discrimination ability and mimics natural human perceptual tendencies. The reasons for this guideline are discussed more in section 1.1.4.

1.1.3.2 Colorblindness and color perception

Another common "bug" in the visual system are mutations that change (or remove entirely) the cones in the retina. Such mutations are commonly termed "colorblindness" and encompass many different types of mutations, shifts and deletions that affect color perception in the visual system. These mutations affect up to 5% of the population, and are generally more common in males than they are in females, as two of the three genes producing cones are found on the X chromosome. Evolutionarily, these mutations are maladaptive for gathering plants (as finding red berries within green leaves is more difficult with the most common types of colorblindness), but may be adaptive for seeing camoflauged objects (Morgan et al., 1992). In statistical graphics, however, these mutations often disrupt perception of standard color schemes used in maps, heatmaps, and divergent color scalings.

In the natural world, many strategies can be used to compensate for colorblindness; the most common of these strategies is to look for textural variation instead of color variation (which may be why camoflauged objects are easier to see), but these strategies fail when viewing abstract, constructed visual stimuli, such as graphics, which may not have textural variation which corresponds to color variation. Compounding this problem, the rainbow color schemes that are commonly used are particularly vulnerable to misinterpretation by colorblind viewers.



(a) Protanopia (b) Deuteranopia (c) Tritanopia

(d) Protanomaly (e) Deuteranomaly (f) Tritanomaly

Figure 1.10: Rainbow color scheme with simulations of each of the six common types of color deficiency. The original image, on the left, is from Light and Bartlein (2004). The top row of pictures on the right show simulations of the map when cones are entirely missing. The bottom row of pictures show simulations of the map when each cone is altered due to genetic mutation.

Figure 1.10 shows a map using a rainbow color scheme (first shown in Light and Bartlein (2004)) and simulated images showing what that map would look like to those with missing cones (-anopia) and cones with altered wavelengths (-anomaly). Simulations[3] show that rainbow color schemes are incredibly difficult for color-deficient individuals to read, because the opposite ends of the color spectrum appear extremely similar with mutations to the first or second cones. In addition, categorical rainbow color schemes require longer fixation times to identify regions of importance (Lewandowsky et al., 1993) and result in decreased recall accuracy compared to monochrome categorical color schemes. Light and Bartlein (2004) provides color schemes that are more appropriate for those with color deficiencies, but not all of these schemes are appropriate for all types of color blindness. Silva et al. (2011) suggest many tools to recommend appropriate color schemes for colorblind users as well as tools to preview graphics as they might

[3]provided by http://www.color-blindness.com/coblis-color-blindness-simulator/

look to color-deficient or colorblind users.

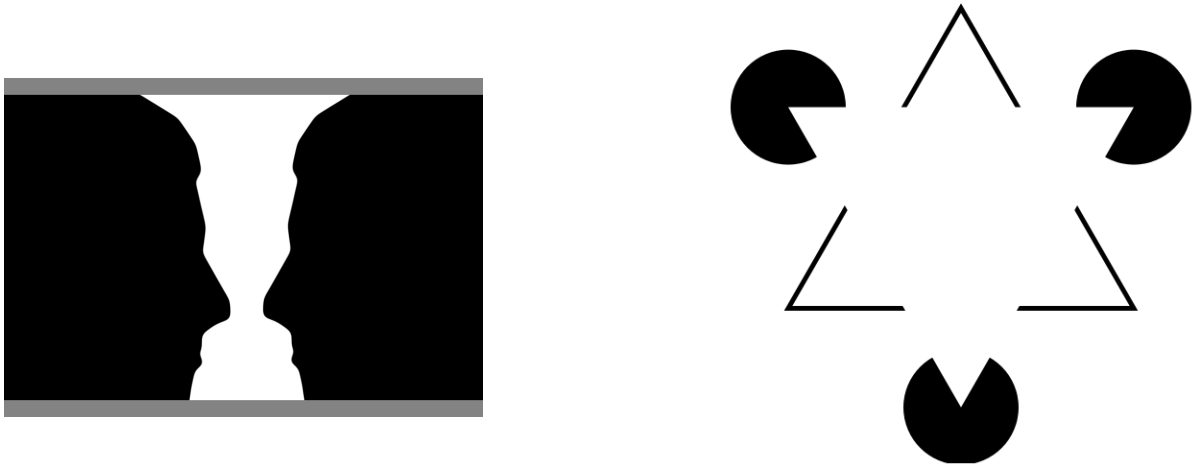
Appropriately colored maps and graphs are not only useful for those who have impaired color vision, they can also be much easier to read for those with normal vision. Treinish and Rogowitz (2009) suggest that color schemes which utilize the range of human color vision appropriately produce more aesthetically pleasing graphs and more accurately convey data in a form appropriate for the human perceptual system. Even some dual-color schemes may be problematic, as some evidence (Lewandowsky et al., 1993) suggests that these schemes are not infrequently inverted when encoded in memory, and are suboptimal for those with limited color perception as well as when printed in monochrome. Psychologically, color schemes which utilize both hue and intensity (for instance, transitioning from blue to red through white) require binding of two features, increasing encoding latency and the possibility of recall errors. Where possible, schemes utilizing only one feature (generally intensity, to preserve accessibility) are preferrable.

1.1.3.3 Optical Illusions

The "software" programs presented in section 1.1.2 are generally efficient at completing everyday tasks: navigating the environment, avoiding predators (lions or cars, as the case may be), and identifying situations and objects relevant to the task at hand. These heuristic-style approaches produce suboptimal results when applied to more artificial tasks, such as reading statistical graphics. As a result, it is important to understand where conflicts between sensation and perception may occur, so that these conflicts can be dealt with or avoided entirely. In this section, we will discuss several optical illusions and explanations for their occurence based on the visual system.

Physiological Illusions The illusions shown in figure 1.4 are illusions which occur due to the wiring of the brain. These illusions can generally be avoided in statistical graphics, but are difficult to counteract once they occur, as the illusion is literally hard-wired into the brain.

Gestalt Illusions Some illusions occur due to a conflict of gestalt principles. Two of these illusions are shown in figure 1.11: the figure/ground illusion, and the illusory contour illusion.



(a) Figure-Ground Illusion

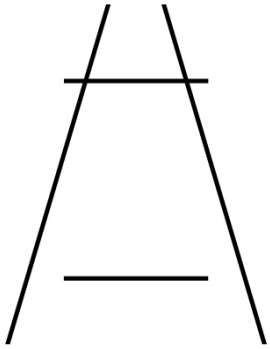
(b) Kanizsa Triangle Illusion

Figure 1.11: Illusions due to misapplied or ambiguous Gestalt rules.

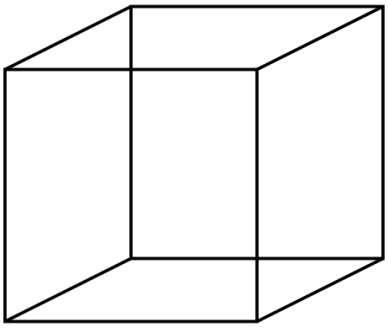
The figure/ground illusion depends on the color of the top and bottom edges of the picture; if the edges are black, the vase appears to be the central part of the image; if the edges are white, the faces appear to be the central part of the image. When the edges are omitted, the image seems to oscillate between a vase and the profile view of two heads. This is a result of ambiguity in identifying which part of the image is the background; when the top and bottom edges are present, that cue is sufficient to resolve the illusion. The Kanizsa triangle demonstrates the Gestalt principles of good form and continuity: We perceive objects that are partially obscured by a floating white triangle, even though no such triangle actually exists. The illusory triangle produces an image that is much simpler (3 circles, a black triangle outline, and a white triangle) than the objects that are actually displayed (3 partial circles and three V shapes arranged pointing in toward a central point). In addition to the existence of the illusory contour, we also perceive some depth to the image; that is, the white triangle is perceived as being above the other components of the image (Coren and Porac, 1983), as this is the only way to make sense of the set of stimuli in a simple fashion. In general, these gestalt principles make

sense of the natural world, but when applied to artificial contexts, they occasionally produce unexpected results.

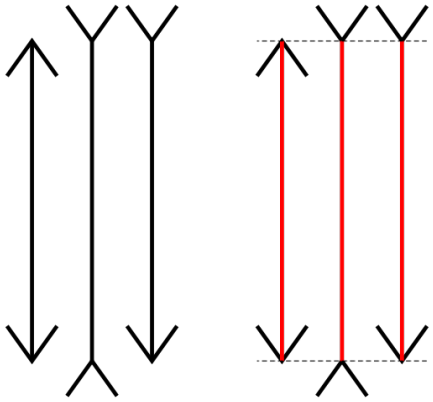
Depth illusions Other optical illusions occur due to the optimization of the visual system for three-dimensional perception. These three-dimensional heuristics can produce unexpected or misleading results when applied to two dimensional objects.



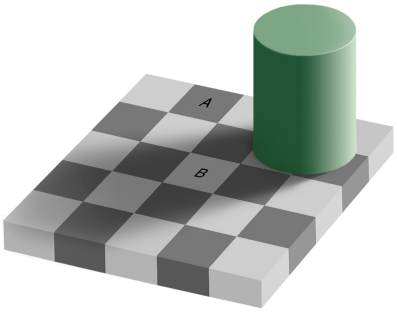
(a) Ponzo Illusion



(b) Necker Cube



(c) Muller-Lyer Illusion



(d) Color Constancy Illusion

Figure 1.12: Illusions due to misapplied depth perception.

Figure 1.12 contains four of the more interesting optical illusions that result from ambiguous figures that trigger depth cues. The Ponzo illusion (figure 1.12a) suggests that the top line is longer than the bottom line, because of the implied convergence of the two vertical lines (to understand the natural scene that produces this illusion, consider railroad tracks converging at the horizon). The Necker Cube, shown in figure 1.12b, can be seen such that the top-right face

is closest to the viewer or alternately such that the bottom-left face is closest to the viewer. Due to the ambiguity in the image, it will often seem to "flip" when the viewer temporarily loses focus on the image (Gregory, 1997). The Muller-Lyer illusion (figure 1.12c) is generally believed to result from misapplied depth cues as well - the left-most image would occur in nature as the exterior corner of a building, the middle image would occur when viewing an interior corner of the same building, further away from the viewer (Ward et al., 1977; Gregory, 1968; Fisher, 1970). As a result of the illusion, the middle line appears to be longer than the first or third lines. Figure 1.13 shows the first two parts of the illusion in a context which removes the ambiguity through additional depth cues. The additional cues result in the resolution of the illusion. Finally, the color constancy illusion shown in figure 1.12d suggests that the square marked A is much darker than the square marked B, even though the two squares are the same color. This illusion results from our experiences with depth and shadows: square B is perceived to be the same color as the lighter-colored squares outside the shadow, while square A is perceived to be the same color as the other dark squares in the tile pattern, regardless of the actual color due to the shadow.

Figure 1.13: The Muller-Lyer illusion in a non-ambiguous three-dimensional context.

Depth illusions in particular result from a conflict between our experience with the three-dimensional world and the appearance of two-dimensional ambiguous stimuli. The Necker cube "flips" because there are two physical objects that could produce the same retinal image, the Muller-Lyer illusion exists because our experience with the three-dimensional world is harnessed

inappropriately for a two-dimensional figure, and the color-constancy illusion exists because our brains automatically correct a pseudo three-dimensional image to represent the reality of that image in the real world. These conflicts occur in statistical graphics as well; Chapter 2 explores statistical graphics that trigger a three-dimensional heuristic in the brain and lead to misleading conclusions.

There are other optical illusions that have the potential to appear in statistical graphics but are not easily classified (or necessarily easily explained). Two of these illusions are considered below.

Other Important Optical Illusions Certain illusions do not lend themselves to simple classification. While many illusions are the result of multiple concurrent processes in the brain, these illusions may not even be fully understood. The Poggendorff illusion, shown in figure 1.14, is one such illusion.

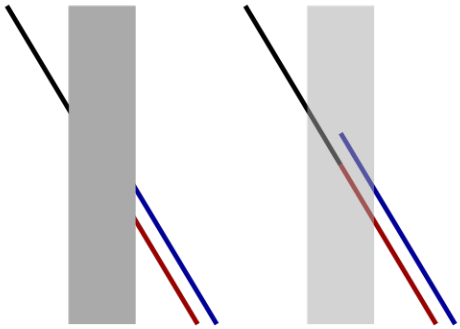
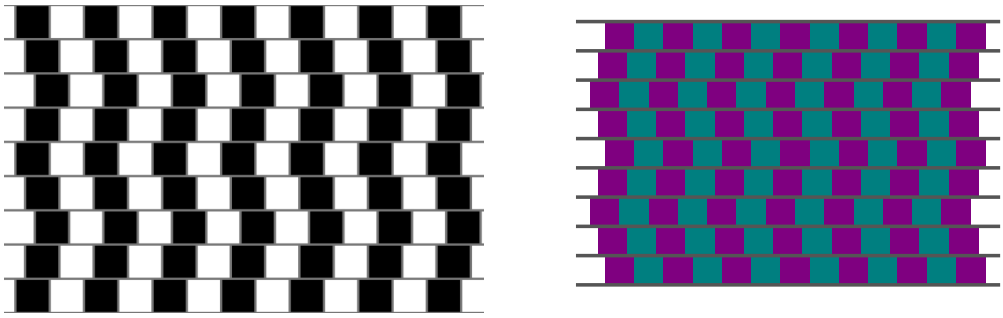


Figure 1.14: The Poggendorff Illusion. The left figure shows a black line which is obscured by a grey rectangle; it appears that the blue line would intersect the black line if the rectangle were not present. In fact, the black line and the red line are co-linear, and the blue line is parallel to both other lines.

Gregory (1963, 1997) suggests the Poggendorff illusion is similar to the Muller-Lyer illusion in that it results from misapplied depth cues, but Green and Hoyle (1963); Ward et al. (1977) found no evidence that participants viewed this illusion in any three-dimensional context. Instead, Green and Hoyle (1963) suggests that the illusion results from a tendency to perceive acute angles as less acute and obtuse angles as less obtuse than the image suggests. As the illusion disappears as the angle of the line segment approaches horizontal, this seems to be a

reasonable explanation, but it is almost certainly not complete (Morgan, 1999), as the illusion survives in forms which do not preserve the acute angle intersections. Regardless, this illusion can make it difficult to read line graphs (Amer, 2005; Poulton, 1985) if proper precautions (use of reference lines and grid lines) are not taken.

The second of these illusions is the cafe wall illusion, shown in figure 1.15, named because this tile pattern is apparently common in cafes.



(a) Original Cafe Wall Illusion (b) Isoluminant Cafe Wall Illusion

Figure 1.15: The Cafe Wall illusion. The lines between rows of black and white tiles are parallel but appear to be tilted. The second image shows the isoluminant version, which mitigates some of the illusion but does not entirely eliminate the effect.

The cafe wall illusion is in part due to the contrast between light and dark zones (as in the mach bands and hermann grid illusion), and much of the illusion is resolved when the black and white tiles are replaced with isoluminant colors, but some of the illusion still remains. Westheimer (2007) suggests that this portion of the illusion is because the position of a black-white border will be biased to appear closer to the black side of its physical location, an effect which is compounded in the cafe wall illusion to produce the appearance of tilted lines. This illusion, while not explicitly found in most statistical graphics, shows that even simple (and pleasant) configurations of geometric objects can wreak havoc in the brain under the right circumstances. In fact, the illusion is so simple that it is also known as the "Kindergarten illusion", but its cause is sufficiently complex that it has not been fully explained by psychologists or neuroscientists.

1.1.4 Cognitive Load

It is fairly well established that statistical graphics are useful in part because they can replace long tables of data, summarizing information in a form that is much easier to understand

and mentally maniuplate; as such, it would not be out of line to suggest that graphics require less cognitive load than tables of the same data. When we discuss cognitive load, we typically mean the limits in short-term memory, cognitive manipulation, attention, and mental bandwidth that bound our ability to take in new information and draw conclusions from that information. In this section, we will briefly discuss some of these considerations.

Short Term Memory A famous paper in memory and cognition (Miller, 1956) suggests that active memory can contain only 7 (plus or minus two) chunks of information. A chunk of information could be a single letter or number, a meaningful collection of several letters or numbers (e.g. a word or an area code), or an association. This limitation is important in designing information for graphical consumption. For instance, the number of categories in legends should be limited to 7, to allow a viewer to store the associations within the legend and then use that information to understand the graph. Abuse of this limitation is referred to as a "color mapping attack" in Conti et al. (2005), a paper detailing the various ways to "attack" a human visualization system. Similarly, viewers should not be expected to remember more than 7 "chunks" of information from a single graph. Due to these limitations in memory, when a single color scale is used to represent more than one order of magnitude of variation, using a logarithmic scale provides more optimal information scaling than using a linear color scale (Sun et al., 2012; Varshney and Sun, 2013).

Information Integration Integrating multiple dimensions of information (or mentally combining multiple graphics) is another area which can strain the ability of the brain to utilize information effectively. Well-constructed graphs can help the brain to integrate information by connecting points across dimensions (through the use of regression lines, clustering, etc.), which creates "chunks" of information that can then be stored in memory in a more compressed format. These chunks are useful because they allow people to draw conclusions from multiple sets of data across multiple dimensions (Gattis and Holyoak, 1996). Poorly created graphics may make this task harder or even promote the encoding of misleading chunks; for instance, data that is overplotted may obscure the important trend and may also produce chunks which

lead to the wrong associations being stored in memory. This integration limitation is very much related to short-term memory, but is also constrained by mental effort limitations and processing capacity. As a result, it is important to reduce the effort required to integrate multiple graphics.

Attention Human attention is limited; thus visualizations which do not focus attention on important aspects of the data are likely to confuse the reader. Tukey (1977) said "The greatest value of a picture is when it forces us to notice what we never expected to see". When there are too many salient features to notice anything in particular, attention is split too many ways to gain useful information from the picture. Graphics should present data in a controlled fashion, so that focused attention is rewarded with useful information taken from the graph. Conti et al. (2005) describes graphs that do not follow this principle as "processing attacks", in that the overload the "CPU" with needless calculations and mental manipulations that are ultimately futile to understanding the data.

The consequence of the limits of human perception and processing capacity is that there is a limited amount of information one can expect to portray graphically; thus graphics should be designed to most efficiently communicate information so that this cognitive overload does not occur. The next section presents studies which examine the perception of graphs and charts directly across a wide range of perceptual levels and experimental conditions.

1.2 Statistical Graphics

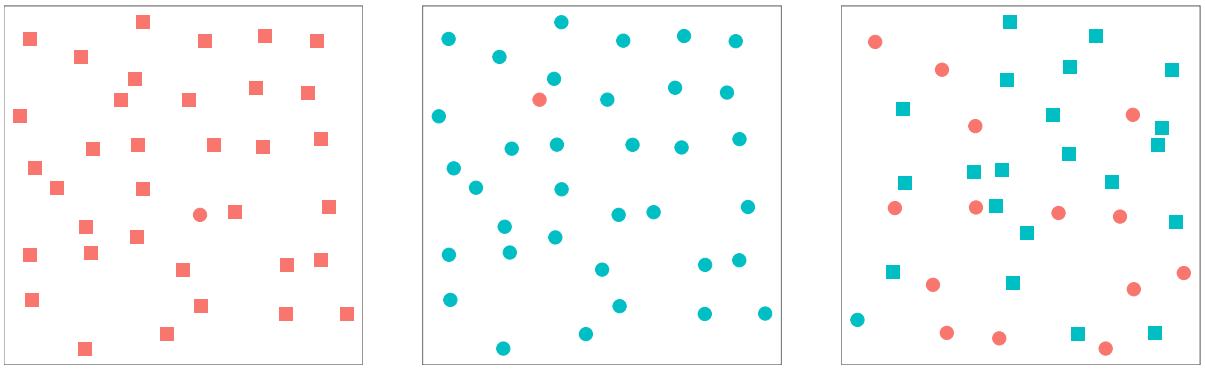
Psychologists who study graphical perception are generally concerned with the underlying mechanisms of effects within the brain, and thus study very simple graphics and lower-level perceptual effects. In statistics, the literature is somewhat more variable; Cleveland and McGill (1985) produced the seminal paper on the subject, but outside of that work, there are relatively few papers that examine the accuracy of judgments made from graphs through user studies that mimic the way graphs are used in practice. There have been a few papers in other disciplines; business and communications researchers occasionally study graphs and charts as well. As a result, the literature in this area is scattered across many disciplines. In order to organize

this section effectively, we will begin with the lower-level graphical perception literature and conclude with studies that have more external validity and are applicable to statistical practice. For the purposes of the following sections, lower-level perception research is research which involves either extremely simple graphical elements (or non-graphical research which applies directly to graph perception processes), or which does not require attention (pre-attentive features of graphics). These experiments provide some information, but are less informative than experiments which utilize more complex graphics in realistic scenarios.

1.2.1 Preattentive Perception of Statistical Graphics

Much of the lower-level research within the statistical graphics literature has been performed by those within the psychological community that study human information processing. In particular, Healey and Enns (1999, 2012) have produced several papers studying the accuracy of conclusions viewers can make after less than 1/2 of a second of viewing an image. As discussed in section 1.1.2.1 and demonstrated in figure 1.16, certain features do not require individual focus to process; these features are called preattentive and can be detected on the first glance (typically within 250ms). Healey's work focuses on determining which features can be detected in a pre-attentive fashion, and whether a hierarchy of features exists when these features are combined. Healey suggests that for three-dimensional displays, the 3d layout is determined first, surface structure and volume are determined next, followed by object movement (if present), luminance gradients, and color; he suggests that if there are conflicts between these 5 levels, priority is given to an earlier process (Healey and Enns, 2012). Healey and Enns (1999) showed that if visualizations are carefully constructed to conform to the architecture of the human visual system (isoluminant colors, removing certain background patterns from textural arrays), visual estimation tasks can be performed preattentively. The experiment also revealed an interference effect between texture and color that corresponds to previously documented interference between preattentive features (Treisman, 1985). Figure 1.16c demonstrates the interference effect; it is much easier to locate the target point in figure 1.16a or figure 1.16b than it is to locate the target in figure 1.16c.

Healey's work on preattentive perception is interesting, and provides a reasonable approach



(a) Shape (b) Color (c) Interference

Figure 1.16: Shape and color are detected preattentively in figures (a) and (b), but interfere in figure (c) so that location of the target in (c) is no longer preattentive.

to creating graphics compatible with the human visual system, but his work is largely focused on multidimensional displays and his focus on preattentive processes limits the applicability of his work to statistical graphics. In particular, most graphics are created with the idea that a viewer will spend more than one second looking at the graph, so not all features need to be preattentive to be useful. In addition, many of the multidimensional displays he designs are very load-intensive to understand; with so many additional dimensions, the viewer must spend considerable time understanding the scales and legends which correspond to each variable (feature integration over multiple dimensions is time intensive and generally not preattentive). In the next section, we will examine the literature concerning higher-level graphical perception, including perception at the attentive level and which types of graphs are more accurately perceived by viewers.

1.2.2 Conscious Perception of Statistical Graphics

Graph perception from a statistician's point of view is more focused on the attentive stage of perception: When asked to answer a question using a graph, what parts of the graph are useful, and how is the information transferred from the image to working memory in the brain? Several models have been proposed to describe this process; of these, the set of "task models" and "integration models" seem to be most consistent with empirical evidence.

1.2.2.1 Models of Graph Perception

These task-based models suggest that task-based graphical perception, e.g. using a graph to answer a specific question involves several stages of information processing (Ratwani et al., 2008).

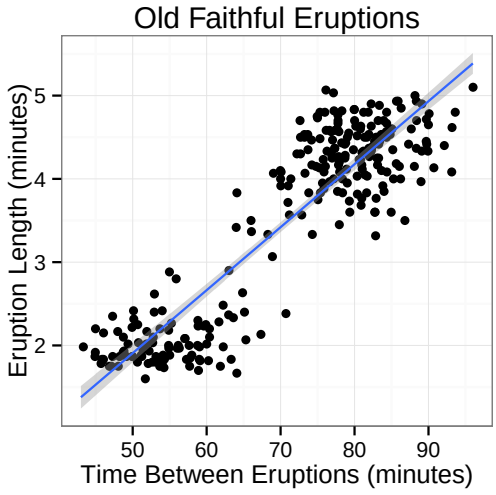
1. Parts of the question are read several times
2. The graph is searched for relevant information, with focus shifting from the graph axes to the main part of the graph and back again (pattern recognition)
3. Once information is found, the focus shifts between the important part of the graph and the legend several times in order to keep the relevant information in working memory (conceptual relations produce quantitative meaning from visual features)
4. The question is answered and the participant moves on to another task (the question is related to the encoded quantitative features)

These steps are illustrated in figure 1.17.

Working within this task-oriented framework, researchers have explored the "search" portion of the task-based model, the information integration portion of the model, and the types of graphs which facilitate both the "search" and "integration" portions of the task. Integration models modify the above sequence to allow for more complex graphical relationships to be assimilated, such that a viewer cycles between stages (2) and (3) several times in order to encode different portions of the graph. The time required for each of these steps may also change in accordance with the reader's familiarity with the task and graphic style; those who are more familiar with similar graphics may be able to encode information faster and in larger chunks and thus answer the question more quickly (Carpenter and Shah, 1998).

Analyzing graphs using task-based models emphasizes the importance of spatial relationships between graphical elements. The gestalt laws of proximity and similarity dictate that items which are close together or physically similar (the same shape or color) are perceived as a group; this spatial perception creates "chunks" of the graph which may be encoded as single objects and thus reduce the mental bandwidth necessary to process the image. Figure

Question: What is the relationship between the length of the eruption and the time between eruptions for Old Faithful?

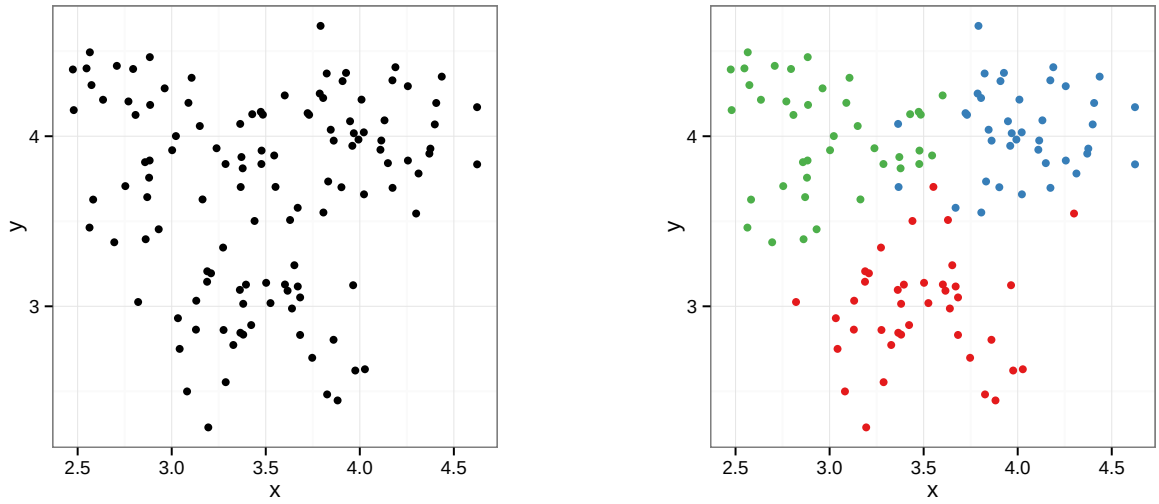


Sample mental steps:

1. Understand the question, identify "length of eruption" and "time between eruptions" as things to search for in the graph.
2. Look for "length of eruption" on the axes, determine that the y coordinate contains that information. Look for "time between eruptions" on the axes and determine that the x coordinate contains that information. Verify that these quantities are indeed what is being sought by re-reading the question.
3. Establish that as the time between eruptions increases, the length of the eruption increases. Note that there seems to be a bimodal distribution of points.
4. Answer the question: As the time between eruptions increases, the length of the eruption seems to increase.

Figure 1.17: Task analysis of a simple graph-based task. The graph shows the length of the eruption of Old Faithful as a function of the waiting time between eruptions, with a corresponding sample "mental dialog" of the perceptual tasks involved in answering the question in response to the graph, according to Shah and Miyake (2005).

1.18 shows the advantage of "chunks" in graphs, as the second graph shown is much easier to describe and understand than the first graph, even without the contextual meanings of the variables. Of graphics that present information of similar complexity, graphics that require less effort to understand and search for relevant information are preferrable (Cleveland and McGill, 1985). More complex models of the graphical perception process suggest that data are integrated on a visual level and then further integrated on a cognitive level, to form successive clusters of information. Once these clusters are formed, further information can be integrated by comparing and contrasting different clusters to understand the higher-level meaning in the graph (Ratwani et al., 2008). Graph types which cater to this hierarchical clustering mechanism may be more easily understood by viewers than graphs that do not provide information in a manner easily assimilated by the human brain. Based on this information, facets of graphs may be particularly useful for mapping multidimensional data to provide "chunks" of information in a relevant manner that can be integrated into the viewer's working conceptual understanding



(a) Proximity-based grouping

(b) Proximity and Similarity-based grouping

Figure 1.18: The utility of chunking in graph perception. Graph (a) could potentially be described as three distinct clusters of points rather than 120 individual points, but it is much easier to draw conclusions from graph (b), which has colored points that clearly show the grouping structure in the data. The second figure would more probably be encoded and described as three groups of 40 points, which serves as a form of mental data compression.

of the dataset. Additionally, color schemes and appropriate labeling of graph features which reduce the amount of work necessary to integrate numerical information from a legend into the visual representation of the graph facilitate graphical inference (Carpenter and Shah, 1998).

From a statistical perspective, much of the literature involved in understanding graph perception from a task-analysis point of view focuses on simple graphics, such as side-by-side bar graphs or line graphs, and straightforward tasks of reading information from the graph accurately, rather than examining model assumptions or making inference beyond the data. The psychologocial mechanisms involved in processing simple graphics are perceptual, and typically require direct comparisons rather than mental manipulation in order to satisfy the tasks posed by the researchers (Trickett and Trafton, 2006). More complicated graphics and more sophisticated tasks may require comparison of two or more distinct graphs and may utilize working memory and spatial reasoning; these situations are not as well studied (Shah and Miyake, 2005). We begin first with the simple tasks of graph comprehension, and will summarize work with more complicated graphics at the end of this section.

1.2.2.2 Perception of Simple Graphs

A series of experiments by Cleveland and McGill (1984, 1985) studied basic perceptual tasks in graphical perception to produce a relative ordering of graphical elements by the accuracy of participant conclusions. This ranking is shown in Table 1.1. Other researchers (Kosslyn, 1994) have collapsed this ranking into position/length/angle and area/volume, as the difference in accuracy between categories 1, 2, and 3 is small compared to categories 4 and 5.

Rank	Task
1	Position (common scale)
2	Position (non-aligned scale)
3	Length, Direction, Angle, Slope
4	Area
5	Volume, Density, Curvature
6	Shading, Color Saturation, Color Hue

Table 1.1: Cleveland and McGill (1984, 1985) ordering of graphical tasks by accuracy (adapted from both papers and Shah and Miyake 2005). Higher-ranking tasks are easier for viewers than low-ranking tasks and should be preferred in graphical design.

The particular task required of participants in experiments also has an effect; Simkin and Hastie (1987) found that readers were more accurate in determining position when presented with a bar graph, but when readers were presented with a pie chart, they were more accurate at determining proportional judgments (using angles). This finding contradicts Cleveland and McGill (1984) to some degree and suggests that the experimental design and specific task are important in evaluating these sorts of user studies; the contradictory results also suggest that graph type is an important influence in determining what information viewers encode from the graph. This conflict also illustrates that the user's attention and past experience influence the judgments they make from a given graph: when participants were asked to provide a summary of the graphic, their answers depended on the type of display: bar charts elicited a comparison judgment, pie charts elicited proportional judgments (Simkin and Hastie, 1987). Similarly, when presented with a line graph, viewers are more likely to summarize the graph in terms of the slope of the trend line (even when the x-axis is discrete); when presented with a bar graph, viewers summarize the information using discrete comparisons (Carswell and Wickens, 1987; Shah and Miyake, 2005). The task and the graph format interact to influence viewer

perceptions, thus, when creating graphics, statisticians should match appropriate graphical formats to meaningful conclusions about the data.

The task requirements are mediated by the limits of human processing ability. Chernoff faces, once proposed for visualizing multidimensional data, are difficult to read because viewers are unable to store the legend and the image in working memory; comparisons must be made serially and with conscious attention (Shah and Miyake, 2005), and data features may not map well to relatable facial features (Lewandowsky and Spence, 1989b). Similarly, while color does not generally correspond to precise quantitative information, certain color schemes utilizing hue, saturation, and brightness together can provide an implicit numerical ordering that does not place exceptional demands on working memory (Shah and Miyake, 2005). Color schemes which correspond to everyday situations (e.g. using a blue to red scale for low to high temperatures) may also reduce the demand on working memory (though such a scale may be problematic due to principles of color perception). While specific numerical judgments would still require selective attention, the "gist" of a graph using such schemes may be understood fairly quickly.

It is important to consider working memory when constructing graphical scales (particularly when utilizing a discrete scale for categorical data), but it is also important to consider feature selection and discriminability as well. Color is generally believed to be preferrable for representing strata on a scatterplot (Cleveland and McGill, 1984), but Lewandowsky and Spence (1989a) found that if color is not available or appropriate, shapes, intensity, or discriminable letters may be utilized without a significant decrease in accuracy. Discriminable letters are those which do not share physical features such as closure and symmetry, such as the letters H, Q, and X; confusable letters, such as H, E, and F, are associated with significantly less accurate perception. Demiralp et al. (2014) synthesized experimental evaluations of stimuli to create "perceptual kernels" describing the perceived distance between values; multidimensional scaling of the resulting distance matrix produces four distinct groups of shapes which share features (triangles with various degrees of rotation, squares and diamonds, and non-convex shapes such as x, +, and *). This separation suggests that feature integration underlies many of the processing speed effects found in studies examining discrete palettes and scales: palettes which are composed of confusable shapes, letters, or colors will require more processing time

(and decrease accuracy) compared with discriminable palettes.

Other graph features can also influence viewer inferences: multiple studies suggest that our mental schematic for a graph is most consistent with a 45° trend line (Cleveland et al., 1988; Tversky and Schiano, 1989). "Banking to 45°" is a commonly-cited recommendation for optimal graphics (it is also quite old, according to Wickham (2013)), and does have some limited utility in reducing the strength of the line-width illusion (a more thorough discussion of this heuristic is provided in Chapter 2). Axis scale transformations can make it easier for viewers to spot outliers of data conforming to skewed distributions (though this does require some domain-specific knowledge of statistical distributions), and appropriately labeled graphs can reduce the working memory requirements by reducing the number of "back-and-forth" comparisons required to pass information into working memory (Shah and Miyake, 2005).

While graph perception is commonly limited by working memory considerations, there is some evidence that we perceive and process graphical information differently than numerical algorithms: Bobko and Karren (1979) found that participants underestimated correlation coefficients when estimating correlation strength, particularly under unequal variance; their estimates were in fact more closely aligned with r^2. In addition, visual estimation tends to discount the effect of perceived outliers, producing a more robust estimator than numeric estimators designed for that purpose (Lewandowsky and Spence, 1989b). Finally, some evidence suggests that when visually estimating lines of best fit, we fit the slope of the first principal component rather than the least squares regression line; that is, we consider variability in x and y rather than only considering variability in y(Mosteller et al., 1981). While these studies do not offer theoretical explanations or attempt to provide causal explanations, human perception of data displays does appear to differ from computational exploration in meaningful ways.

These studies indicate that it is important to consider the cognitive processing of statistical graphics as well as the data used to generate these graphics: the type of graph, color scheme, annotations, aspect ratio, legends, and axis transformations can all influence the amount of mental processing required to draw conclusions from a graph, as well as the types of conclusions that graph viewers are likely to draw. Many of these features were studied in relative isolation, using simple graphs that may lack real-world context. More complex, domain specific graphs

may create higher cognitive load and recruit previously acquired knowledge; experiments using simple, bland graphics may not be applicable to more complex graphics meant for experts. What follows is a summary of the relatively sparse literature on these sorts of real-world graphics.

1.2.2.3 Perception of Complex, Domain-Specific Graphs

Carpenter and Shah (1998) showed that graph comprehension time increased when the number of distinct x-y functions (i.e. nonparallel sloped lines) increased, even if the same data was represented. The density of these functions also had an impact: dense graphs with multiple intersecting trend lines took more time to interpret than dense graphs with parallel trends or sparse graphs with intersecting trend lines. This supports the idea that the information conveyed in the graph must be read into working memory before the graph can be described or used for inference; more complex graphs would take more time to understand and internalize. Additional factors can also influence the ease with which graphs are perceived and understood in real-world scenarios. Gattis and Holyoak (1996) found that graphs were more accurately perceived when the dependent variable was on the y axis and the independent variable was on the x axis, even when the perceived IV and DV were manipulated using a cover story. In even more complex visualizations, Trickett and Trafton (2006) found that meterologists and other domain experts would mentally superimpose graphs from memory on visible graphs, utilizing spatial processing rather than manipulating a physical interface. These interactions demonstrated complex spatial manipulation to assimilate information from multiple graphs, particularly when the information provided in the graphs conflicted with prior information, either from the meterologist's own domain knowledge or verbal information provided during the course of the study. While the procedures used in this study rely on verbal descriptions of mental processes (i.e. the meterologist speaking aloud as they process each graph and map to assimilate information), the evidence is sufficient to suggest that in addition to working memory and the visual processing performed by the brain, some complex graphs also utilize spatial processes (and the corresponding brain regions) to perform complicated overlays and mental transformations. By designing such complicated graphs to more easily facilitate such

mental operations, it is possible that more effective spatial visualizations could make these graphs more accessible.

Complicating the research into more complex graphs, there are many different types of complexity that can affect graphics. There may be differences in how processing occurs for large amounts of data, but it could also be that more complex x-y relationships could also require more mental effort. In addition, multiple relationships can be depicted simultaneously, either because of underlying groups in the data or because multiple related trends are depicted on the same graph (though this is widely acknowledged as bad practice in statistical graphics). Finally, the mental complexity of the task required of the graph viewer can also factor into the amount of time and effort required to complete a task using a graph. These different types of complexity interact with the graph format; for instance, line graphs are less effected by increasing complexity than bar graphs (Tan, 1994), and bar graphs are more affected by increasing complexity than pie charts for ratio judgments (Hollands and Spence, 1998).

Finally, complex graphs often facilitate different types of participant tasks; rather than simple numerical judgments or information lookup, complex graphs may encourage (or require) viewers to use prior knowledge and interpretation skills. These additional complications make experimental study of complex or domain-specific graphs more difficult. Many of these problems (types of complexity, expanded tasks, prior knowledge) make further work in this area somewhat difficult. One concept that facilitates studies examining the relationship between complex data and graph format in statistical graphics is the grammar of graphics, which we discuss in the next section, along with other experimental methodology useful for understanding how people perceive statistical graphics.

1.3 Testing Statistical Graphics

1.3.1 Basic Psychophysics Methodology

Psychophysics studies are generally concerned with the ability to detect a stimulus (or a difference between two stimuli). Many classic psychophysical methods are still used in studies today (for an example, see Chapter 2). Several of these methods are mentioned here; for a

more thorough review, see Goldstein (2009b).

Method of Limits The method of limits seeks to determine the level of intensity at which a stimulus is just barely detectable. A series of trials is used, with each trial starting at either the lower or upper range of intensity and incrementally moving towards the opposite end of the range; for each point, the observer indicates whether they can detect the stimulus. At the end of several trials, the detection limits are averaged to produce a measured absolute threshold. Figure 1.19 demonstrates this process.

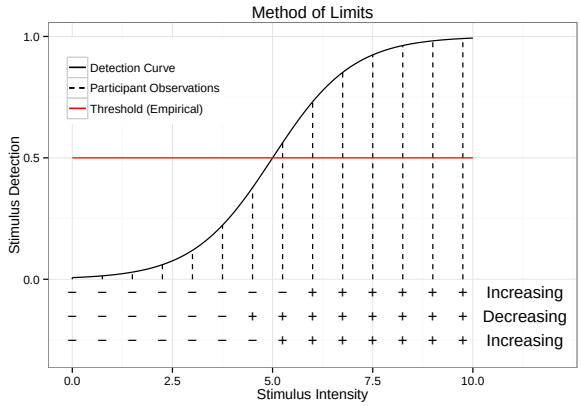


Figure 1.19: Demonstration of the method of limits. In this experiment, three trials were performed, two trials starting from 0 and increasing, and one trial starting at 9.5 and decreasing. The empirical detection threshold is the threshold at which detection occurs 50% of the time, and is shown in red.

Method of Adjustment This method is similar to the method of limits, except that the stimulus intensity is adjusted by the observer (not the experimenter) in a continuous manner until the observer can just barely detect the stimulus. This procedure may be repeated several times, with trials averaged to produce a mean value for the absolute threshold. Figure 1.20 demonstrates this procedure.

Measuring the Difference Threshold The difference threshold, discussed in section 1.1.3.1, is the smallest detectable difference between two stimuli. This threshold can be measured using either the method of limits or the method of adjustment. Rather than increasing the absolute intensity of the stimulus as discussed above, two stimuli are given: one with constant

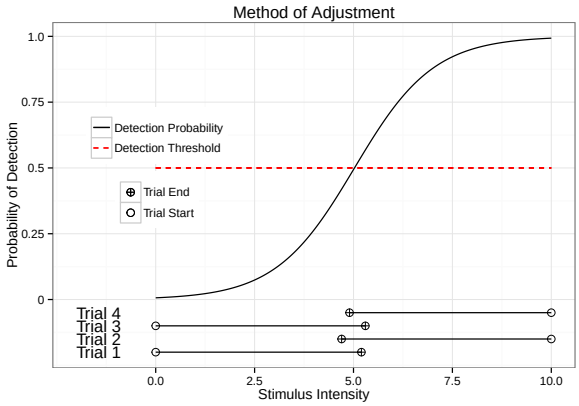


Figure 1.20: Demonstration of the method of adjustment. In this experiment, four trials were performed, two trials starting from 0 and increasing, and two trials starting at 10 and decreasing. The empirical detection threshold is the threshold at which detection occurs 50% of the time, and is shown in red.

intensity and one whose intensity may vary either continuously or incrementally (depending on the method utilized). The participant is instructed to identify the point at which the two stimuli are indistinguishable (if the varied stimulus is approaching the constant stimulus) or the point at which the two stimuli are distinguishable (if the varied stimulus is diverging from the constant stimulus).

Magnitude Estimation Magnitude estimation studies are used to measure the perceptual intensity of two different stimuli. For example, a participant might be shown a series of two lights, and asked to assign a number to describe how bright each light is. These numerical values would then be compared to the actual light intensity (as measured by a digital sensor or by the input voltage) to determine how perceived brightness corresponds to actual intensity. Many stimuli measured this way produce power-law functions that exhibit response compression (doubling the actual intensity corresponds to a much smaller change in perceived brightness) or response expansion (doubling the actual intensity corresponds to a change in perceived intensity that is more than double the original intensity).

1.3.2 Testing Images using Psychological Paradigms

In addition to the psychophysics methods outlined above, there are testing paradigms within psychology that are applicable to the study of statistical graphics. These include experimental methods such as visual search and eye tracking, as well as experimental control procedures that may be important in graphics studies that are similar to cognitive psychology studies. We will first discuss the experimental methods, and then briefly discuss some common control procedures that may be applicable to the study of statistical graphics.

1.3.2.1 Experimental Methods

Many psychological experiments utilize straightforward methods to address hypotheses in perception, such as asking participants to make numerical judgments based on presented stimuli. These methods are quite useful, but not particularly difficult or domain-specific. In this section, we discuss two domain-specific methods for understanding perception of visual stimuli: visual search and eye tracking.

Visual Search Simply put, visual search methods involve presenting a participant with many distractor stimuli and one or more target stimuli, and asking the participant to locate the target stimuli. Time is measured between the initial stimulus presentation and the participant's answer; participant accuracy is also considered in more complicated visual search settings. This procedure allows researchers to measure simple preattentive stimuli and can also be utilized for more complicated tasks that require attention (Anderson and Revelle, 1983). One example of a visual search task is shown in figure 1.21; a common preattentive search task is shown in figure 1.16c.

Visual search tasks can be used to measure the efficiency of a participant's visual search abilities (and focus on a task) to serve as a baseline for more complicated visual tasks. They can also be used to examine feature binding and common mistakes that may indicate relevant distractor stimuli. Even when reaction time is not directly measured, these tasks are typically given under time pressure, to establish a baseline performance that is below 100% performance on the task. This time pressure allows experimenters to avoid response compression, so that

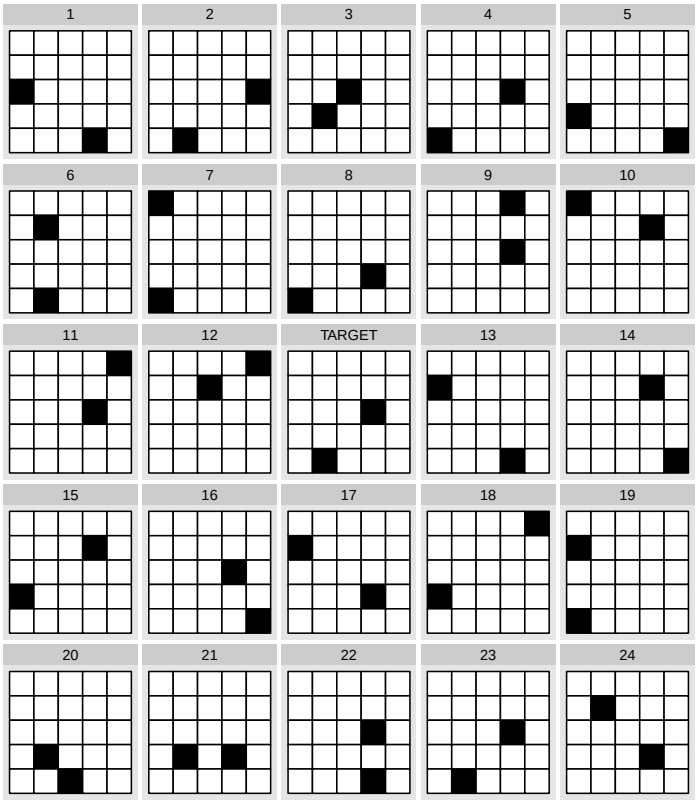


Figure 1.21: Visual Search Task. Participants were asked to locate the figure (1-24) most similar to the central "target" figure.

the number of questions completed within the time limit provides an approximate measure of response time.

Eye Tracking Eye tracking studies are often utilized in order to understand which parts or features of an image participants focus on, and in what order they examine the image components. Eye tracking studies were heavily used in order to refine the task-based models of graphical perception; they allowed researchers to understand that participants had to iterate between the question and different parts of the graph in order to assimilate all of the represented information into working memory. Figure 1.22 shows one lightweight eye-tracking assembly. The camera allows researchers to track the direction of the pupil and thus infer gaze direction.

Eye tracking studies have been performed on statistical graphics as well (Zhao et al., 2013), utilizing a visual search task and examining which graphics participants compared to determine the target plot.

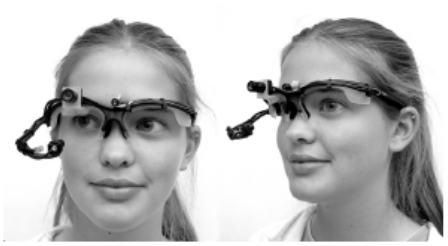


Figure 1.22: Eye tracking equipment (Babcock and Pelz, 2004). The cameras allow researchers to determine what part of a scene the wearer is viewing.

In some psychological experiments, the straightforward approach to a task can produce biased responses from participants. Human perception is highly reliant on expectations and past experience, and as a result, experimenters must take care to reduce undesired biasing effects in order to appropriately control experiments. Some of these considerations are discussed in the next section.

1.3.2.2 Experimental Control Procedures

While not all of the experimental control procedures discussed here are appropriate for every experiment, they do demonstrate the degree of control many experiments require to measure small psychological effects. The variation in the human brain and in cognitive strategies (and the sample size constraints of testing in humans) requires a large degree of experimental control in order to minimize the effects of population variance. Some of the biases of the human brain as well as strategies to address these biases are described below.

Habituation The human visual system is attracted to novelty; odd, bizarre, or new sights attract more attention than ordinary, run of the mill scenes. Habituation describes the process of becoming less interested in a stimuli; as this occurs, the mind begins to enter "auto-pilot" and attention to the task at hand becomes less focused. In infants, this habituation process is used to determine whether there is a perceived difference between two stimuli; in adults, this process is not typically as useful to the experimenter. To avoid habituation, experiments should generally consist of somewhat varied tasks in order to maintain participant attention. The psychophysics experiments described in figures 1.19 and 1.20 can be vulnerable to this

problem; to overcome habituation, trials often alternate in direction or start at random points along the intensity spectrum.

Masking Images can persist on the retina for a period after the image is no longer available; this phenomenon is called persistence of vision. In order to control the time in which the stimulus is visible, psychological experiments often will show a mask immediately after an image in order to "erase" the retina. This degree of control is often useful in experiments which focus on the preattentive stage of perception, but persistence of vision is not likely to affect experiments which take place in the attentive stage of perception (i.e. images shown for more than .5 seconds). A sample mask is shown in figure 1.23.

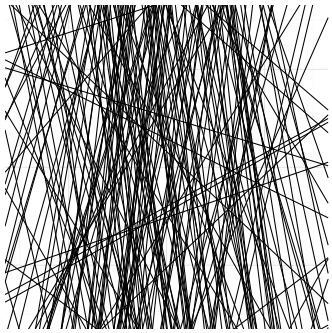


Figure 1.23: Sample mask used to "erase" the retina in some psychological experiments which test stimuli for short time periods (<1 s). The mask removes any afterimage the participant might have, ensuring that the stimuli is only available for the specified period.

Priming Broadly, priming is a technique that can be used to subconsciously bias a participant towards a certain conclusion. In cognitive psychology research, priming can be used to test word association (i.e. participants are quicker to identify an apple if they have just heard the word "fruit" than if they heard the unrelated word "pen"); in statistical graphics, priming effects are more likely to occur due to instructions or examples provided to participants at the start of a testing session. If an initial example contains notable outliers, participants are more likely to look for and recognize graphs with outliers than graphs with other notable features. Examples must be designed in such a way to avoid activating these priming affects as much as possible.

There are many other psychological mechanisms that may impact participant performance; the mechanisms presented here are simply some of the more salient considerations in experimental design for statistical graphics.

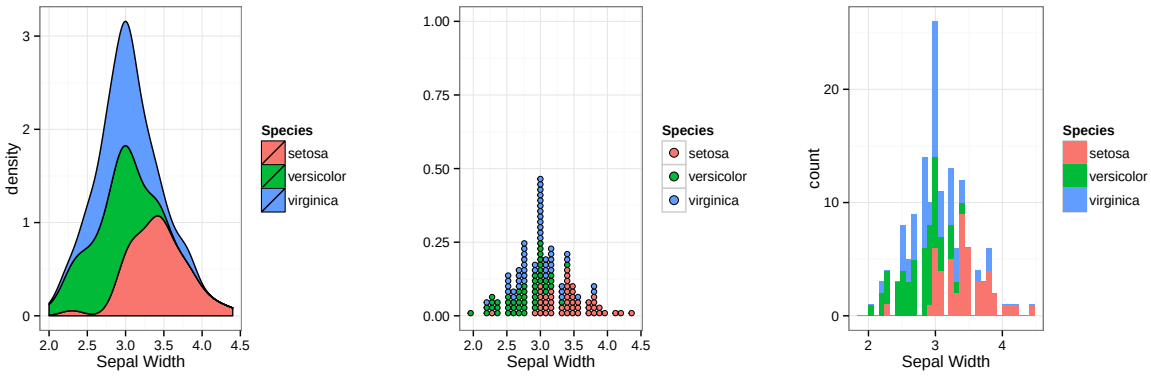
1.3.3 Testing Statistical Graphics

This section details the tools specific to the testing of graph perception within the field of statistics. Cleveland and McGill (1985) studied statistical graphics from a largely psychological perspective, but their findings have been widely utilized in the field of statistics; however, it has been 20 years since that paper, and the field has developed within statistics since that time. Two major developments, the grammar of graphics and the lineup protocol, are particularly important for future research into the perception of statistical graphics.

1.3.3.1 The Grammar of Graphics

The grammar of graphics, detailed in Wilkinson et al. (2006), is a framework for describing a graphic in terms of its basic component pieces. An implementation of the grammar of graphics for R, `ggplot2`(Wickham, 2009, 2010), provides a useful tool for manipulating graphics to test in an experimental setting. 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 1.24 shows three plots created using the same data and different geometric objects, and figure 1.25 provides the ggplot2 code to create the plots [4]. Comparing these graphics experimentally would be reasonably simple, and the grammar of graphics helps to control the extraneous variables introduced by utilizing different plot types. In addition, the grammar of graphics approach to transformations and scales allows us to easily test judgments made utilizing different axis transformations and color scales to compare perceptual accuracy (Hofmann et al., 2012).

[4] These plots are terrible from a psychological perspective, but serve to illustrate the versatility of the grammar of graphics. In general, stacked density plots, histograms, and dot plots are bad for making numerical comparisons (Cleveland and McGill, 1985).



(a) Stacked Density Plot (b) Dotplot (c) Histogram

Figure 1.24: Three different plots of iris data, created using the grammar of graphics

```
# Stacked Density plot
ggplot(data=iris, aes(x=Sepal.Width, fill=Species)) +
  geom_density(position="stack")

# Dotplot
ggplot(data=iris, aes(x=Sepal.Width, fill=Species)) +
  geom_dotplot(method='histodot', stackgroups=TRUE)

# Histogram
ggplot(data=iris, aes(x=Sepal.Width, fill=Species)) +
  geom_histogram(position="stack")
```

Figure 1.25: ggplot2 code to produce figure 1.24

1.3.3.2 Testing Statistical Graphics using Lineups

One useful tool for testing statistical graphics is the concept of a lineup. Lineups combine the psychological notion of visual search tasks with the statistical concept of hypothesis testing: Participants are provided with a number of plots of the same form, one using real data and the rest generated using resampling methods. If participants identify the target plot (the plot with real data), this is considered similar in nature to a significant hypothesis test at a given α level (generally, there are 20 plots, so $\alpha = 0.05 = 1/20$). Figure 1.26 shows a sample lineup.

In addition to the visual inference protocols lineups were designed to fulfill (Buja et al., 2009), they also provide a method to easily quantify (on a statistical level) the "power" of a plot; if two lineups are generated from the same data, but one allows participants to more

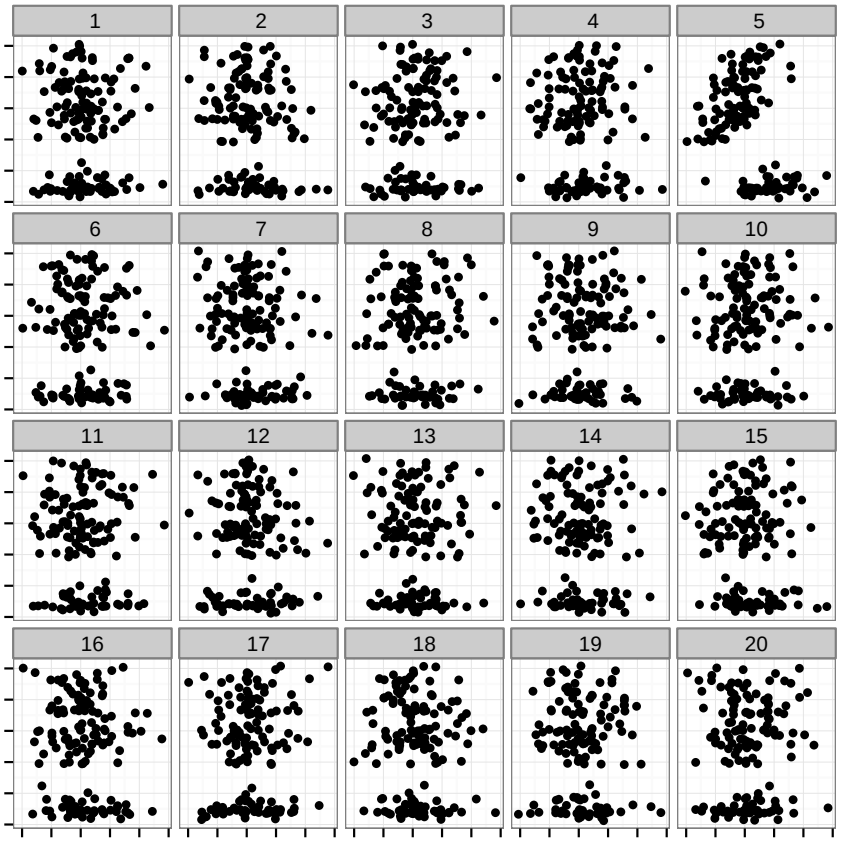


Figure 1.26: Lineup of the iris data, comparing sepal width to petal width. The target data is in plot 5, other plots generated by permuting petal width.

frequently detect the target plot, then that lineup provides more perceptual power. The lineup protocol provides a useful tool for examining some of the issues discussed for complex, domain specific graphs. When combined with the grammar of graphics approach (Wickham et al., 2010), lineups have the potential to be extremely useful for studying the perception of graphs which present the same data in different forms.

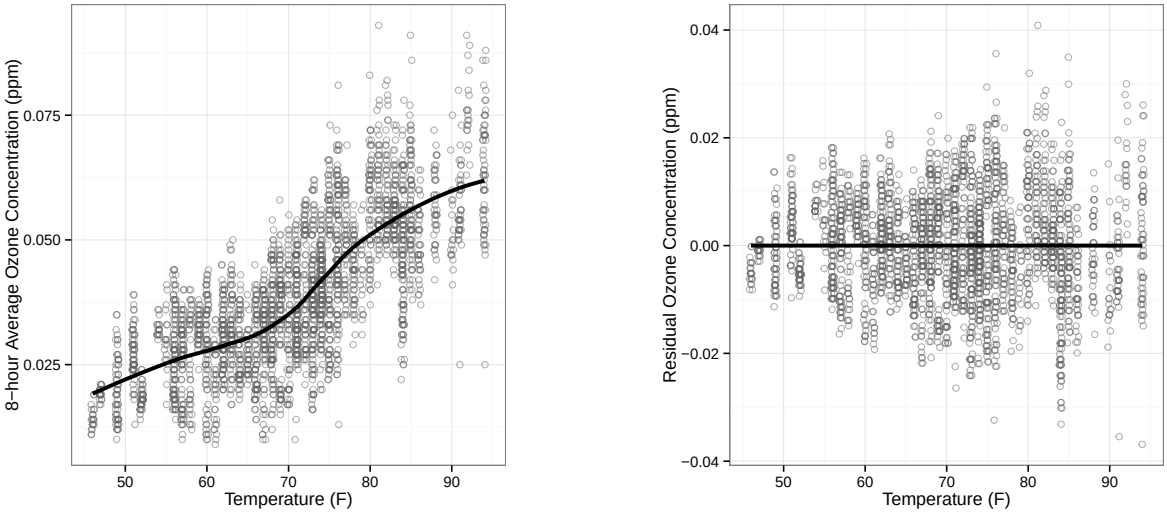
CHAPTER 2. SIGNS OF THE SINE ILLUSION – WHY WE NEED TO CARE

Accepted to the Journal of Computational and Graphical Statistics, July 2014

2.1 Introduction

Graphics are powerful tools for summarizing large or complex data, but they rely on the main premise that any graphical representation of the data has to be "true" to the data (see e.g. Tufte (1991); Wainer (2000); Robbins (2005)). That is, a measurable quantity of a graphical element in the representation has to directly reflect some aspect of the underlying data. Generally, we see a lot of discussion on keeping true to the data in the framework of (ab)using three dimensional effects in graphics. Tufte (1991) goes as far as defining a *lie-factor* of a chart as the ratio of the size of an effect in the data compared to the size of an effect shown, with the premise that any large deviations from one indicate a misuse of graphical techniques. Computational tools help us ensure technical trueness – but this brings up the additional question of how we deal with situations that involve innate inability or trigger learned misperceptions in the audience. In this paper we want to raise awareness for one of these situations, show that it occurs frequently in our dealings with graphics and provide a set of strategies for solving or avoiding it.

As a first example let us consider the relationship between ozone concentration and temperature. Ozone concentrations were measured from 21 locations in the Houston area (Environmental Protection Agency, 2011), and temperature data is provided by the NCDC (National Climate Data Center, 2011) site at Hobby International Airport, located near the center of Houston.



(a) Scatterplot of Ozone and Temperature in Houston, 2011. A loess fit is overlaid to show the overall trend.

(b) Scatterplots of Ozone and Temperature de-trended according to the loess fit in (a).

Figure 2.1: Scatterplots of Ozone and Temperature in Houston, 2011. The increase in variability over the temperature range is more pronounced in the de-trended plot on the right.

Figure 2.1a shows daily measurements of 8-hour average ozone concentration and temperature at several sites in Houston, for days in 2011 with temperatures above 45°F and dew points of less than 60°F. A loess smooth line is added for reference. These types of plots are often used to give an overview of the relationship between two variables. The trendline summarizes this relationship, while the points show raw measurement to allow an assessment of the overall size of the data, the amount of (marginal) variability presented, as well as the (conditional) variability along the trendline. It is the latter task that we cannot satisfactorily complete. While we might agree that there is an increase in variability of ozone concentrations for temperatures above 80°F, we will not doubt homogeneity elsewhere based on figure 2.1a.

This evaluation changes when considering figure 2.1b: the scatterplot shows a loess based de-trended residual of temperature. A previously almost invisible increase in variability of ozone measurements with increasing temperatures now becomes apparent.

This phenomenon, caused by the change in the slope of the trend line, is known as the *sine illusion* in the literature on cognition and human perception or *line width illusion* in the statistical graphics literature.

The illusion is a frequent occurrence in statistical graphics, and displays should therefore be thoughtfully considered to minimize its effect visually and acknowledge its influence. In the cognitive literature, Day and Stecher (1991) first documented the illusion in the context of vertical lines along a sinusoidal curve. Figure 2.2 shows a sketch of this: line segments are centered evenly spaced along the curve. Line segments are of equal length but appear longer in the peaks and troughs due to the illusion. The parameters that influence the strength of the illusion are the amplitude of the curve and the length of the line segments. As the length of the line segments increases, the apparent difference in the length of the line segments decreases. Any modification that increases the change in slope under which the curve appears, such as an increase in the amplitude of the curve or a more extreme aspect ratio, reinforces the apparent difference in line lengths.

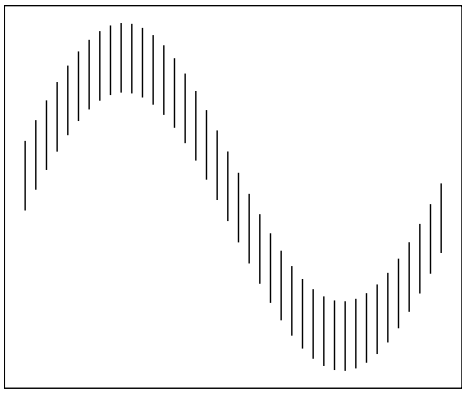
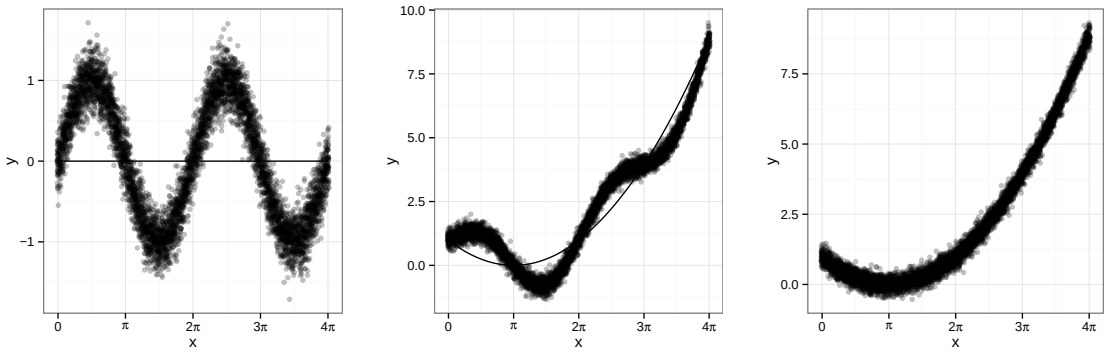


Figure 2.2: The original sine illusion, demonstrated on evenly spaced vertical lines centered around a sinusoidal curve of $f(x) = \sin(x)$. The lines in the peak and trough of the curve appear to be longer than in the other regions.

More recently the illusion has been shown in non-sinusoidal curves (Cleveland and McGill, 1984; Schonlau, 2003; Robbins, 2005; Hofmann and Vendettuoli, 2013), but the underlying effect seems to be the same, in the sense that the illusion is not triggered by the periodic nature of the underlying trendline but only by changes to its slope. Figure 2.3 shows three panels, which all exhibit the illusion. From left to right, the trend stems from (a) a periodic function, (b) a periodic component added to an exponential function, and (c) an exponential function on its own. While all three graphs seem to show nonconstant variance along the main trend; in reality, it is constant. Clearly, the illusion does not rely on the periodicity of the

function for which it was named, but is a symptom of the change in curvature that comes with the periodicity.



(a) Seasonality, No Trend (b) Seasonality and Trend (c) Trend, No Seasonality

Figure 2.3: Set of three scatterplots of simulated data with constant variance. Plot (a) shows seasonality without any underlying trend, (b) shows seasonality superimposed on a quadratic trend, and (c) shows a quadratic trend without seasonality. Though all three sets of simulated data have constant variance, none of the variances appear constant due to the sine illusion.

Next, we give an overview of the perceptual and statistical literature regarding this illusion.

2.1.1 The Sine Illusion in Statistical Graphics

The sine illusion demonstrated in figures 2.1 and 2.2 has been frequently noted in statistical graphics, though usually not as an optical illusion. Rather, the problem is typically identified as the difficulty of visually subtracting two curves, and the resulting erroneous conclusions when this process goes awry. Figure 2.4 presents the possibly oldest example of this common phenomenon (Playfair, 1786; Playfair et al., 2005): Playfair's chart of the balance of trade between England and the East Indies shows time series of the trade value for imports and exports between the countries in the 18th century. The shaded area on the chart is named "balance against England", suggesting that the difference between the lines is of main importance. This difference in trading values is encoded as the difference between the lines along the vertical axis. However, the vertical distance between two lines provides a much less visually salient cue than the orthogonal width between the lines. This results in an underestimation (Cleveland and McGill, 1984) of the difference in trades around 1763, which is of a much higher (about 1.5

fold) magnitude as around 1770, but appears much smaller. In more modern visualizations, bivariate area charts and "stream graphs" (Byron and Wattenberg, 2008) commonly produce the illusion (see an example at http://bl.ocks.org/mbostock/3894205).

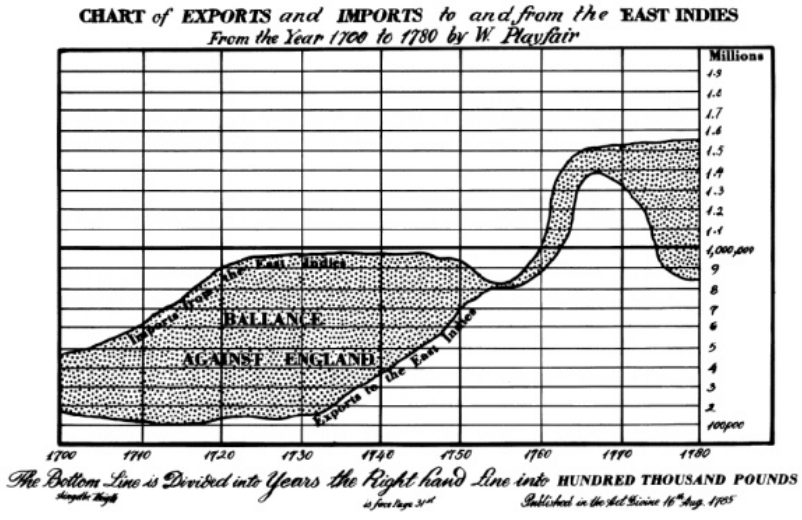


Figure 2.4: Playfair's graph of exports to and imports from the East Indies demonstrates that the line width illusion is not only found on sinusoidal curves but is present whenever the slope of the lines change dramatically. The increase in both imports and exports circa 1763 does not appear to portray as large of a deficit as that in 1710, even though they are of similar magnitude.

2.1.2 Perceptual Explanations for the Sine Illusion

While not thoroughly examined in the sensation and perception literature, the sine illusion has been classified as part of a group of geometrical optical misperceptions related to the Müller-Lyer illusion (Day and Stecher, 1991) or the Poggendorf illusion (Weintraub et al., 1980), which puts the illusion into the framework of context-based illusions. Day and Stecher (1991) suggest that the sine illusion occurs due to misapplication of perceptual experience with the three-dimensional world to a two-dimensional "artificial" display of data.

Experience with real-world objects suggests that the stimulus of figure 2.2 is very similar to a slightly angled top view of the 3-dimensional figure of a strip or ribbon describing waves in a third dimension, such as e.g. a road does on rolling hills. This is sketched out in figure 2.5a. Our experience suggests immediately that changes in the width of the road are unlikely and

resolves the illusion. While figure 2.5a shows the line segments slightly angled towards each other, figure 2.5b shows a variation of the same plot with a vanishing point set further away from the viewer. This makes the line segments almost parallel to each other and therefore more closely resembles the sketch of figure 2.2, in which the sine illusion was originally presented.

(a) Perspective plot of sine illusion

(b) Perspective plot, vanishing point near infinity.

Figure 2.5: Two different perspective projections of the same data responsible for the sine illusion. The first projection angles the lines and appears more natural, but the second projection suggests that the lines do not need to be angled to create the same three-dimensional impression.

Recreating the three-dimensional context of the sine illusion might resolve the distortion, even if increasing the dimensionality of a graph is generally not recommended (Tufte, 1991; Cleveland and McGill, 1984) (though Spence (1990) suggests that in certain cases additional dimensions are not misleading). While creating a three-dimensional projection of two-dimensional data might counteract the illusion, the process of projecting the data accurately into a higher dimension is not simple. The projection that best resolves the illusion likely is highly subjective and influenced by choices of angle and color gradient for depth cues. As there is not a single three-dimensional projection that corresponds to the two-dimensional data, this approach would only produce further visual ambiguity.

Further complicating the situation, the illusion itself is insidious – we trust our vision implicitly, to the point that when we understand something, we say "I see". This trust in our visual perception is seldom called into question, for our perception is optimized for interaction with a three-dimensional world. Artificial two-dimensional situations (such as graphs and pictures) may accurately represent the data and still produce a misleading perceptual experience.

The contextual cues of the overall trend are critical to the sine illusion's effect; the illusion only holds when a substantial portion of the graph is considered simultaneously, which

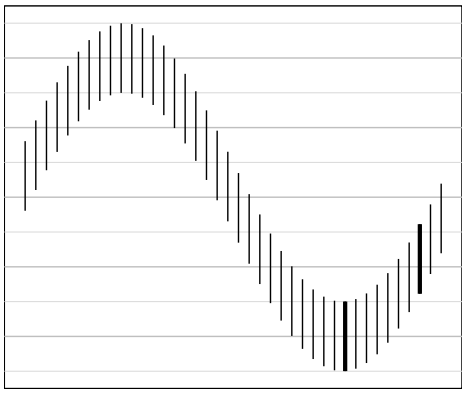


Figure 2.6: The sine illusion with two individual lines highlighted. Horizontal grid lines do not help to resolve the illusion, even though they provide a clear basis for comparison of line lengths. Readers are much better at assessing the length of the two singled out line segments; they are equal.

triggers our innate ability of perceiving one whole rather than the individual parts it consists of (principle of grouping; Wolfe et al. (2012)). Considering only two line segments at a time resolves the illusion. The bold lines in figure 2.6 are clearly of the same length. Comparisons of individual line lengths is visually a fairly simple task, and is done with a relatively high accuracy (Cleveland and McGill, 1984). Day and Stecher (1991) contains a more thorough discussion of how much surrounding context is required for the illusion to persist.

2.1.3 Geometry of the Illusion

In figure 2.2 we have seen that the our preference in evaluating line width is to assess *orthogonal* width rather than the difference along the vertical axis. Figure 2.7 demonstrates the change in orthogonal width as the slope of the line tangent to the graph of f changes; these changes correspond to our perception of apparent line length.

The illusion is most pronounced in regions where the angle between the orthogonal and the vertical line is large. Changes to the aspect ratio therefore have a major impact on the strength of the sine illusion. Any change that alleviates the difference between perceived width and the perpendicular width, such as banking to 45° (Cleveland et al., 1988), will alleviate the effect but not completely overcome it. The perceived length of the vertical line changes with the angle of the line perpendicular to the slope of $\sin(x)$, suggesting that the sine illusion stems

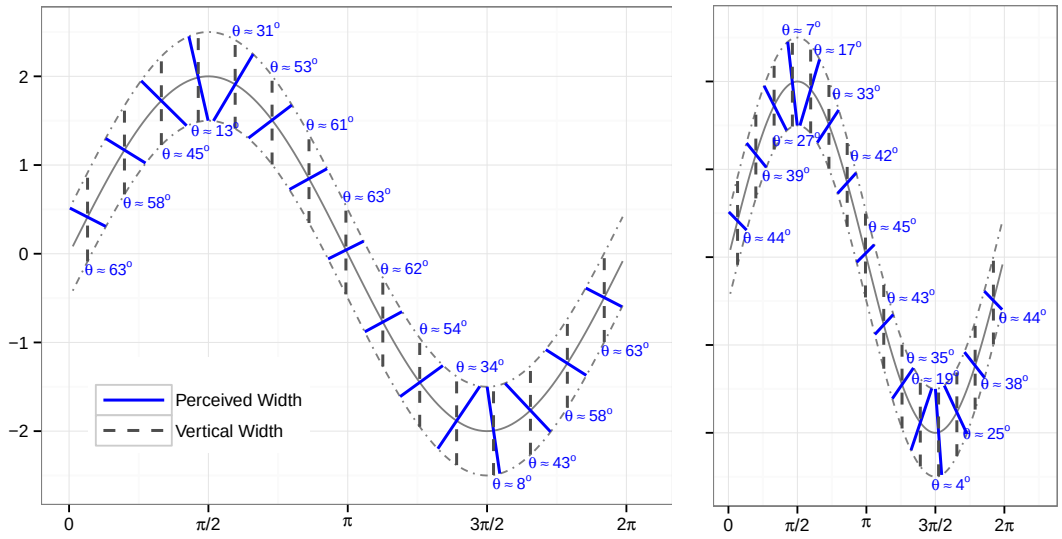


Figure 2.7: The sine illusion with lines orthogonal to the tangent line at $f(x)$. The perception that the vertical length changes with $f(x)$ corresponds to changes in actual orthogonal width due to the change in the visual (plotted) secant angle. The strength of the perceptual effect depends in part on the aspect ratio of the graph, as shown in the second image, which has an aspect ratio of 2 compared to the first figure's aspect ratio of 1. This correspondingly multiplies the strength of the effect by 2.

from a conflict between the visual system's perception of figure width and the mathematical judgement necessary to determine the length of the vertical lines.

Our preference for assessing figure width based on the orthogonal width suggests that the underlying illusion may be a function of geometry rather than some unknown visual or neural process that occurs subconsciously. In this case it may be possible to correct the graphical display for the illusion to minimize its misleading effect. A geometrical correction that –at least temporarily– counteracts the illusion would be a valuable tool in visual analysis, as this illusion very persistently affects our judgment of very common tasks such as e.g. the assessment of conditional variability of data along a trend line.

Simply raising people's awareness of the presence of this illusion is not enough, as it is incredibly difficult, if not impossible, to overcome this illusion even when we are aware of its presence: our brains simply cannot "un-see" it.

What follows is a compilation of several approaches to correct for or mitigate the effect of

the illusion. Our primary intention here is to demonstrate the persaviness of the illusion is and the extreme measures necessary to remove its effect.

2.2 Breaking the Illusion

The sine illusion is caused by a conflict between vertical width, which is the width that we want onlookers to assess visually, and orthogonal width, which is the width that the onlooker perceives. This difference can be expressed as a function in the slope of the underlying trend line. This provides the basis for adjusting the vertical width for the perceived orthogonal width.

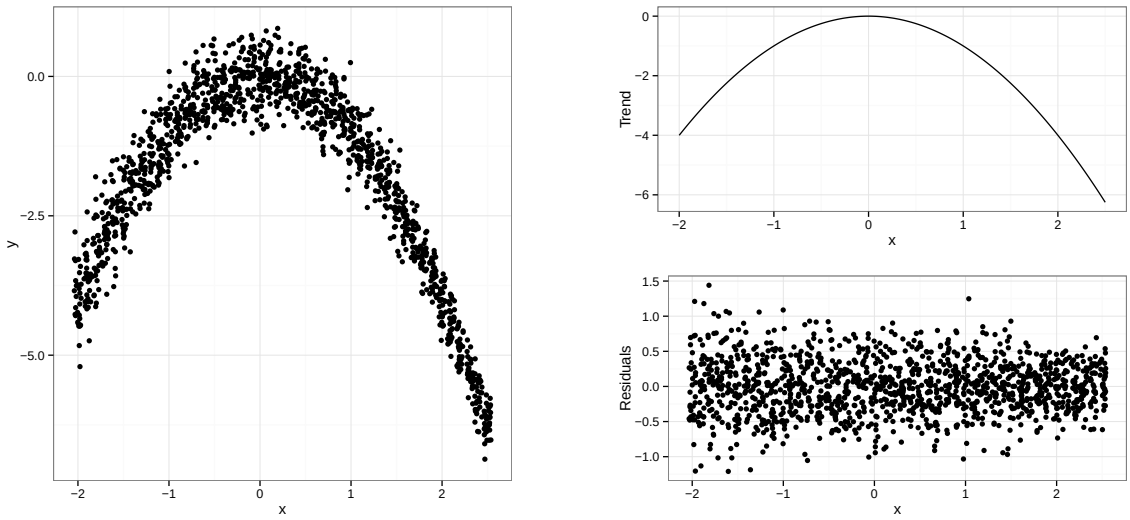
We consider the following three approaches:

1. separating the trend and the variability,
2. transformation of x: adjusting the slope to be constant by reparametrizing the x axis, and
3. transformation of y: adjusting y values to make conditional variability appear correctly by adjusting accordingto orthogonal width.

Each of these ideas is discussed in more detail in this section.

2.2.1 Trend Removal

Cleveland and McGill (1984, 1985) discuss the perceptual difficulty of judging the difference between two curves plotted in the same chart, and alternatively, recommend to display the difference between the two curves directly. This is in line with recommendations for good graphics to 'show the data' rather than make the reader derive some aspect of it (e.g. Wainer (2000)). In particular, de-trending data to focus on residual structure is the generally accepted procedure for assessing model fit. Figure 2.8(a) shows a scatterplot of data with a trend. A loess smooth is used to estimate the trendline. A visual assessment of variability along this trendline might result in a description such as 'homogeneous variance or slightly increasing variance for negative x, followed by a dramatic decrease in vertical variability for positive x'. Once the residuals are separated from the trendline as shown on the right hand side of the



(a) Data

(b) Trend and Residuals

Figure 2.8: Describe the conditional variability of the points along the x axis in (a). Is your description consistent with the residual plot in (b)?

figure, it becomes apparent that this first assessment of conditional variability was not correct, and the decreasing variance along the horizontal axis becomes visible.

While the illusion is not apparent when trend line and variability in the residual structure are shown separately, the separation makes it more difficult to evaluate the overall pattern in the data, as we must base any judgment on two charts; either by combining information from two graphs or by mentally re-composing the original graph (at which point, the sine illusion becomes a factor). To minimize cognitive demands we ideally want to tell the whole story with a single graph, in particular because in many situations we may not be able to show multiple graphs.

Additionally, removing the trend requires an initial model, making any plots produced using that fit conditional on the assumptions necessary to obtain that model fit. In many situations, this may be undesireable. In particular, we typically view the data before fitting even a rudimentary model, and the sine illusion may influence even these initial modeling decisions.

2.2.2 Transformation of the X-Axis

As the sine illusion is driven by changes in the slope of trends between variables, we can counteract the illusion by removing these changes, transforming the x axis such that the absolute value of the slope is constant and forcing the corresponding orthogonal width to represent the conditional variability. In order to describe this transformation of the x axis mathematically, let us assume that the relationship between variables X and Y is given by a model of the form

$$y = f(x) + \varepsilon,$$

where f is some underlying function (either previously known or based on a model fit). Further let us assume that f is differentiable over the region of observed data.

Ideally, the correction would force all lines to appear under the same slope, i.e. we want to find a transformation $T(x)$ of x, such that $f(T(x))$ is a piece-wise linear function, where each piece has the same absolute slope. This transformation has an effect similar to "banking to 45°" in a piecewise manner.

Let a and b be the minimum and maximum of the x-range under consideration. Then for any value $x \in (a, b)$ the following transformation results in a function with constant absolute slope:

$$(f \circ T)(x) = a + (b - a) \left(\int_a^x |f'(z)| dz \right) / \left(\int_a^b |f'(z)| dz \right), \tag{2.1}$$

2.2.2.1 Derivation of the X Transformation

As the slope is determined by the aspect ratio, we are free to choose it and w.l.o.g. we get for each piece T_i:

$$f(T_i(x)) = \pm ax + b_i.$$

This means that T_i is essentially an inverse of function f, with each piece defined by the intervals on which the inverse of f exists: let $\{x_0 = \min(x), x_1, ..., x_{K-1}, x_K = \max(x)\}$ be the set of values with local extrema enhanced by the boundaries of the x-range, i.e. $f'(x_i) = 0$ for $i = 1, ..., K - 1$ and $f'(x) \neq 0$ for any other values of x. Then each interval of the form (x_{i-1}, x_i) defines one piece T_i of the transformation function $T(x)$. We will define T_i now as a

combination of a linear scaling function and the inverse of f, which we know exists for interval (x_{i-1}, x_i).

Let function $s = {}_{[a,b]}s^{[c,d]}$ be the linear scaling function that maps the interval (a, b) linearly to the interval (c, d). This function is formally defined as

$$s(x) = {}_{[a,b]}s^{[c,d]}(x) = (x - a)/(b - a) \cdot (d - c) + c \text{ for all } x \in (a, b).$$

Note that the slope of function s is given as

$$s'(x) = (d - c)/(b - a).$$

Two scaling functions can be evaluated one after the other, only if the image (i.e. y-range) of the first coincides with the domain (i.e. x-range) of the second. This consecutive execution results in another linear scaling:

$${}_{[e,f]}s^{[c,d]}\left({}_{[a,b]}s^{[e,f]}(x)\right) = {}_{[a,b]}s^{[c,d]}(x)$$

In our situation let the scaling function s be given as:

$${}_{[c,d]}s^{f([x_{i-1},x_i])}(x) = f(x_{i-1}) + (x - c)/(d - c) \cdot (f(x_i) - f(x_{i-1}))),$$

where $f([x_{i-1}, x_i])$ is defined as the interval given by $(\min(f(x_{i-1}), f(x_i)), \max(f(x_{i-1}), f(x_i)))$. Note that s has either a positive or negative slope depending on whether $f(x_{i-1})$ is smaller or larger than $f(x_i)$, respectively.

Then the transformation in the x-axis, $T(x)$ is defined piecewise as a combination of T_i, where each T_i is given as:

$$T_i(x) \quad = \quad f^{-1}\left({}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x)\right). \tag{2.2}$$

Using this definition for the transformation makes $f(T(x))$ a piece-wise linear function with parameters c_i and d_i, i.e. for $x \in (c_i, d_i)$ we have

$$f(T(x)) = f(f^{-1}({}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x))) = {}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x).$$

Correspondingly, the slope of $f(T_i(x))$ is $(f(x_i) - f(x_{i-1})))/(d_i - c_i)$. In order to make the slope the same on all pieces T_i of T, we need to define c_i and d_i with respect to the function values

on the interval (x_{i-1}, x_i). There are various options, depending on how closely the x-range of T should reflect the original range: for $[c_i, d_i] = \text{range}\,(f([x_{i-1}, x_i]))$ the new x-range is the range of f on (x_{i-1}, x_i), but with the advantage that the scaling function simplifies to the identity or a simple shift.

In order to preserve the original x-range, we need to invest into a bit more work for the scaling. With an identity scaling, each T_i maps from the range of f on (x_{i-1}, x_i) to the same range. Overall we can therefore set up the function T to map from the interval given by the sum of the function's 'ups' and 'downs', i.e. $(0, \sum_{i=0}^{K} |f(x_i) - f(x_{i-1})|)$, to the range of f on (x_0, x_K). This ensures that all pieces $f(T_i)$ have the same slope (of $|1|$).We can then use another - global - linear scaling function to map from the range of x, i.e. interval (x_0, x_K) to $(0, \sum_{i=0}^{K} |f(x_i) - f(x_{i-1})|)$, yielding a transformation function T of

$$T(x) = (f^{-1} \circ {}_{[c_i,d_i]}s^{f([x_{i-1},x_i])} \circ {}_{(x_0,x_K)}s^{(0,\sum_{i=0}^{K}|f(x_i)-f(x_{i-1})|)})(x),$$

where c_i and d_i are given as

$$c_i = \sum_{j=0}^{i-1} |f(x_j) - f(x_{j-1})| \text{ and } d_i = \sum_{j=0}^{i} |f(x_j) - f(x_{j-1})|.$$

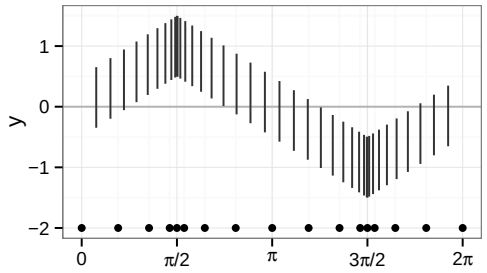
We can write the difference $|f(x_j) - f(x_{j-1})|$ as $\int_{x_{j-1}}^{x_j} |f'(z)|dz$. This shows equation (2.1).

2.2.2.2 Weighting the X Transformation

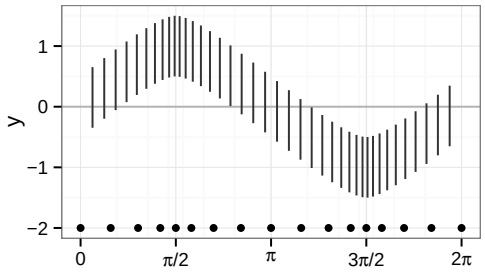
As the sine illusion depends on changing slope in the overall trend, re-parametrizing the x-axis in terms of the slope will make the data appear under a constant slope, thereby removing the effect of the illusion, while the transformed x-axis is changed from a linear representation of the x values to a 'warped' axis that continuously changes the scale of x to compensate for the changes in the slope. To emphasize this change in scale along the x axis, dots are drawn at the bottom of the chart to show the transformation's effect on equally spaced points along the x-axis.

Results from this transformation are demonstrated in Figure 2.9a.

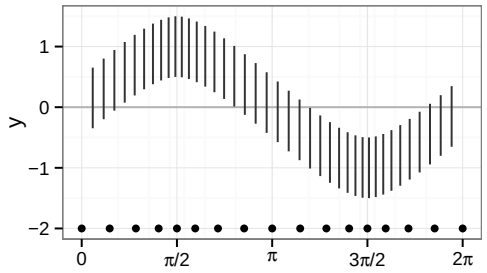
While the transformation in equation (2.1) effectively removes the appearance of changing line lengths, we can see in practice that the illusion can be broken by a much less severe



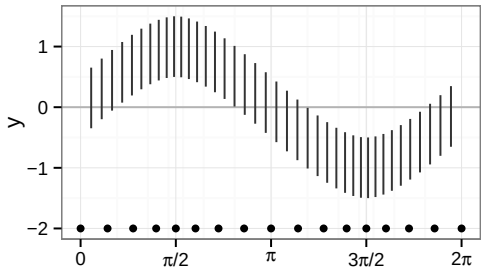
(a) X axis transformation based on eqn. (2.1), corresponding to weighting of $w = 0$.



(b) Weighted Transformation, $w = 1/2$ (based on eqn. (2.3))



(c) Weighted Transformation, $w = 1/3$



(d) Weighted Transformation, $w = 1/4$

Figure 2.9: Examples of X axis transformations in the sine curve. Dots at the bottom of the graph show the transformation's effect on equally spaced points along the x-axis. Different amounts of weighting w correspond to differently strong corrections. In (a), x-spacing of the lines changes the extant width such that the absolute value of the slope is uniform across the whole range of the x axis resulting in the largest amount of correction. (b) - (d) reduce the correction in (a) towards successively more uniform spacings in x while still breaking the effects of the illusion.

transformation of the x axis. For that we introduce a shrinkage factor $w \in (0, 1)$ that allows a weighted approach in counteracting the illusion as:

$$(f \circ T_w)(x) = (1 - w) \cdot x + w \cdot (f \circ T)(x) \tag{2.3}$$

Note that for $w = 1$ the x-transformation is completely warped, while smaller values of w indicate a less severe adjustment against the sine illusion. Under weaker transformations, the data more closely reflect the original function $f(x)$. Figures 2.9b - 2.9d show the effect of different shrinkage coefficients w. As w decreases, the lines become more evenly spaced and the illusion begins to return.

The extent to which we can shrink the adjustment back to the original function varies with the aspect ratio of the chart and the shape of the function. It might also be influenced by the audience's experience with the sine illusion, resulting in very subjective choices of an "optimal

weighting" for specific situations which minimizes distortion and maximizes the correspondence between inferences made from the data and inferences made using the visual display.

Note, that we only make use of the transformation T in the form of $f \circ T$. This allows us to avoid an explicit calculation of the transformation T, which in particular involves a computation of the inverse of f leading to potentially computation-intense solutions.

2.2.2.3 X Transformation Demonstration

In the example of the Ozone data shown in figure 2.1, we can base a transformation of the x-axis on a loess fit of ozone concentration in daily temperature. Loess is particularly convenient for this transformation, as it enforces continuity conditions including differentiability of the fitted function; software allows us to obtain fits of both the function values and their derivatives.

Figure 2.10 shows the original data side-by-side with the transformed x-axis, demonstrating not only the effect of transformation of the x-axis, but also that the transformation is not overly misleading in this example. The granularity of the data in this example provides an implicit measure of the strength of the transformation along the x-axis and the transformation is also clearly evident in the labels along the x-axis.

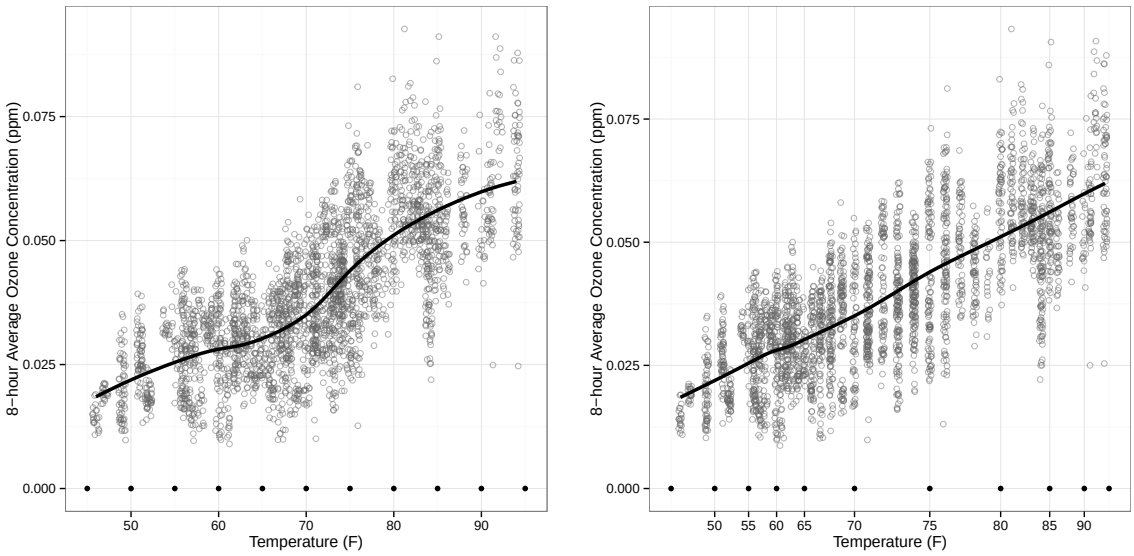
2.2.3 Transformation in Y

Understanding the geometry of the sine illusion leads to another approach to counteracting the conflict between the orthogonal width and the vertical length of the segment.

Let again the function f describe the general relationship between variables X and Y.

As sketched out in figure 2.11 we want to first find the orthogonal (extant) width in a point $(x_0, f(x_o))$ on the graph, which corresponds to the perceived width, and then correct the vertical width accordingly to match with the audience's expectation.

The orthogonal width (see sketch in figure 2.11) is given as the line segment between endpoints $(x_1, f_1(x_1))$ and $(x_2, f_2(x_2))$, where f_1 and f_2 denote the vertical shifts of function f by $-\ell/2$ and $\ell/2$, respectively, where ℓ is defined as the overall line length, $\ell > 0, \ell \in \mathbb{R}$.



(a) Original Data

(b) Transformed X Axis

Figure 2.10: Original data and data after x-transformation. The increasing variance is easier to see when x has been transformed, because the slope is now uniform.

These endpoints are determined as the intersection of the line orthogonal to the tangent line in $(x, f(x))$ and the graphs resulting from the vertical shifts of f.

The function describing the orthogonal line through $(x_o, f(x_o))$ is given in point-vector form as

$$\begin{pmatrix} x_o \\ f(x_o) \end{pmatrix} + \lambda \begin{pmatrix} f'(x_o) \\ 1 \end{pmatrix},$$

for any real-valued λ. The advantage of using point vector form is that it allows us to solve for

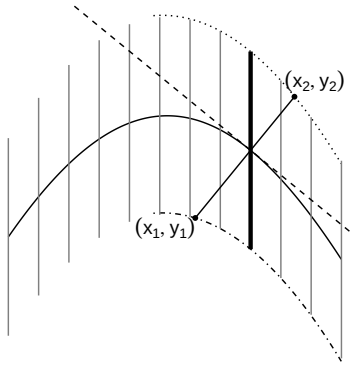


Figure 2.11: General correction approach. This approach may require numerical optimization to obtain exact solutions for (x_1, y_1) and (x_2, y_2).

parameter λ easily, which gives us easy access to the extant (half-)widths, as:

$$|\lambda|\sqrt{1+f'(x_o)^2}. \tag{2.4}$$

Eqn. (2.4) describes the quantity that we perceive rather than the quantity that we want to display ($\ell/2$), which leads us to a general expression of the correction factor as

$$\ell/2 \cdot \left(|\lambda|\sqrt{1+f'(x_o)^2}\right)^{-1}.$$

Note that this yields in general two solutions: one for positive, one for negative values of λ corresponding to upper and lower (half-)extant width.

In order to get actual numeric values for λ, we need to find end points of the extant line width as solutions of intersecting the orthogonal line and the graphs of f_1 and f_2. We find these end points as solutions in x and λ of the system of equations:

$$x - x_o = \lambda f'(x_o) \tag{2.5}$$

$$f(x) - f(x_o) = -\lambda \pm \ell/2 \tag{2.6}$$

Note that the above system of equations involves function values $f(x)$, which implies that solving this system requires numerical optimization for any but the most simple functions f.

In the following two sections we make use of Taylor approximations of first and second order to find approximate solutions to end points as sketched out in figure 2.12.

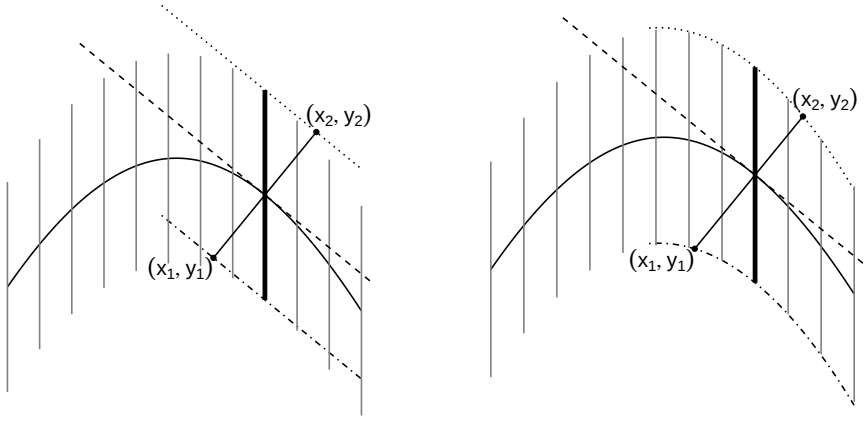
2.2.3.1 Linear Approximation to $f(x)$

For the linear approximation we make use of $f(x) \approx f(x_0) + (x - x_0)f'(x_0)$, which together with equations 2.5 and 2.6 yields a correction factor in x_0 of

$$\ell_{\text{new}}(x_0) = \ell_{\text{old}}\sqrt{1+f'(x_0)^2}.$$

Note that the linear method gives the same result as a varying slope extension from a trigonometric approach suggested by Schonlau (2003) and used in Hofmann and Vendettuoli (2013).

A second-order Taylor polynomial approximation to $f(x)$ additionally accounts for the asymmetry in the extant widths on either side of the center trendline.



(a) Linear Approximation (b) Quadratic Approximation

Figure 2.12: (a) uses a first-order Taylor series approximation to $f(x)$ and (b) uses a second-order Taylor series approximation to $f(x)$. The intersection of the function $f(x) \pm \ell/2$ and the orthogonal line, $(x_1, y_1), (x_2, y_2)$ must be obtained to determine the necessary correction factor.

2.2.3.2 Quadratic Approximation to $f(x)$

Using the approximation $f(x) \approx f(x_0) + f'(x_0)(x - x_0) + 1/2 f''(x_0)(x - x_0)^2$, the system of equations 2.5 and 2.6 simplifies to the following quadratic equation in λ:

$$f''(x_0)f'(x_0)^2\lambda^2 + 2(f'(x_0)^2 + 1)\lambda \pm \ell = 0,$$

which leads us to corrections for the half lengths as:

$$\ell_{\text{new}_1}(x_0) = 1/2 \cdot \left(v + \sqrt{v^2 + f''(x_0)f'(x_0)^2 \cdot \ell_{\text{old}}}\right) \cdot v^{-1/2} \tag{2.7}$$

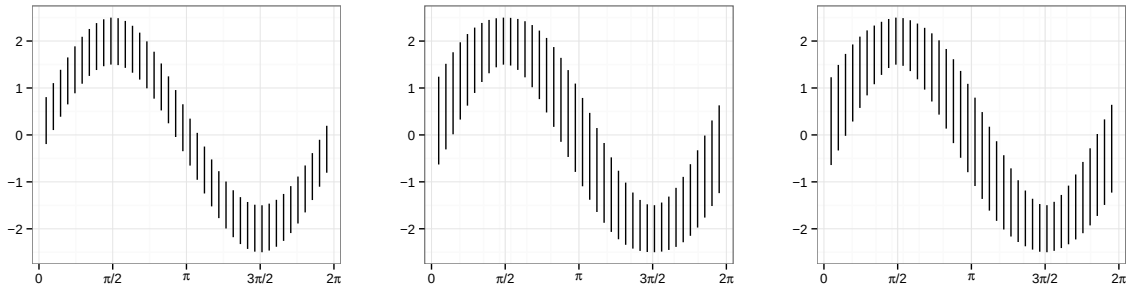
$$\ell_{\text{new}_2}(x_0) = 1/2 \cdot \left(v + \sqrt{v^2 - f''(x_0)f'(x_0)^2 \cdot \ell_{\text{old}}}\right) \cdot v^{-1/2} \tag{2.8}$$

where $v = 1 + f'(x_0)^2$.

2.2.3.3 Reformulation of the quadratic approximation

A quadratic equation in λ of the form

$$a\lambda^2 + b\lambda + c = 0, \tag{2.9}$$



(a) Uncorrected (b) Linear correction (c) Quadratic correction

Figure 2.13: In the quadratic approximation top and bottom segments of the vertical lines are adjusted separately.

where a, b, and c are real-valued parameters the solutions take on the form

$$\lambda_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \overset{*}{=} 2c \left(-b \pm \sqrt{b^2 - 4ac}\right)^{-1}.$$

* if $b \neq \pm \sqrt{b^2 - 4ac}$, i. e. $a, c \neq 0$.

Application to quadratic approximation to f: in the example, we have the following equivalencies:

$$\begin{aligned} a &= f''(x_0) f'(x_0)^2 \\ b &= 2(1 + f'(x_0)^2) \qquad > 0 \text{ for all } x \\ c &= \pm \ell \end{aligned}$$

For a valid solution for the correction factor, we have to assume that λ is a factor that extends the original extant width (in absolute value).

$$\lambda_{1/2} = \ell \left(v + \sqrt{v^2 \pm f''(x_0) f'(x_0)^2 \cdot \ell}\right)^{-1}$$

for $v = 1 + f'(x_0)$. This gives the results as shown in equations (2.7) and (2.8)

Adjusting the top and bottom segments of the vertical lines separately so that the extant width is constant breaks the illusion, but slightly distorts the sinusoidal shape of the peaks.

Figure 2.13 shows the correction factor based on a quadratic approximation compared to the untransformed data. Unlike the linear solution, the half-segments here are not necessarily of the same length, and thus there are separate correction factors for each half-segment.

2.2.3.4 Mathematical Properties of the Y Transformation

The quadratic correction breaks whenever the expression in the square root of eqn. (2.7) becomes negative, i.e. whenever $v^2 \pm \ell \cdot f''(x) \cdot f'(x)^2 < 0$. This happens for combinations of large values of ℓ, which signify a large vertical extent, or large conditional variability $E[Y|X]$, and simultaneous large changes in the slope of the main trend, i.e. large values of the curvature $f''(x)$. In the linear approximation of f the same situation leads to a massive overcorrection of the vertical lines, changing the shape of the 'corrected' function beyond recognition.

Similar to the correction of the x-axis, we can use a weighted approach to find a balance between counteracting the illusion and representating the original data:

$$\ell_{new_w}(x) = (1 - w) \cdot \ell_{old} + w \cdot \ell_{new}(x) \tag{2.10}$$

2.3 Transformations in Practice – a User Study

In order to more fully understand the sine illusion and test the proposed corrections, we created an applet to allow users to investigate the illusion's prominence with respect to its parameters. Users can examine the sine illusion by changing line length, the function's amplitude, and compare corrections in x-axis and y-values to uncorrected data. All corrections proposed in this paper are implemented in the applet located at http://glimmer.rstudio.com/srvanderplas/SineIllusion/.

We employed a second applet to collect data on users' preferences on the amount of correction used, i.e. we are interested in identifying a range of 'optimal weights' in each of the corrections. This applet presents users with a graph that is the result of a correction in x or y with a randomly selected starting weight value. Users are asked to adjust the graph until the illusion (a) is no longer apparent (adjustment of weights from the bottom) or (b) becomes visible (adjustments of weights from the top).

Both applets are implemented in `shiny` (RStudio Inc., 2013).

The graphs in the data-collection applet are adjusted using a plus/minus button to either increase or decrease the amount of correction used. Underlying this adjustment is the value of

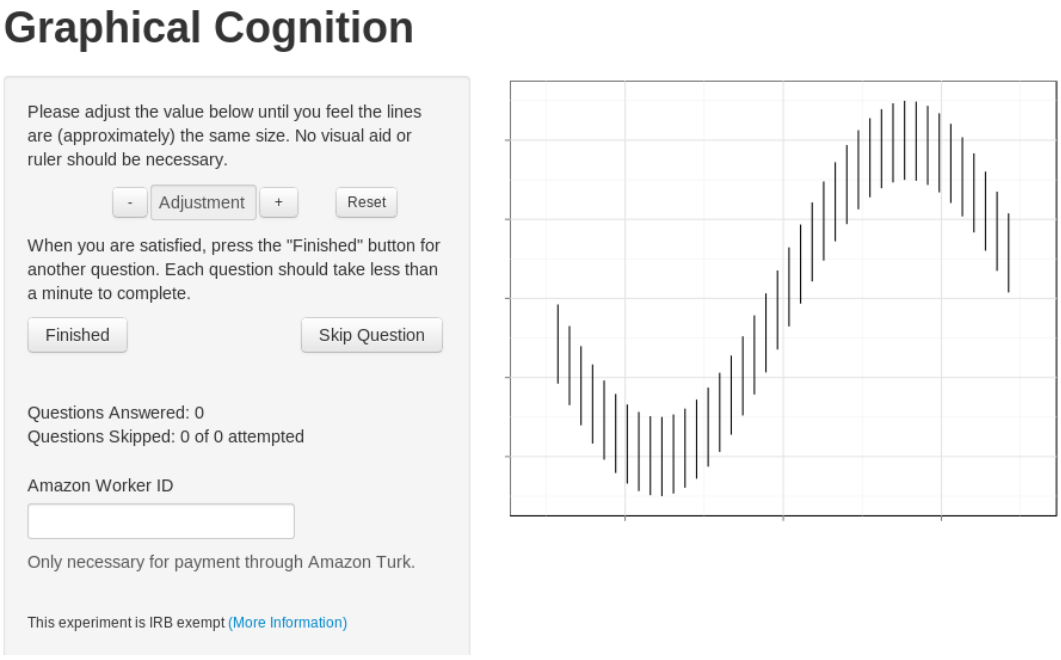


Figure 2.14: Screenshot of the shiny application used to collect information of observers' preference with respect to an optimal correction for the illusion under each of the transformations discussed in the previous section.

the weight w as defined in eqns. (2.3) and (2.10). The numerical value of w was hidden from the user to prevent anchoring to a specific numerical value.

A low initial weight (w_0 close to 0) indicates that the amount of correction is low and the response from a trial like this will give us an idea of the minimal amount of weight necessary to break the illusion, while a high initial weight (w_0 close to 1) indicates that the data is fully corrected. We asked participants to change the amount of adjustment until the lines appear to be the same length assumes that the correction is overcorrecting in practice, and a response from this type of trial gives us an upper boundary for the amount of weighting preferred. Generally, responses from the two different types of trials do not result in the same threshold weight, but rather lead to a range of acceptable weights.

It is of additional interest to determine whether and how much these optimal weights are subject-specific or population-based, whether they depend on the initial weight, and how much within-subject variability we find compared to between-subject variability.

Figure 2.14 shows a screenshot of the applet used to collect user data. This applet is available online at http://glimmer.rstudio.com/srvanderplas/SineIllusionShiny/. Line length

and function are controlled in this app, and we used the linear transformation for adjusting y values; the transformation does not break under any combination of parameters tested in this experiment.

We deployed the applet to participants recruited online, collecting their responses and other metadata. The results of the analysis suggest that the correction factors in X and Y are both preferrable to uncorrected data, but that a full correction is not necessary to break the illusion.

2.3.1 Study Design

The study aims to determine the range of "optimal" transformation weights for each transformation type. Psychophysics methodology typically approaches threshold estimation by using the method of adjustment (Goldstein, 2009b), where stimuli are provided showing states both above and below the hypothesized optimal value and participants adjust the stimuli until the stated goal is met (in this case, until the lines appear to have equal length). It is expected that there will be a difference in user-reported values from below and from above, and these values are typically averaged to produce a single threshold value. Beyond averaging these values, we use a mixed model to compare user responses for different starting points in a more continuous fashion, incorporating some of the advantages of the method of constant stimuli to more robustly estimate the range of optimal transformation weights. For a review of general psychophysics methodology, the method of adjustment, and the method of constant stimuli, see Goldstein (2009b).

The study is set up as a fractional factorial design of correction type (x or y correction) and starting weight w_0. Each participant is asked to evaluate a total of twelve situations, six of each correction type. Starting weights were chosen as follows: each user was given a trial of each type starting at 0 and 1. The remaining four trials of each type had starting weights chosen with equal probability from 0.25 to 0.75 (see figure 2.15). We decided to have a higher coverage density for starting weights around 0.6 after a pilot study indicated a preference for that value. Using a distribution with a wide coverage allows us to more fully explore the space of plausible weights w while focusing on the $(0, 1)$ interval and enabling precise estimation of the optimal weight in the region indicated by the pilot study.

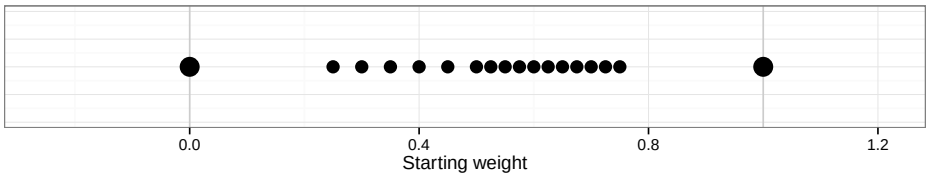


Figure 2.15: Overview of possible starting weights. Weight values are discrete, but staggered so as to provide fine-grained adjustments around 0.6 and more coarse discriminatory information toward the outside.

A trial begins with the presentation of a graph at the chosen starting weight w_0. Participants adjust the graph using increment and decrement buttons. A trial ends with the participant clicking the 'submit' button, at which point the weight for the final adjustment is recorded. This provides a clear starting value and ending value, allowing us to assess the range of optimal values for each participant. In addition to starting weight, correction type, and anonymized user-specific data (partial IP address, hashed IP address, and hashed browser characteristics), each incremental user chosen weight is recorded with a corresponding timestamp. The user-specific browser data is sufficient to provide a 'fingerprint' to distinguish and recognize individual users (or rather their computer settings) in an anonymous fashion.

Each participant is provided with two initial "training" trials in which the graph of the underlying mean function is superimposed on the line segments to give participants some idea of the function the lines represent. This approach was taken to reduce incidences of extremely high correction values under the X transformation, as large adjustment values do not change the impression of same line length, but the resulting function bears little resemblance to a sine function, see figure 2.16 for examples of overcorrection.

2.3.2 Results

Participants were recruited from Amazon Mechanical Turk and the reddit community.

As this study was conducted outside a laboratory setting, we can not gauge a participant's willingness to follow the guidelines and put in their best effort. This, besides potential technical issues (server outage, speed of response) make a careful selection of data going into the analysis unavoidable. The following exclusion criteria were used:

- Participants did not interact with the applet: we required participants to use the adjust-

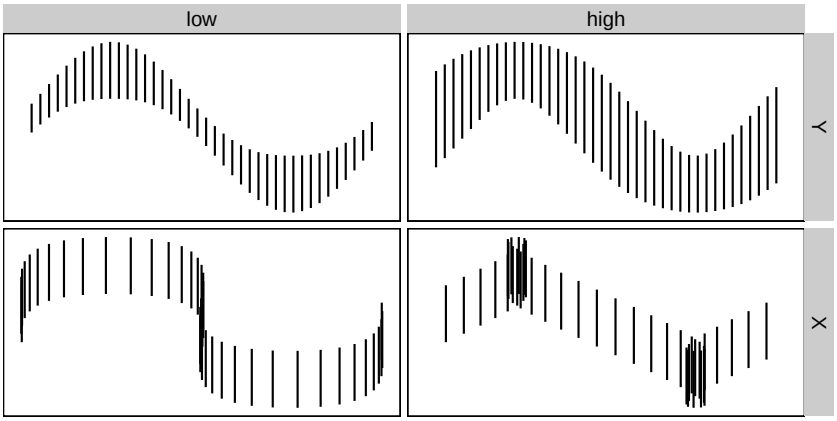


Figure 2.16: Transformation weights outside of the intervals $[-2.5, 3.5]$ for y and $[-2, 2]$ for x produce figures which do not maintain the underlying function shape (in x) or which are composed of extremely uneven length lines (in y). Trials with final results that were more extreme than these examples were excluded from the analysis.

ment at least once in order to include data for this trial (592 trials removed).

- Participants finished fewer than four trials: while participants were asked to complete twelve trials, some did not finish all of those. In order to stabilize predictions of random effects, participants' data were excluded if there were fewer than four trials (78 out of a total of 203 participants).

- Out-of-bounds results: weights leading to severely over- or undercorrected results were excluded from the analysis. For trials to adjust Y-values, weights outside of $[-2.5, 3.5]$ show dramatically unequal line lengths; weights from X-transformations outside the range of $[-2, 2]$ do not preserve the underlying function shape and concavity. Figure 2.16 shows results at the threshold of acceptability. Only more severely distorted results were excluded from the analysis (12 of the X and 5 of the Y trials out of 1227 trials remaining after application of other criteria).

The following analysis is based on the cleaned data, consisting of 125 participants with 1210 valid trial results. The psychophysics model shown in figure 2.17 is based on weighted averages (by adjustment type) of all trials with starting weights $w_0 = 0$ and 1.

According to this analysis, the optimum transformation value for x is 0.35, and the optimum transformation value for y is 0.45. Figure 2.17 shows the estimates and 95% Wald intervals for

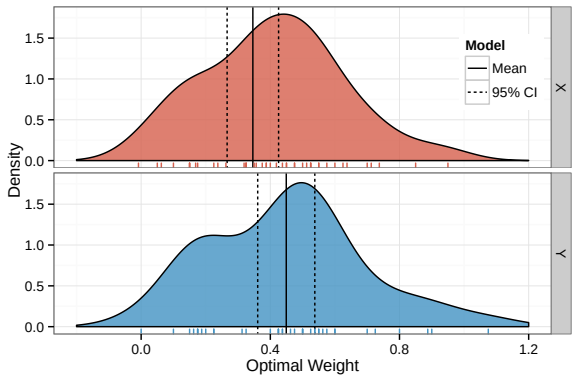


Figure 2.17: Estimated density of participant-level means using the standard psychophysics method of limits analysis. The overall means are both near 0.4, however, there is quite a bit of user-level variability.

the mean, as well as estimated density of participant-level responses.

While these results suggest that the transformation is useful and that complete transformation is not necessary, we can get more precise bounds on the range of acceptable transformation weights using a linear model that can incorporate starting points other than 0 and 1, and at the same time allow for user-specific variability.

In order to account for user-level variability, we fit a random effects model for the adjusted weight value as a function of starting weight and trial type, with a random intercept for each participant.

Let W_{ij} denote the final adjustment to weight by participant i, $1 \leq i \leq 125$, on trial j, $1 \leq j \leq n_i$. We model the final weight W_{ij} as a function of the correction type $T(i,j)$ (where $T(i,j) \in \{X, Y\}$), and starting weight X_{ij}, with a random intercept for participant to account for subject-specific ability:

$$\begin{aligned} W_{ij} &= \alpha_{T(i,j)} + \beta X_{ij} + \gamma_{i,T(i,j)} + \epsilon_{ij} \\ \gamma_{iX} &\overset{\text{i.i.d.}}{\sim} N(0, \eta_X^2),\ \gamma_{iY} \overset{\text{i.i.d.}}{\sim} N(0, \eta_Y^2), \\ \epsilon_{ij} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ and } \mathrm{Cov}(\gamma, \epsilon) = 0 \end{aligned} \tag{2.11}$$

$\alpha_{T(i,j)}$ is either α_X or α_Y, describing the lower threshold of the acceptable range for each of the types of correction, while $\alpha_X + \beta$ and $\alpha_Y + \beta$ describe the upper thresholds for the respective correction.

We can therefore interpret β as the length of the interval of plausible weights. Additionally, this allows the interpretation of the quantity $(\alpha_* + \beta/2)$ as equivalent to the estimate of the optimal weight based on the psychophysics methodology.

The fitted model parameters are shown in tables 2.1 and 2.2.

Transformation	Threshold	Parameter	Estimate	95% C.I.
X	Lower	α_X	0.097	(0.045, 0.150)
	Upper	$\alpha_X + \beta$	0.625	(0.570, 0.682)
Y	Lower	α_Y	0.143	(0.097, 0.188)
	Upper	$\alpha_Y + \beta$	0.671	(0.626, 0.718)

Table 2.1: Fixed effect estimates of model (2.11) for the boundaries for reasonable weights. In parentheses, 95% parametric bootstrap confidence intervals are given based on model (2.11) (N=1000).

Groups	Correction	Parameter	Estimate	95% C.I.
Participant	X	η_X	0.171	(0.167, 0.247)
Participant	Y	η_Y	0.145	(0.107, 0.179)
Residual		σ	0.304	(0.290, 0.317)

Table 2.2: Overview of random effects for model (2.11), including 95% confidence intervals based on parameteric boostrap results (N=1000).

Table 2.2 gives an overview of the variance estimates. 95% confidence intervals are, based on 1000-fold parametric bootstrap of model 2.11. All variance components are significant and relevant; variability within a single individual's trials is about half the size of variability across participants.

We use parametric bootstrap to generate responses for each correction type and each participant from the model, which we use to both create user-level densities, population-level densities, and bootstrap intervals for model parameters.

The variability of the random effects for each trial type is similar; but the model benefits significantly from allowing separate random effects for individual's variability by correction type (0.1452394 and 0.1705474 for Y and X transformations, respectively, as opposed to 0.3044344 for the overall variability). The interaction between starting weight and trial type was not significant, however, and was thus removed from the model (p-value = 0.9009749).

Figure 2.18 gives an overview of the relationship between starting weights and user-preferred

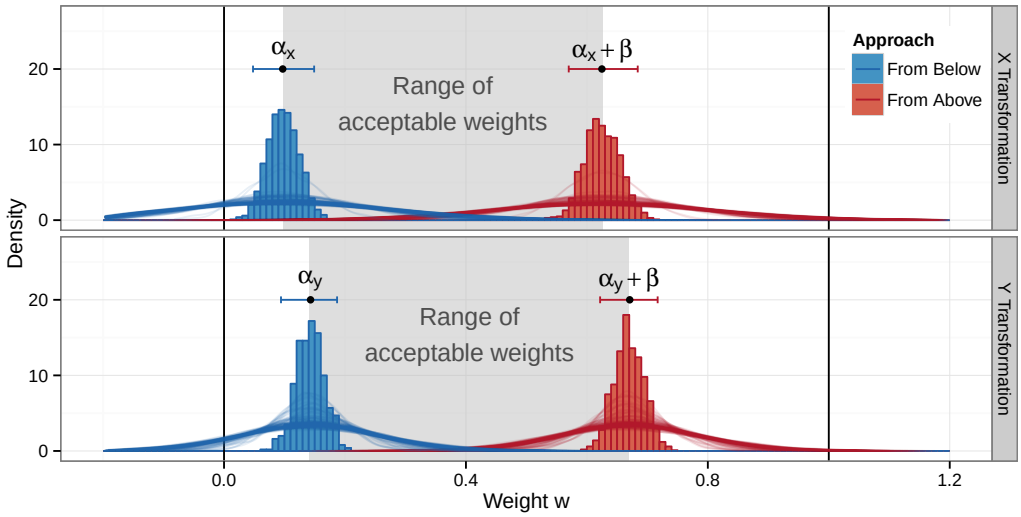


Figure 2.18: Simulation results from the fitted model, facetted by correction type. Fixed effects results are shown as histograms; the red values display the results when starting from an uncorrected plot and are concentrated around $w = 0.1$ for X and $w = 0.14$ for Y; the blue values represent user-chosen weights when starting from a fully corrected plot and are concentrated around $w = 0.63$ for X and $w = 0.67$ for Y. Additionally, 95% bootstrap intervals are shown as horizontal line segments above the histograms; these intervals are for the lower and upper bounds of the "preferred weight interval" tested in the experiment. User-level density curves show the individual variability around fixed effects α_* and $\alpha_* + \beta$.

weight values. Higher starting weights are associated with higher user-submitted values, and lower starting weights are associated with lower user-submitted values.

The ranges of optimal weights are similar under both transformations. Boundaries for the X transformation are slightly lower than boundaries for Y.

Bootstrap simulations for each of the coefficients suggest that the range of optimal w is between 0.098 and 0.625 for x and 0.142 and 0.67 for y, where the lower value is the estimate starting at $w = 0$ and moving up, and the upper value is the estimate starting at $w = 1$ and moving down. This suggests that either correction is preferrable to an uncorrected graph, and that a weighted correction is preferrable to the fully corrected graph, as neither 0 nor 1 is contained in any overall interval. In addition to showing the strength of the correction, this experiment also demonstrates the strength of the illusion itself: a correction appears more uniform than the uncorrected values, even though the corrected values are not uniform and the uncorrected values are completely uniform.

2.4 Application: US Gas Prices

Figure 2.19 shows daily gas prices for a time frame between 1995 to 2014 as published in the Energy Information Administration's historical database of gas prices (EIA, 2014b). This data includes prices for all three grades of gasoline as well as two chemical formulations which are sold in different geographic areas across the United States (for more information, see (EIA, 2014a)).

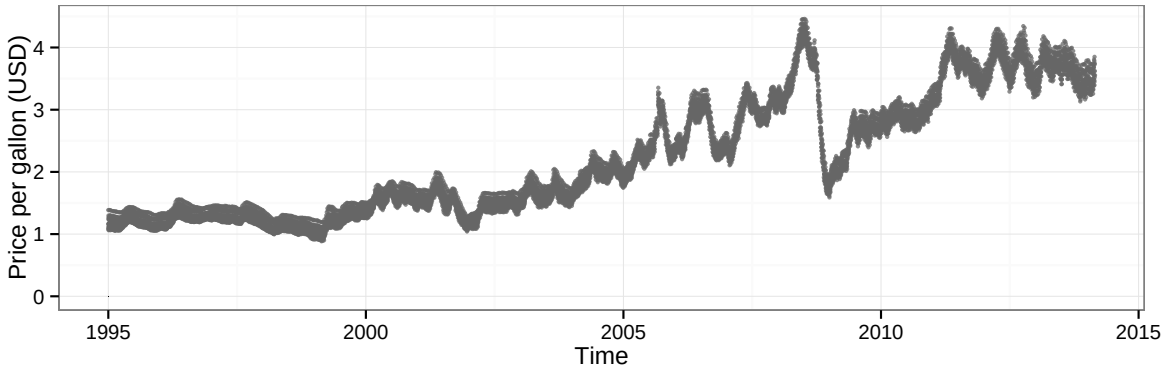


Figure 2.19: US Gas prices from 1995 to 2014. Gas prices steadily increase over the time frame, with some dramatic short-term developments. Peaks and troughs seem to exhibit more variability in daily prices than times of dramatic changes. This is an effect of the sine-illusion, which hides a fairly steady increase in variance in daily gas prices over time.

There is a clear increase in daily gas prices over time as well as several dramatic price changes. These developments mask the steady increase in variance shown in figure 2.20. Instead, we perceive an increase in variability in the frequent ups and downs along the overall trend. In particular, the strong decrease in gas prices at the end of 2008 seems to be associated

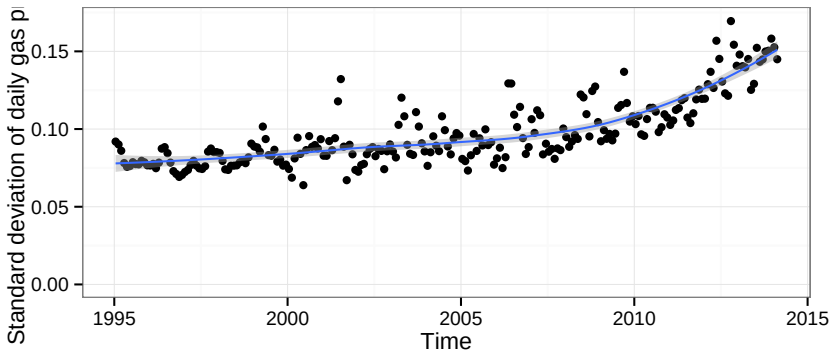


Figure 2.20: Standard deviation of daily gas prices between 1995 and 2014. The doubling of the standard deviation over the time frame is masked in figure 2.19.

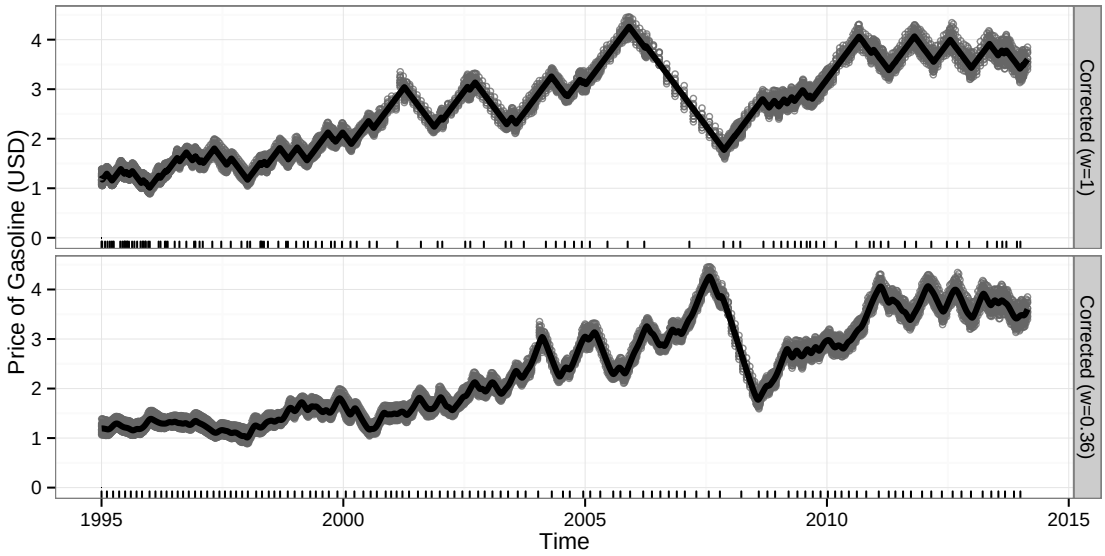


Figure 2.21: Gas price data corrected using the X transformation with $w = 1$ and with $w = 0.36$.

with a low variance. This is an effect of the sine illusion, and the actual variability in Oct 2008 is higher than previous months. In order to better judge variability along the trendline we applied the two different corrections to this data.

For either of the corrections we use a trendline fit based on smoothing splines, which provides the necessary first and second derivatives.

Figure 2.21 shows the results from the X transformation applied to the gas prices. The figure on top is a fully corrected version, while the one below only uses $w = 0.36$, the midpoint of the range of experimentally determined acceptable values, for the transformation. At $w = 1$, the transformation is severe, but it becomes clear that the variance between 1995 and 2000 is lower than it is between 2009 and 2014. When $w = 0.36$, the transformation is much less noticable but yields a near-constant absolute slope of the fitted line.

The minor effect of the weighted transformation on individual x-values contrasts with the effectiveness of the transformation in reducing the illusion; this is best seen in the fitted line, which is distinctly (piecewise) curved in the uncorrected data and appears to be much more piecewise linear in the corrected data, even at the reduced weighted value.

Similar to the X transformation, the Y transformation highlights local fluctuation in the variability of daily gas prices much more than the untransformed data. Figure 2.22 shows Y

transformations for the data. Again, we show a full transformation (top) and a transformation based on the midpoint of the previously determined acceptable region of $w = 0.40$. in the full transformation it is clear that the variance is nearly constant between 1995 and 2000 and then begins to increase with the price of gas. When $w = 0.40$, the transformation is much less noticable, and the resulting y-axis scale is much more similar to the uncorrected data.

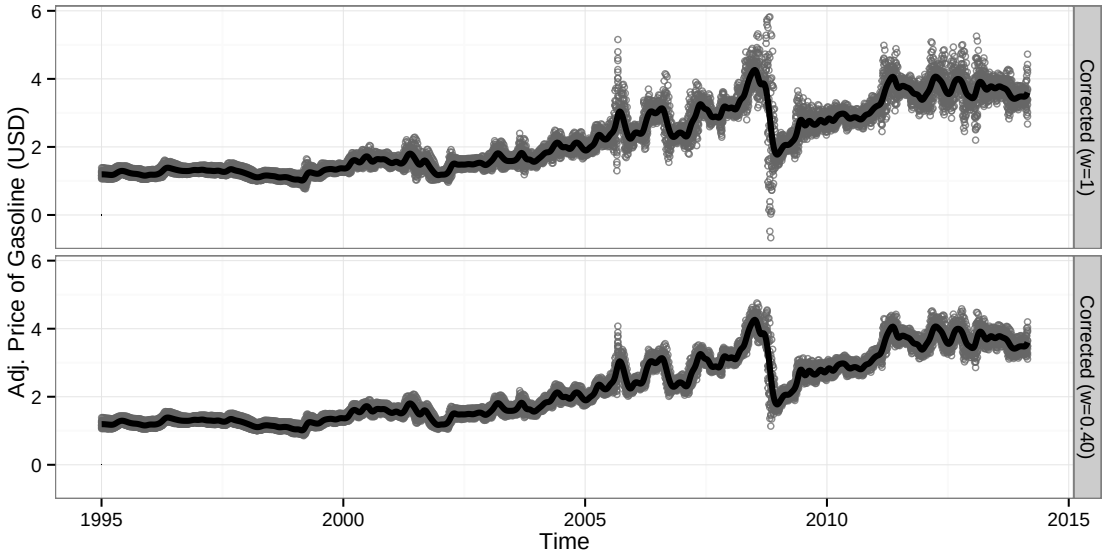


Figure 2.22: Gas price data corrected using the Y transformation with $w = 1$ and with $w = 0.40$.

2.5 Conclusions

The sine illusion is a persistent and powerful illusion that is very difficult to counteract without modifying the visual stimulus directly. While systematically modifying the data is uncommon in the statistical world, this approach is not out of place in the visual arts or architecture; as far back as 400 BC the builders of the Parthenon ensured a straight appearance of the columns from afar, by widening columns at the center, thereby counteracting the effects of the Hering illusion (Howe and Purves, 2005; Hering, 1861). Similarly, painters often exaggerate color hues used in shadows to account for color constancy in the brain. The systematic modifications we suggest here are also comparable to chloropleth maps, which scale a region's area based on some other variable.

We cannot counteract the illusion and represent the data visually without an intervention

that is drastic enough to counteract the three-dimensional context the sine-illusion induces. The proposals in this paper for transformations in x and y provide the means to temporarily correct the data as a diagnostic measure, perhaps using an applet or R package for that purpose. These corrections are significant not only because of their implications for statistical graphics, but because previous attempts to resolve optical illusions using geometry have not met with success (Westheimer, 2008). These corrections are only a first step and could be improved upon; currently, the corrections break down for extreme (secant) values, but multiple iterations of the correction procedure will likely resolve some of these issues (though iteration removes the convenience of a functional form for the transformation). Similarly, the y corrections proposed here extend the line lengths (or for actual data, increase the deviation from the smooth line) – some normalization might make the necessary corrections less noticeable.

Our primary goal is to raise awareness of the illusion and its implications for statistics; the use of plots to guide the modeling process can leave us vulnerable to overlooking changes in the variance due to the illusion. While best practice has been to plot the residuals separately, this removes the context of the data and is not practical before there is a model. In addition, viewer attention spans may be limited if multiple graphs are presented. The proposed transformations require only a nonparametric smooth, maintain the context of the data, and are readily interpretable.

The data for this study was collected with approval from IRB-ID 13-257.

CHAPTER 3. THE CURSE OF THREE DIMENSIONS: WHY YOUR BRAIN IS LYING TO YOU

Intended for submission to IEEE Transactions on Applied Perception

3.1 Introduction

Graphics are one of the most powerful tools researchers have to communicate results to wide audiences. They are easier to understand than tables or verbal descriptions(Larkin and Simon, 1987), easier to remember than words alone (Mayer and Sims, 1994), and provide information that can be perceived and used with minimal additional cognitive load (Zhang and Norman, 1994). Conveying numerical relationships using spatial information makes use of the brain's visual processing ability, freeing working memory to interpret and make connections between graphics and written interpretations. Informative graphics differ from visualizations, sketches, and diagrams in that they present visual summaries of data, using summary functions to map the data graphically, preserving the relationship between two variables using spatial information. That is, unlike visualizations, sketches, and diagrams, spatial relationships presented in information graphics are functions of the data, representing numerical quantities (within the limits of image resolution) accurately.

In fact, the primary principle of informative graphics is that a chart should accurately reflect the data (Tufte, 1991). Tufte argues that some graphics (many of which might be better classified as visualizations) do not accurately reflect the data because they blend artistic rendering with numerical information, sacrificing numerical accuracy for visual appeal. In order to quantify the loss of numerical accuracy, Tufte created the lie factor, which compares the effect size shown in the image to the effect size shown in the data, so that a lie factor

much greater than 1 indicates a picture that over-emphasizes an effect, and a lie factor much smaller than one indicates a picture which minimizes an effect (values between .95 and 1.05 are typically acceptable). While Tufte's lie factor is an effective measurement of the accuracy of the transition from data to graphics, it does not give us any insight about the transition between a chart and the brain, that is, whether the mental representation of the chart is accurate.

Ideally, charts not only represent the data accurately, but also allow readers to draw accurate conclusions. Generally, the human visual system is quite good at accurately interpreting charts (Cleveland and McGill, 1984; Kosara and Ziemkiewicz, 2010), but we need to be aware of contextual misperceptions that lead us to the wrong conclusion. While there are relatively few examples of the effect of optical illusions and other misperceptions on information graphics, Amer (2005) and Poulton (1985) have documented the effect of the Poggendorff illusion on line graphs in different contexts. In this paper, we examine a situation in which low-level human perceptual processes interfere with making accurate judgements from displays and suggest an experimental methodology for estimating the psychological "lie factor" of a chart due to a specific conceptual misperception: the sine illusion.

3.1.1 The Curse of Three Dimensions

The human visual system is largely optimized for perception of three dimensions. Biologically, binocular vision ensures that we have the necessary information to construct a functional mental representation of the three-dimensional world, but even in the absence of binocular information the brain uses numerous heuristics to parse otherwise ambiguous two-dimensional retinal images into meaningful three-dimensional information. Predictably, however, these heuristics are not without drawbacks; the same two-dimensional neural representation might correspond to multiple three-dimensional objects, as in the well-known Necker Cube (Gregory (1968); shown in figure 3.1). Additionally, the same three-dimensional object often has infinitely many two-dimensional representations, for instance, when viewed from different angles. Many optical illusions occur due to the transition from a two to three dimensions (or from three dimensions to two dimensions)(Gregory, 1968). The necker cube has a single two-dimensional representation corresponding to two three-dimensional objects which are both equally salient.

As a result, the brain does not prefer one interpretation over the other and instead continuously switches between interpretations. Impossible objects, such as the Penrose triangle (Penrose and Penrose, 1958), are two-dimensional images of objects that appear to be impossible in three dimensions (sometimes, these objects can be represented in three dimensions, but only appear to be impossible from a certain angle). Impossible objects produce a conflict between the brain's three-dimensional representation of a two-dimensional figure and the brain's experience with the physical world. This conflict between the constraints of physical reality and a depicted image can be quite compelling and is an important component in the work of artists like M.C. Escher (Seckel, 2007).

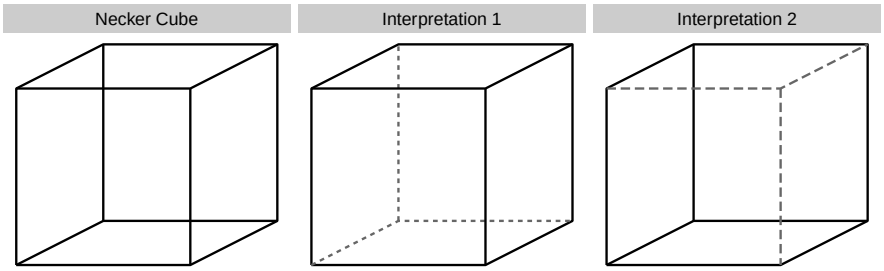


Figure 3.1: The Necker Cube is a so-called "ambiguous object" because two different transparent objects produce the same retinal image (and thus the same perceptual experience). Commonly, the image seems to transition instantaneously from one possible mental representation to the other.

In non-illusory contexts, experience with the real world informs the choice between multiple possibilities of rendering a three-dimensional object from the same two-dimensional representation. This indicates that processing occurs "top down" in that our previous experience influences our current perceptions. Without this top-down influence, the brain would not be able to map a two-dimensional image back to a three-dimensional object. One of the most well studied examples of the influence of top-down processing is the Müller-Lyer illusion, shown in figure 3.2.

In the Müller-Lyer illusion, two vertical line segments are shown with arrows extending from both ends; one segment forms an acute angle with the arrows, the other segment forms obtuse angles with the arrows. The line segment adorned with arrows that form an acute angle appears to be shorter than the linen segment which forms an obtuse angle with the arrow

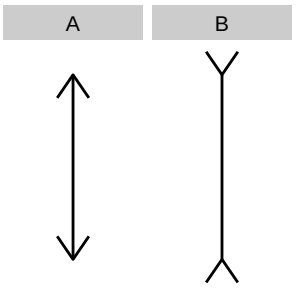


Figure 3.2: The Müller-Lyer illusion. The central segment of figure A is perceived as shorter than the central segment of figure B, even though the two are actually the same length.

segments.

One explanation for the Müller-Lyer illusion (Gregory, 1968) is that the brain interprets the ambiguous lines as a common three-dimensional object common to everyday experience: corners of a room. Figure 3.2A occurs when viewing the outside corner of a rectangular prism, figure 3.2B occurs when viewing the prism from the inside. In regions which do not commonly have rectangular buildings, the illusion is significantly less pervasive (Ahluwalia, 1978). Figure 3.3 provides one possible context that would lead to the Müller-Lyer effect. This real-world experience carries with it an inferred perspective - when the arrows point inward, the object is typically closer than when the arrows point outward, which causes the brain to interpret the outward-pointing figure as larger when the retinal size of the two objects is identical. The perspective cues which contribute to the Müller-Lyer illusion allow for an accurate neural representation of the object in context; when misapplied to two-dimensional stimuli, these cues are responsible for the illusion's effect. This inferred "depth cue" (Gregory, 1968) is reasonably consistent across individuals, suggesting that the phenomenon has a neurological basis.

A similar effect can also be found in the Necker Cube - whichever face appears to be furthest away also seems larger, even though any two parallel faces are equally sized in the image. This approach has proved to be very advantageous for real world scenarios (Gregory, 1968), as pictures of real objects are seldom ambiguous. This strategy also allows for high performance with limited neural bandwidth.

Figure 3.3: Real-world context that gives rise to the Müller-Lyer illusion. The highlighted areas correspond to the parts of the Muller Lyer illusion, and while the two arrows are obviously the same size in the real world, the black arrow takes up much more visual space than the white arrow.

3.1.2 Three Dimensional Context of the Sine Illusion

While the classic Müller-Lyer illusion is seldom a factor in information charts, there are other illusions caused by the interpretation of a two-dimensional stimulus in the context of three-dimensional objects, leading to a distortion in the mental representation of the original stimulus. The sine illusion (also known as the line width illusion: VanderPlas and Hofmann (2014); Day and Stecher (1991)) is one example of this phenomenon which occurs frequently in information graphics.

Figure 3.4 shows the sine illusion in its original form (Day and Stecher, 1991) as straight vertical lines of the same length with a sinusoidal mean function. In this illusion, the vertical lines in the center of the figure appear much shorter than the vertical lines at the peak and trough of the sine curve. The illusion still persists when the image is rotated by 90°. Even when viewers are aware of the illusion's presence, it is close to impossible to overcome mentally.

The problem that the sine illusion presents in information graphics is well documented

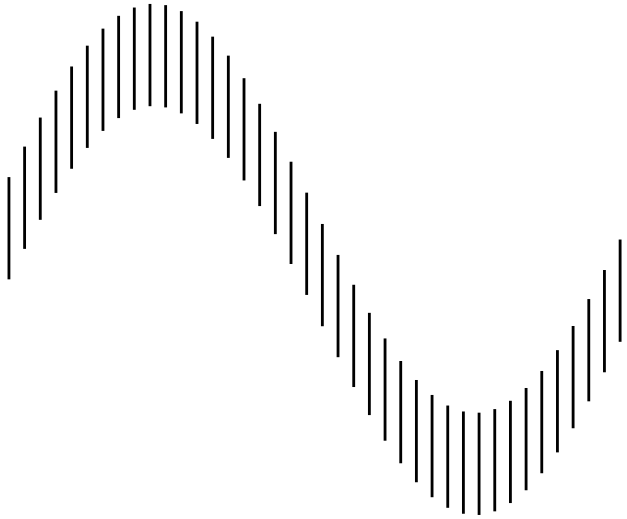


Figure 3.4: The classic sine illusion. Each vertical line has the same length, though the lines at the peak and trough of the curve appear longer.

(Cleveland and McGill, 1984; Robbins, 2005). One such example is the "Balance of Trade" from Playfair's Statistical Atlas (Playfair, 1786), as shown in figure 3.5. The balance of trade in 1765 seems to be approximately the same as the balance of trade in the years immediately preceding that year; this is in fact extremely misleading (using a straightedge along the chart vertically will demonstrate the issue).

In both figures, the illusion appears when the vertical length displayed in the chart does not match the perceived information. Like the Müller-Lyer illusion, the illusion is pervasive and very difficult to "un-see" or mentally correct. The sine illusion, which is also known as the line-width illusion, has also been documented in parallel sets plots (Hofmann and Vendettuoli, 2013) and occurs when there is a nonlinear function with a large change in absolute slope; this change in slope can mask or exaggerate changes in variance. The illusion is also affected by the aspect ratio of the image and the aspect ratio of the chart's coordinate system. An interactive demonstration of the illusion is available at `http://bit.ly/1ldgujL`; manipulating the length of the lines and the amplitude of the underlying sine function also changes the chart's aspect ratio and the perceived strength of the illusion.

The illusion is not dependent on specifically identifying the vertical distance along a line. Figure 3.6(a) shows a scatterplot of data with a trend. A loess smooth is used to estimate the

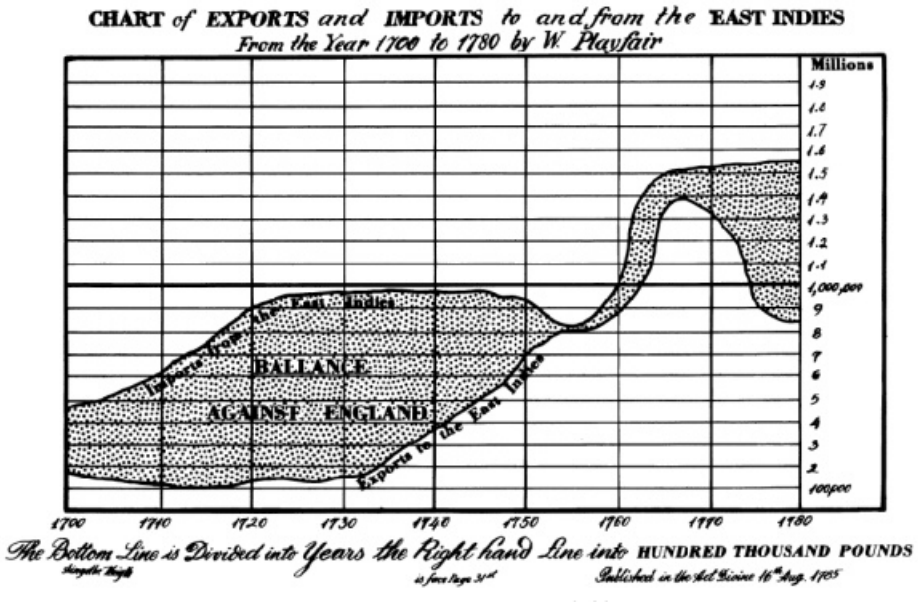
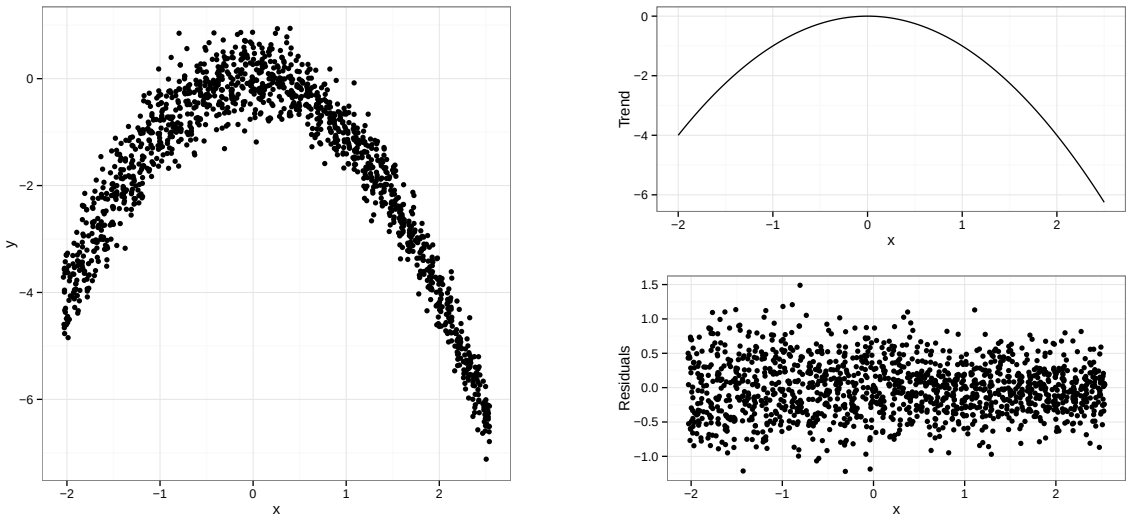


Figure 3.5: Playfair's chart of trade between the East Indies and England, 1700-1780. The trade balance is influenced by the sine illusion: the difference between imports and exports in 1763 does not appear to be the same size as that in 1745, though the vertical distance is approximately the same.

trendline. A visual assessment of variability along this trendline might result in a description such as 'homogeneous variance or slightly increasing variance for negative x, followed by a dramatic decrease in vertical variability for positive x'. Once the residuals are separated from the trendline as shown on the right hand side of the figure, it becomes apparent that this first assessment of conditional variability was not correct, and the steady decrease along the horizontal axis becomes visible.

Cleveland and McGill (1984) determined that comparison of the vertical distance between two curves is often inaccurate, as "the brain wants to judge minimum distance between the curves in different regions, and not vertical distance". While they do not explain a reason for this tendency, introspection does support their explanation: we judge the distance between two curves based on the shortest distance between them, which geometrically is the distance along the line perpendicular to the tangent line of the curve. This comparison holds with scatterplots (such as figure 3.6) because when the points are dense, we examine variability by looking at the upper and lower contours of the data.

Day and Stecher (1991) suggest that the sine illusion is similar in principle to the Müller-



(a) Data

(b) Trend and Residuals

Figure 3.6: Describe the conditional variability of the points along the x axis in (a). Is your description consistent with the residual plot in (b)?

Lyer illusion, attributing it to the perceptual compromise between the vertical extent and the overall dimensions of the figure. The sine illusion is similar to the Müller-Lyer illusion in another way, as well – there are three-dimensional analogues of the two-dimensional image that may influence the perceptual context. One of these contexts is shown in figure 3.7, generated from the same function shown in the two-dimensional analogue, figure 3.4, but with the length projected onto a third dimension. While the images do not match exactly, the similarities are striking. Additionally, the tendency to judge vertical distance using the extant width noted in (Cleveland and McGill, 1985) corresponds to the measurement of depth in the three-dimensional image. The main difference between the first three dimensional image shown in figure 3.7 and the original image is that the lines connecting the top and bottom sections of the curve are slightly angled in the three-dimensional version; this is due to the perspective projection used to create the image and the corresponding angles of rotation chosen such that the entire surface is visible.

As the vanishing point moves further away from the viewer and the 3d projection decreases in strength, the three-dimensional reconstruction of the image converges to figure 3.4. The second image in figure 3.7 shows a weaker 3-dimensional projection that is much closer to figure 3.4, however, the three-dimensional contextual information provided by the shading re-

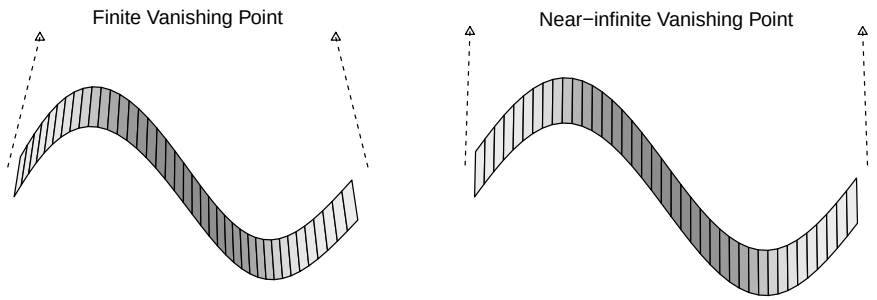


Figure 3.7: Three dimensional context for the sine illusion. The second figure has a vanishing point closer to infinity, and very closely resembles the form of the classic sine illusion.

moves much of the illusion's distortions. This is similar to the Müller-Lyer illusion, as figure 3.3 is not at all ambiguous because the contextual depth information provided by the rest of the surface of the house is sufficient to remove the illusion that the closer corner is in fact larger due to the perspective.

3.1.3 Case Study: Three Dimensions and the Sine Illusion

Further evidence that places the sine illusion firmly into the area of a 3d contextual illusion is given by the non-response to the illusion by individuals with depth-deficient vision. While it is difficult to provide experimental evidence suggesting that the sine illusion is due to depth perception directly, it is possible to examine whether the illusion occurs in people who do not have binocular depth perception. Conditions such as amblyopia (lazy eye) and strabismus (crossed eyes), when not corrected within a critical period during development (Hubel and Wiesel, 1970), can result in weakened or absent depth perception (Henson and Williams, 1980; Parker, 2007; Holopigian et al., 1986). In many cases, use of partial patching and early surgery can correct these problems before the critical period lapses, but before protocols were well established, this was not always completely successful.

We examined the effect of the sine illusion on DW, who has minimal depth perception due to strabismic amblyopia. DW was diagnosed as a young child, and prescribed complete patching to strengthen her intially non-dominant eye. As a result of the patching, DW developed near-independent control over both eyes (doctors now recommend partial patching as a result of this problem). She has 20/20 vision, and can wear glasses to correct the strabismus, but generally does not because they are not necessary for her to see well. DW is right-eye dominant in most

contexts, but is left-eye dominant for driving, and can switch which eye is in focus at will.

We asked DW to view a subset of the sine illusion stimuli used in the experiment described in the next section, as well as the Müller-Lyer illusion, identifying the illusions as having lines that appeared the same length or different lengths (the stimuli are included in the appendix). DW identified both uncorrected sine illusion graphs as having lines of the same length, indicating that she did not appear susceptible to the sine illusion. In addition, DW identified the partially corrected images as having the same line length, indicating that the corrected image would produce similar conclusions as the uncorrected image (in this, she was not different from those with normal binocular vision). In fact, DW only identified the fully corrected $y = \exp(x)$ image as having lines of different length.

In addition to DW's resistance to the effects of the sine illusion, she also was not fooled by the Müller-Lyer illusion, instantaneously identifying the lines as the same length. This suggests that these two illusions are related to the presence of binocular depth perception, perhaps mitigated by experience.

One difference between the sine illusion and the Müller-Lyer illusion that may influence the tendency to see a three-dimensional "ribbon" instead of the two-dimensional sine curve is that the vertical lines in the sine illusion are ambiguously oriented - there is an entire plane of possible three-dimensional reconstructions for each line, and each possible rotation leads to a line of different length. It is this facet of the image that we believe partially contributes to the ambiguity of the image, though it is not a necessary feature for the illusion to persist, as the illusion also can be found in scatterplots and in "ribbon plots" such as figure 3.5.

3.2 Measuring the Psychological Lie Factor Experimentally

3.2.1 The Psychological Lie Factor

The psychological mechanisms which force three-dimensional context onto two-dimensional stimuli are useful adaptations to a three-dimensional world (Gregory, 1968), but they do have disadvantages when applied to abstract two-dimensional stimuli, such as information charts. In order to assess the distortions due to the illusion, we need to quantify this distortion. For

comparison, we will work from Tufte's lie factor (Tufte, 1991, pg 57), which compares the size of an effect in the data with the size shown in a graphic, and is defined in equation 3.1.

$$\text{Lie Factor} = (\text{size of effect shown in chart})/(\text{size of effect shown in data}) \tag{3.1}$$

We will similarly define the psychological lie factor for this illusion as shown in equation 3.2.

$$\text{Psychological Lie Factor} = (\text{size of effect perceived})/(\text{size of the effect shown in the chart}) \tag{3.2}$$

A correction factor which focuses on correcting the psychological distortion caused by the sine illusion is detailed in VanderPlas and Hofmann (2014). This correction factor (the y correction in the aforementioned paper) is applied to the line segments and extends these segments vertically at location x according to equation (3.3), where w represents a weight factor to allow variation of the strength of the correction.

$$(1 - w) + w \cdot \left(1/\cos\left(\tan^{-1}(|f'(x)|)\right)\right) \tag{3.3}$$

A value of $w = 0$ indicates that there is no correction, and a value of $w = 1$ indicates that the graph is fully corrected. Extending this approach, we can over-correct or under-correct the graph, to test whether the geometric derivation of the correction is sufficient to remove the illusion. The lie factor can then be determined by varying w, so that the lie factor of the plot selected so that the lines appear to be "even" indicates the level of psychological distortion (since lines which are in fact not even but appear to be even would indicate that some distortion occurred within the brain). An example of the correction's effect and various weight factors can be seen in figures 3.9 and 3.10.

In order to estimate the psychological lie factor that occurs due to this illusion, we assessed the strength of the illusion experimentally.

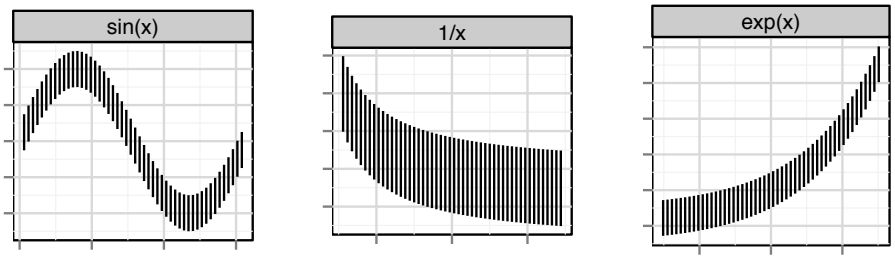


Figure 3.8: Mean functions used during the experiment: $\sin(x)$, $\exp(x)$, and $1/x$. These functions are nonlinear, easily differentiable (for the correction factor), and are similar to trends commonly found in information graphics.

3.2.2 Study Design

The study was designed as a factorial exploration of the factors that contribute to the psychological distortion. We varied the underlying mean function of the stimuli, as well as the strength of the correction described in equation 3.3.

Three underlying mean functions were used: $y = 2\sin(x)$, $y = \exp(x/2)$, and $y = 5/(6x)$; varying these functions allowed us to consider whether the psychological lie factor was influenced by the underlying function. The mean functions, shown in figure 3.8, were chosen from nonlinear functions that occur with relative frequency in statistics which are easily differentiable (due to the correction in equation 3.3), and then refined so that the aspect ratio would be similar for each set of plots (between 0.75 and 0.85). As no x or y units were provided on the graph, these functional modifications served as experimental controls but did not change the information provided to the participants.

In addition to varying the underlying mean function, we also varied the strength of the correction factor (as described by the parameter w in equation 3.3). Experimental stimuli consisting of sets of 6 sub-plots were constructed such that each sub-plot was generated using a different w value between 0 and 1.4. Two of the stimuli used in the experiment are shown in figures 3.9 and 3.11. Figure 3.10 shows the amount of line correction used in each of the sub-plots in figure 3.9, and the (ordered) w values and corresponding lie factors are shown in table 3.1 (row 4).

For each of the stimuli, participants were asked to answer the question: “In which graph is

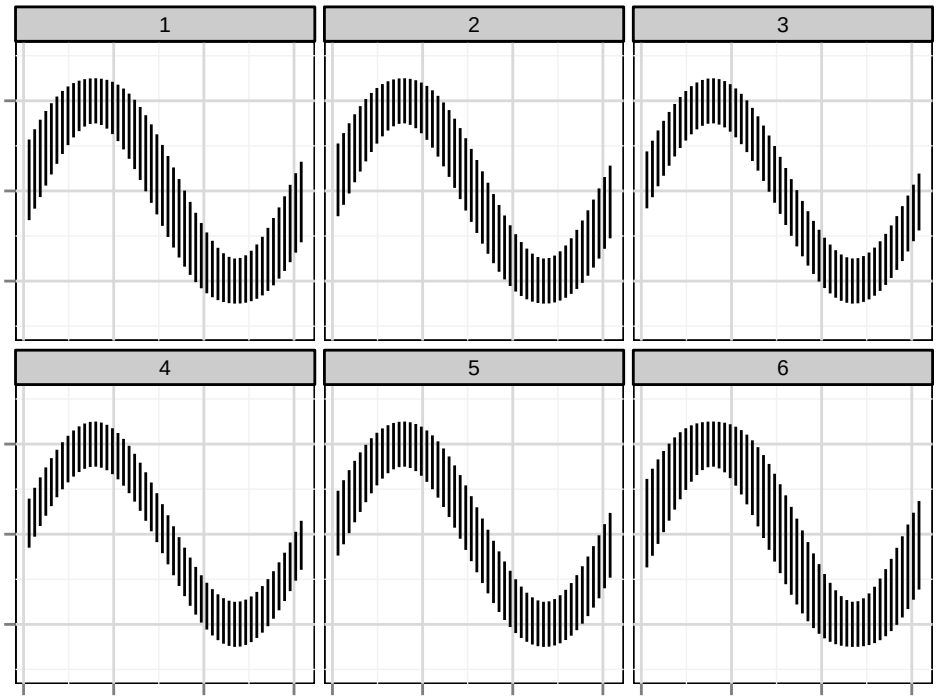


Figure 3.9: One of the charts presented to participants through Amazon Mechanical Turk. Figure 3.10 shows the actual differences in line lengths. This chart corresponds to set #4 in table 3.1. The plots are shown in random order, however; plot #1 corresponds to $w = 0.9$, plot #2 to $w = 0.7$, plot #3 to $w = 0.3$, plot #4 to $w = 0.1$, plot #5 to $w = 0.5$, and plot #6 to $w = 1.1$.

the size of the curve most consistent?". The phrasing 'size of the curve' was chosen deliberately so as not to bias participants to explicitly measure line lengths.

Figure 3.11 shows another set of these stimuli using a different underlying mean function with the same underlying weight values. As the slope of the mean function has changed, the illusion does appear to be slightly less misleading. Table 3.3 (plot 4) shows the weight values (ordered from least to greatest) and corresponding lie factors for this plot; they are lower than the lie factors corresponding to the sine illusion plots shown in figure 3.9 even though the weight values are the same. The illusion is still present even though the mean function has changed;

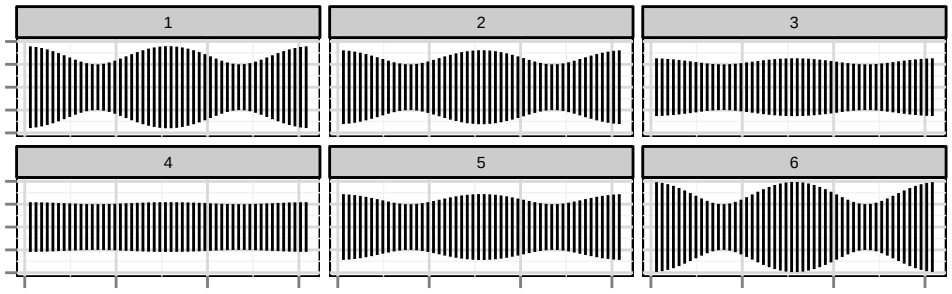


Figure 3.10: De-trended line lengths for figure 3.9, demonstrating the distortion present due to the correction factor in each sub-plot. Comparing the distortion in the chosen sub-plot to the undistorted data produces an estimate of the psychological lie factor.

our goal is to determine whether the psychological distortion is similar despite the difference in the underlying function.

3.2.3 Calculating the Psychological Lie Factor

In order to quantify the psychological lie factor D for each sub plot $k = 1, ..., 6$, we took the ratio of the maximum line length to the minimum line length shown in the plot. As participants were to choose the plot with lines that had the most even length, a large value of D indicates significant distortion.

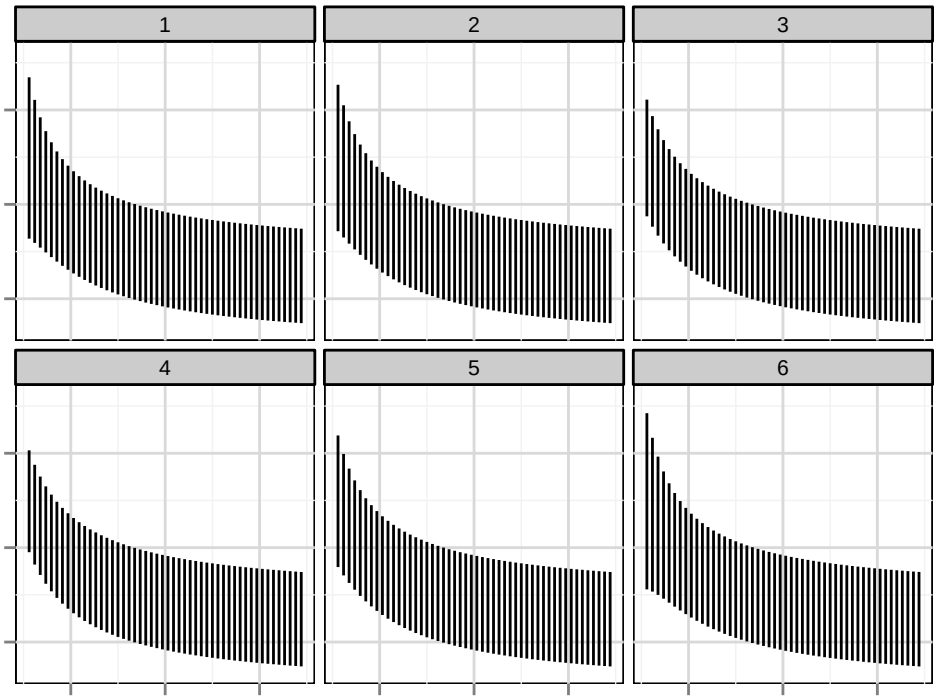


Figure 3.11: Another chart presented to participants through Amazon Mechanical Turk. This chart corresponds to set #4 in table 3.3. The plots are shown in random order, however; plot #1 corresponds to $w = 0.9$, plot #2 to $w = 0.7$, plot #3 to $w = 0.3$, plot #4 to $w = 0.1$, plot #5 to $w = 0.5$, and plot #6 to $w = 1.1$.

$$D_k = (\text{maximum line length in sub-plot } k)/(\text{minimum line length in sub-plot } k) \tag{3.4}$$

The definition of D in equation 3.4 does deviate somewhat from our extension of Tufte's definition of the lie factor described in equation 3.2, but this is by necessity, as an uncorrected plot would show a difference of 0, which would wreak havoc on any experimental measure of the lie factor, as that 0 would be in the denominator. Our modification preserves the interpretation of Tufte's lie factor, while adapting the computation for use in this experimental setting.

From an internal pilot study, we expected values around $w = 0.8$ to be sufficient to break the illusion, but did not know whether this would generalize to those outside of the information

graphics community. In order to pinpoint the weight value necessary to correct the illusion more precisely, we chose twelve sets of 6 weight values each that were used to produce test plots similar to that shown in figure 3.9. These sets of weight values were chosen to allow greater precision estimates closer to $w = 0.8$, while still covering the range of w between 0 and 1.4. The sets of w used are shown in table 3.1, along with corresponding lie factors D_k for stimuli with underlying function $sin(x)$ (other functions used will have different D_k due to the nature of the correction factor). Plots with weight values spread over the full range of w tested were considered "test" plots that could be used for verification purposes, while plots with weight values concentrated near $w = 0.8$ were considered to have higher difficulty (because the sub-plots were very similar). This allowed us to estimate w, and thus D, with higher precision while still exploring the entire parameter space.

Table 3.1: Ordered weight factors and corresponding lie factors for the sine curve stimuli sets, as computed using equation 3.4.

set	diff	Weight (w)						Lie Factor (D) for sin(x) plots					
		sub-plot						sub-plot					
		1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.00	0.20	0.40	0.80	1.25	1.40	1.00	1.18	1.35	1.71	2.11	2.24
2	0	0.00	0.15	0.35	0.80	1.20	1.40	1.00	1.13	1.31	1.71	2.06	2.24
3	1	0.00	0.20	0.40	0.60	0.80	1.00	1.00	1.18	1.35	1.53	1.71	1.88
4	1	0.10	0.30	0.50	0.70	0.90	1.10	1.09	1.27	1.44	1.62	1.80	1.97
5	2	0.05	0.30	0.50	0.65	0.80	1.00	1.04	1.27	1.44	1.57	1.71	1.88
6	2	0.10	0.30	0.55	0.70	0.85	1.00	1.09	1.27	1.49	1.62	1.75	1.88
7	3	0.40	0.60	0.70	0.80	0.90	1.05	1.35	1.53	1.62	1.71	1.80	1.93
8	3	0.35	0.65	0.75	0.85	0.95	1.05	1.31	1.57	1.66	1.75	1.84	1.93
9	4	0.35	0.50	0.60	0.70	0.80	0.95	1.31	1.44	1.53	1.62	1.71	1.84
10	4	0.40	0.55	0.65	0.75	0.85	1.00	1.35	1.49	1.57	1.66	1.75	1.88
11	5	0.50	0.65	0.75	0.80	0.90	1.00	1.44	1.57	1.66	1.71	1.80	1.88
12	5	0.50	0.60	0.70	0.75	0.85	1.00	1.44	1.53	1.62	1.66	1.75	1.88

Each participant was presented with eleven sets of graphs (each "set" consisting of 6 separate plots), consisting of one "easy" test set, five stimuli sets of difficulty level 1 through 5 with the sine curve as the underlying function, and another five graph sets (also of difficulty levels 1 to 5) with either the exponential or the inverse curve as the underlying function. After the easy introductory chart, which was presented first, the order of the plots was randomized across

Table 3.2: Ordered weight factors and corresponding lie factors for the exponential curve stimuli sets, as computed using equation 3.4.

set	diff	Weight (w)						Lie Factor (D) for exp(x) plots					
		sub-plot						sub-plot					
		1	2	3	4	5	6	1	2	3	4	5	6
3	1	0.00	0.20	0.40	0.60	0.80	1.00	1.00	1.21	1.42	1.63	1.84	2.05
4	1	0.10	0.30	0.50	0.70	0.90	1.10	1.11	1.32	1.53	1.74	1.95	2.16
5	2	0.05	0.30	0.50	0.65	0.80	1.00	1.05	1.32	1.53	1.69	1.84	2.05
6	2	0.10	0.30	0.55	0.70	0.85	1.00	1.11	1.32	1.58	1.74	1.90	2.05
7	3	0.40	0.60	0.70	0.80	0.90	1.05	1.42	1.63	1.74	1.84	1.95	2.10
8	3	0.35	0.65	0.75	0.85	0.95	1.05	1.37	1.69	1.79	1.90	2.00	2.10
9	4	0.35	0.50	0.60	0.70	0.80	0.95	1.37	1.53	1.63	1.74	1.84	2.00
10	4	0.40	0.55	0.65	0.75	0.85	1.00	1.42	1.58	1.69	1.79	1.90	2.05
11	5	0.50	0.65	0.75	0.80	0.90	1.00	1.53	1.69	1.79	1.84	1.95	2.05
12	5	0.50	0.60	0.70	0.75	0.85	1.00	1.53	1.63	1.74	1.79	1.90	2.05

difficulty level as well as function type. The test chart consisted of a set of six sine curves with a very low level difficulty level, and was used as an introduction to the testing procedure. Participants were asked to select a single plot out of the 6 plots presented as having lines which were the most "even".

3.2.4 Data Collection

Participants for the study were recruited through the Amazon Mechanical Turk web service, which connects workers with tasks that are not easily automated for a small fee. In exchange for completing at least 11 trials, participants were paid $1. Given the anonymity of web-based data collection, we informed participants that a unique IP address was required to participate in the experiment; responses from duplicate IP addresses with different Turk IDs were grouped and only the first response was paid. This procedure was used to lower the probability of a single user completing the task multiple times, in order to ensure that we could accurately estimate variation among individuals.

1598 responses from 115 users at 110 unique IP addresses were collected. We removed data from participants who did not complete at least 10 trials (allowing for one trial to be skipped or otherwise not completed), and of the collected responses, 30 trials from 4 participants were removed.

Table 3.3: Ordered weight factors and corresponding lie factors for the inverse curve stimuli sets, as computed using equation 3.4.

set	diff	Weight (w)						Lie Factor (D) for $1/x$ plots					
		sub-plot						sub-plot					
		1	2	3	4	5	6	1	2	3	4	5	6
3	1	0.00	0.20	0.40	0.60	0.80	1.00	1.00	1.14	1.28	1.43	1.57	1.71
4	1	0.10	0.30	0.50	0.70	0.90	1.10	1.07	1.21	1.36	1.50	1.64	1.78
5	2	0.05	0.30	0.50	0.65	0.80	1.00	1.04	1.21	1.36	1.46	1.57	1.71
6	2	0.10	0.30	0.55	0.70	0.85	1.00	1.07	1.21	1.39	1.50	1.60	1.71
7	3	0.40	0.60	0.70	0.80	0.90	1.05	1.28	1.43	1.50	1.57	1.64	1.75
8	3	0.35	0.65	0.75	0.85	0.95	1.05	1.25	1.46	1.53	1.60	1.67	1.75
9	4	0.35	0.50	0.60	0.70	0.80	0.95	1.25	1.36	1.43	1.50	1.57	1.67
10	4	0.40	0.55	0.65	0.75	0.85	1.00	1.28	1.39	1.46	1.53	1.60	1.71
11	5	0.50	0.65	0.75	0.80	0.90	1.00	1.36	1.46	1.53	1.57	1.64	1.71
12	5	0.50	0.60	0.70	0.75	0.85	1.00	1.36	1.43	1.50	1.53	1.60	1.71

In addition to the requirement that participants complete at least 10 trials, the participant also was required to complete at least 4 trials of a specific underlying function for those trials to be included. This condition ensured that for any specific function, there were enough trials to estimate an initial effect; an additional 26 trials were excluded based on this condition.

There were 106 users who completed at least 10 trials each (providing enough data that we could accurately fit individual-level parameters), and those users completed 1542 trials which were used for this analysis. Though participants were asked to participate in 11 total trials, some participants continued to provide feedback beyond the eleven trials required to receive payment through Amazon. For any subsequent responses we randomly selected one of the 32 possible stimuli. This approach allowed us to collect some data in which a single participant provided responses to all three underlying functions. We did not exclude data based on user responses to avoid biasing our conclusions; the possibility of significant variability between individuals in the preference for a specific w value was too large to filter out even those who chose plots corresponding to $w = 1.4$.

Due to the experimental design, all participants completed trials with underlying function $y = \sin(x)$, for a total of 815 trials. As each participant who completed only the required 11 trials saw charts with either $y = 1/x$ or $y = \exp(x)$ as the underlying function, each of these

functions had fewer trials; 316 and 411, respectively.

3.2.5 Analysis

Psychological "Lie Factor"

As the strength of the correction varies across the horizontal range of the curve, we quantify the psychological distortion as the ratio of the maximum line length to the minimum line length for each sub-plot k: $D_k = l_{\max}/l_{\min}$. For a given set of 6 sub-plots, j, D_{jk} would denote the sub-plot distortion factor. Let D_{ijk} denote the distortion factor corresponding to participant i's choice of sub-plot k in stimulus set j during a single trial. In this experiment, there are 32 stimuli sets (2 test sets + 10 sets for each of 3 underlying functions), so $1 \leq j \leq 32$.

The correction for the sine illusion by default extends the line segments, so that if the initial line segments were all of length 1, the correction will produce corresponding line segments of length greater than or equal to 1. In addition, due to the underlying functions we have chosen, the minimum line length (assuming a starting line length of 1) after correction is approximately 1; this allows us to simplify D_k as

$$D_k = \text{maximum line length in plot } k$$

Without any correction for the sine illusion, this factor is, like Tufte's lie-factor, equal to one. Values above one indicate that at least in some areas of the curve line segments are extended.

We compute this quantity for each sub-plot in each stimulus presented to the participant. The participant's choice therefore provides us with an estimate of what value of D constitutes the most consistent line length (out of the set shown). As each set of 6 plots is not guaranteed to contain a plot with $w = 0$, corresponding to constant length, choosing a plot with $D = 1.4$ indicates more distortion if there is a sub-plot with $w = 0$ ($D = 1$) present than if least distorted sub-plot present has $w = 0.4$ ($D = 1.2$ for plots of $y = \sin(x)$). Correcting for this bias, the set of $\{D_{..k}\} = \{D_{..1}, ..., D_{..6}\}$ that is available to choose from produces an estimate of the overall psychological 'lie factor' as

$$P_{ij} = D_{ijk} / \min_{1 \leq k \leq 6} D_{ijk} \tag{3.5}$$

for each plot and each participant. This normalization does conservatively bias the results, effectively shrinking the effect size we observe, but without the normalization we would be biased in favor of finding a significant result. Furthermore, while we could normalize relative to $\max_{1\leq k\leq 6} D_{ijk}$, most stimuli sets contain $D > 1.85$ (for $y = \sin(x)$), where more than half of the stimuli sets do not contain $D \approx 1$. This is due to the range of w values chosen for the stimuli sets, as we had pilot data suggesting that $w = 0.8$ was a commonly preferred weight value.

That is, P is the ratio of the lie factor of the chosen plot to the smallest lie factor available in the set of available plots. When the set of available plots contains an uncorrected plot, $P = D$; when an uncorrected plot is not presented, $P < D$, since each $D_j >= 1$, $j = 1, ..., 6$. This transformation is a conservative approach to estimating the lie factor, but allows us to show a variety of scaled transformations and estimate the effect with more granularity around $w = 0.8$.

By considering each participant's answers for the plot with the most consistent line length, we can obtain an estimate of the psychological distortion from the sine illusion on an individual level. Estimating distortion factors for each participant facilitates comparison of these estimated values to determine whether the illusion is a product of an individual's perceptual experience or whether there is a possible underlying perceptual heuristic for the sine illusion common across the majority of participants. If the illusion is a learned misperception rather than an underlying perceptual "bug", we would expect there to be considerable variability in the estimated individual lie factor P_i for each unique participant i, $1 \leq i \leq 123$, as it is likely that personal experience varies more widely than perceptual heuristics and their underlying neural architecture.

Each set of w values as defined in table 3.1 corresponds to a value of P as defined in equation 3.5. We test for only a set of discrete values of w, which is reflected directly in the number of different values of P we can observe. This approach allows us to use a finite set of stimuli for testing, so that we can explicitly control the range of w displayed in each set of plots. To mathematically model a continuous quantity (the real domain of possible P values) using discrete data, we employ a Bayesian approach to model an overall psychological lie factor

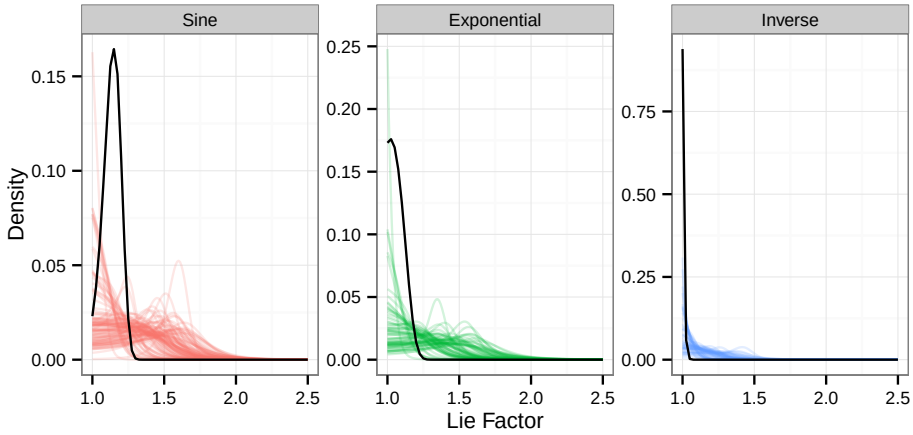


Figure 3.12: Estimated densities for θ_i, shown in color, with the estimated overall density for θ shown in black. Individuals have extremely similar posterior distribution of θ_i, and even different functions have similar $\hat{\theta}$, suggesting a common underlying mental distortion.

θ and individual participant lie factors θ_i.

Plots used in the experiment have factors P ranging between 1 and 2.5, so we can use a truncated normal data model for participant i viewing plot j, with $P = p_{ij} \sim N(\theta_i, \sigma)$ and independent flat priors $\pi(\theta) = 0.4$ and $\pi(\sigma) = 2.5$ for $\sigma \in [.1, .5]$. These prior distributions $\pi(.)$ represent our expectations of the values of θ and σ before the experiment; assigning them constant values indicates that we had little useable knowledge about the joint or marginal distributions of θ and σ before the experiment was conducted. Using Bayesian estimation, we can then obtain posterior distributions for θ_i and θ, the individual and overall mean lie factors. We are not particularly interested in the actual values of σ, but the additional parameter is a useful tool to better estimate possible values for θ.

3.2.6 Results

The posterior density of θ for each function is shown in figure 3.12, along with separate posterior densities for each individual θ_i. θ is reasonably similar for all three functions, suggesting that while function type may moderate the size of the effect, the illusion occurs regardless of function type. Individual curves have different variability due to the number of trials completed, and are necessarily more spread out as there is less data with which to estimate the individual posterior distributions.

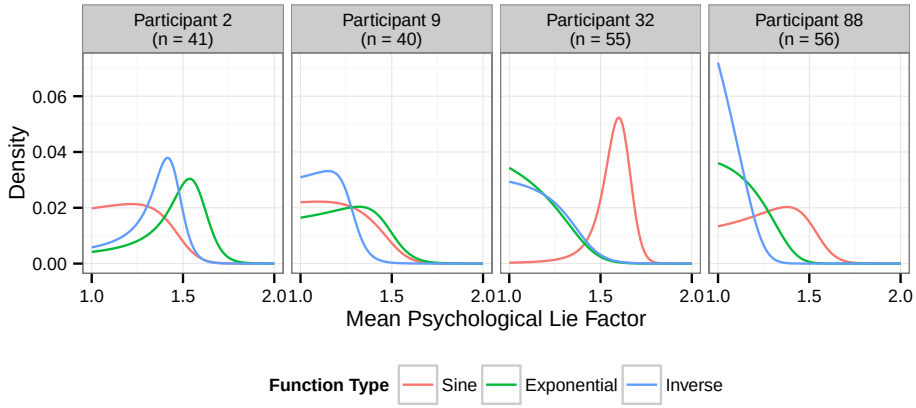


Figure 3.13: Posterior distributions for θ_i for four of the participants who completed at least 6 trials of each of the three function types.

On an individual level, figure 3.13 shows the posterior density for θ_i for four of the participants who completed at least 6 trials in each category. While in many cases, the most probable θ_i is similar across trials, individuals do seem to have been somewhat more affected by the illusion when the underlying function was sinusoidal, though this may reflect a discrepancy in the number of trials rather than a stronger illusion. Alternately, as the illusion depends on variable slope, it is possible that the monotonic exponential and inverse stimuli induced a weaker three-dimensional context. The posterior densities for these individuals appear extreme because they completed more trials (and thus estimates are much more precise); the individuals in question are not necessarily more prone to the illusion than other participants.

In order to appropriately compare intervals for each participant's θ_i (even though participants may have completed different numbers of trials), we simulated 11 new "data points" from our model (thus enforcing a uniform 11 trials per participant for each function type) to get a single new estimate of $\hat{\theta}_i$. For each participant, we generated 1000 of these $\hat{\theta}_i$ and used these simulated values to calculate the 95% credible intervals shown in figure 3.14. These intervals will allow us to consider the variability in θ_i due to participant preference rather than the number of trials a participant completed during the study. Removing this additional variability provides us with the opportunity to consider whether the sine illusion stems from an individual's perceptual experiences or from a lower-level perceptual heuristic.

Posterior predictive intervals for θ_i as shown in figure 3.14 suggest that overall, the θ_i are

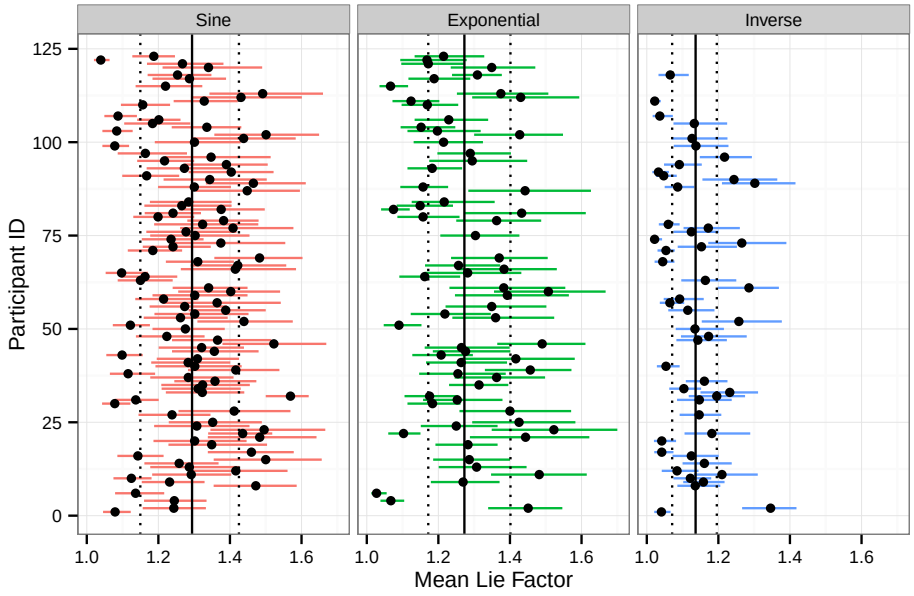


Figure 3.14: 95% posterior predictive intervals for θ_i, calculated for each stimulus type. Vertical lines indicate the median estimate of the overall θ with a 95% credible interval.

similar across individuals. Very few (5 each for $y = \sin(x)$ and $y = \exp(x)$, 16 for $y = 1/x$) of the intervals overlap the region (1, 1.05), which corresponds to an "acceptable" lie factor according to Tufte. This indicates significant distortion for most participants in our experiment, and the marked overlap of the intervals for each participant provides evidence consistent with a common magnitude of distortion. This suggests that there may be some common psychological strategy that is misapplied to the perception of these stimuli.

Comparison of the Preferred Stimuli

Estimates of $\hat{\theta} = E[\theta]$ for each function are 1.31, 1.29, and 1.14 respectively for exponential, inverse, and sine functions, suggesting a similar psychological distortion even for very different functions, though it seems as if the inverse function causes somewhat less distortion, possibly because the correction factor is not as proportionately large. Credible intervals can be found in Table 3.4. As all three of the credible intervals exclude 1.05, there is evidence that a psychological distortion is occurring; that is, there is evidence of a significant psychological lie factor. In addition, the method of adjusting the estimated lie factor we have used here is conservative; it is likely that because most stimuli contain a sub-plot with $w \geq 1$ our estimate

Table 3.4: Credible intervals for the overall θ for exponential, inverse, and sine stimuli.

Function	95% Credible Interval for θ	Median
Sin	(1.0623636, 1.2310909)	1.1372727
Exp	(1.1984091, 1.4643182)	1.3004545
Inv	(1.1735455, 1.3952273)	1.2863636

of θ for the inverse function is low (as the lie factor for fully corrected inverse plots is greater than the corresponding lie factors for the other two functions used in this experiment).

The estimated weight values corresponding to these θ are shown in figure 3.15. In all three cases, the experimentally-corrected plots appear less distorted than the uncorrected plots.

This experiment has demonstrated that the sine illusion results in mis-perception of graphically presented data. In particular, participants tend to see uneven line length when lines are even while missing uneven line length due to the illusion's effect.

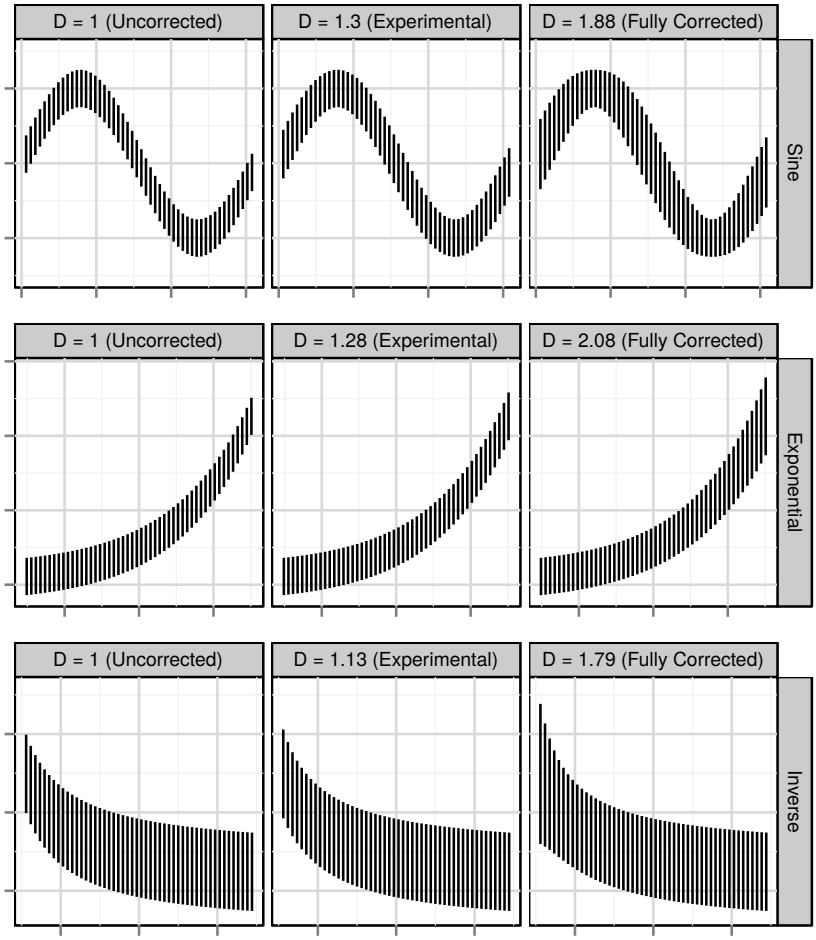


Figure 3.15: Uncorrected, experimentally corrected (according to the median value of θ), and fully corrected stimuli for all three underlying functions used in the experiment. The corrected value shown here is equivalent to the distortion factor D defined as $D = \ell_{max}/\ell_{min}$.

3.3 Conclusions

The sine illusion arises from misapplication of three-dimensional context to a two dimensional stimulus, resulting in nearly unavoidable perceptual distortions. These distortions impact the inferences made from charts which are similar to three-dimensional figures, even when we are not consciously aware that this context exists; the only immunity to the illusion we have found is for those who have never developed binocular depth perception. We have estimated that the illuson produces a distortion of about 135%. This distortion occurs entirely between the retinal image and the mental representation of the object; it is not due to the chart, rather, it is an artifact of our perceptual system.

As Tufte advocated for charts that showed the data without distortion, our goal is to raise awareness of perceptual distortions that occur within the brain itself due to misapplied heuristics. While applying corrections to the data to remove these distortions is somewhat radical, the persistence of the illusion despite awareness of its presence presents a challenge to those seeking to display data visually. In addition, many graph types can induce this illusion (scatterplots, ribbon plots, parallel sets plots), so avoiding a specific type of graph is not an effective solution. The best solution to this problem is to raise awareness: to demonstrate that optical illusions occur within information graphics, and to understand how these illusions arise.

CHAPTER 4. SPATIAL REASONING AND DATA DISPLAYS

Submitted to IEEE Transactions on Visualization and Computer Graphics, January 2015

4.1 Introduction

Data displays provide quick summaries of data, models, and results, but not all displays are equally good, nor is any data display equally useful to all viewers. Graphics utilize higher-bandwidth visual pathways to encode information (Baddeley and Hitch, 1974), allowing viewers to quickly and intuitively relate multiple dimensions of numerical quantities. Well-designed graphics emphasize and present important features of the data while minimizing features of lesser importance, guiding the viewer towards conclusions that are meaningful in context and supported by the data while maximizing the information encoded in working memory. Under this framework, well-designed graphics reduce memory load and make more cognitive resources available for other tasks (such as drawing conclusions from the data), at the cost of depending on certain visuospatial reasoning abilities.

Many theories of graphical learning center around the difference between visual and verbal processing: the dual-coding theory emphasizes the utility of complementary information in both domains, while the visual argument hypothesis emphasizes that graphics are more efficient tools for providing data with spatial, temporal, or other implicit ordering, because the spatial dimension can be represented graphically in a more natural manner (Vekiri, 2002). Both of these theories suggest spatial ability impacts a viewer's use of graphics, because spatial ability either influences cognitive resource allocation or affects the processing of spatial relationships between graphical elements. In addition, previous investigations into graphical learning and spatial ability have found relationships between spatial ability and the ability to read information from

graphs (Lowrie and Diezmann, 2007). However, mathematical ability, not spatial ability, was shown (Shah and Carpenter, 1995) to be associated with accuracy on a simple two-dimensional line graph. Spatial ability becomes more important when more complicated graphical displays are used in comparison tasks: the lower performance of individuals with low spatial ability on tests utilizing diagrams and graphs is attributed (Mayer and Sims, 1994) to the fact that more cognitive resources are required to process the visual stimuli, which leaves fewer resources to make connections and draw conclusions from those stimuli. It is theorized that graphics are a form of "external cognition" (Scaife and Rogers, 1996) that guide, constrain, and facilitate cognitive behavior (Zhang, 1997).

"Lineups" have recently been introduced (Buja et al., 2009; Wickham et al., 2010; Majumder et al., 2013) as a tool to evaluate the statistical significance of a graphical finding. Lineups are also useful in assessing the effectiveness of different graphical displays (Hofmann et al., 2012; Loy et al., 2015). Like their police counterpart, lineups consist of several distractor plots (of randomly generated data) and one target (the data plot).

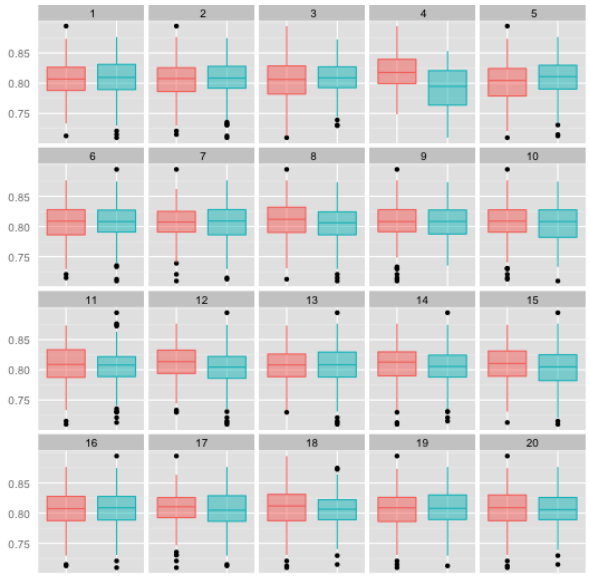


Figure 4.1: Sample lineup of boxplots. Participants are instructed to choose the plot which appears most different from the others. In this lineup, plot 4 is the target plot, because the two groups have a large difference in medians.

Figure 4.1 shows a sample lineup of boxplots; participants are expected to identify the most different among the plots shown. In this example, sub-plot 4 is the target because of the

noticeably different locations of the two boxplot medians.

Lineups provide a quantitative measurement of the effectiveness of a particular plot: if participants consistently identify the target plot rather than the randomly-generated distractors, the plot effectively shows the difference between real data and random noise. This removes much of the subjectivity from user evaluations of display effectiveness, and the procedure is simple enough that it does not generally require participants to be very familiar with data-based graphics. While previous research (Lowrie and Diezmann, 2007; Mayer and Sims, 1994) has examined the link between certain types of graphical perception and spatial skills, it is important to identify any additional visual skills participants utilize to complete the lineup task, as well as better understand demographic characteristics (math education, research experience, age, gender) which may impact performance (Majumder et al., 2014a).

In this paper, we present the results of a study designed to compare lineup performance with visual aptitude and reasoning tests, examining the skills necessary to successfully evaluate lineups. We compare lineup performace to the visual search task (VST), paper folding test, card rotation test, and figure classification test. The VST measures visual search speed(Goldstein et al., 1973), the paper folding and card rotation tests measures spatial manipulation ability, and the figure classification test measures inductive reasoning(Ekstrom et al., 1976); all of these skills are at least peripherally recruited during the lineup task, but some may dominate in predicting performance on the lineup task. We hope to facilitate comparison of the lineup task to known cognitive tests, inform the design of future studies, and better understand the perception of statistical lineups.

In section 4.2 we introduce the tests used in the study and describe how the tests are scored. In section 4.3 we discuss the study results and compare them with scores previously established test, that also take demographic characteristics associated with test scores into account. We discuss multi-collinearity in the study results, and use principal components analysis and linear models to draw some conclusions about the similarity between lineups and aptitude tests. Finally, in section 4.4, we discuss the implications of the study results for the lineup protocol and possible extensions.

4.2 Methods

4.2.1 The Lineup Protocol

The lineup protocol (Hofmann et al., 2012; Wickham et al., 2010; Buja et al., 2009) is a testing framework that allows researchers to quantify the statistical significance of a graphical finding with the same mathematical rigor as conventional hypothesis tests.

In a lineup test, the plot of the data is placed randomly among a set of, generally 19, distractor plots (or *null plots*) that are rendered from data generated by a model, without a signal (a null model). This sheet of charts is then shown to human observers, who are asked to identify the display that is "the most different". If an observer identifies the plot drawn from the actual data, this can be reasonably taken as evidence that the data it shows is different from the data of other plots. Let X be the number of observers (out of n) who identify the data plot from the lineup. Under the null hypothesis that the data plot is not different from the other plots, X has approximately a Binomial distribution (Wickham et al., 2010; Majumder et al., 2013). If k of the observers identify the data plot from the lineup, the probability $P(X \geq k)$ is the p-value of the corresponding visual test. By aggregating responses from different individuals the lineup protocol therefore allows an objective evaluation of a graphical finding.

Additionally, however, we can aggregate the scores from the same individual on several lineups to objectively assess an individual's performance on the lineup task.

For this approach, we derive a score for an individual as follows:

Assume that an observer has evaluated K lineups of size m (consisting of $m-1$ decoy plots and 1 target), and successfully identified the target in n_c of these plots, while missing the target in n_w of them. The score for this individual is then given as:

$$n_c - n_w/(m-1). \tag{4.1}$$

Note that the sum of answers, $n_c + n_w$, is at most K, but may be less, if an observer chooses to not answer one of the lineup tasks or runs out of time.

The scoring scheme as given in (4.1) is chosen so that if participants are guessing, the expected score is 0.

Statistical lineups depend on the ability to search for a signal amid a set of distractors (visual search) and the ability to infer patterns from stimuli (pattern recognition). Depending on the choice of plot shown in the lineup, the task of identifying the most different plot might require additional abilities from participants, e.g. polar coordinates depend on the ability to mentally rotate stimuli (spatial rotation) and mentally manipulate graphs (spatial rotation and manipulation). By breaking the lineup task down into its components, we determine which visuospatial factors most strongly correlate with lineup performance, using carefully chosen cognitive tests to assess these aspects of visuospatial ability.

Demographic factors are known to impact lineup performance: country, education, and age affected score on lineup tests, and all of those factors plus gender had an effect on the amount of time spent on lineups (Majumder et al., 2014a). In addition, lineup performance can be partially explained using statistical distance metrics (Chowdhury et al., 2014), but these metrics do not completely succeed in predicting human performance, in part due to the difficulty of representing human visual ability algorithmically.

One of the most useful features of the lineup protocol is that it allows researchers to conclusively determine which graphics show certain features more conclusively by providing an experimental protocol for comparing graphics based on the accuracy of user conclusions. In addition, lineups provide researchers with a rigorous framework for determining whether a specific graph shows a real, statistically significant effect by comparing a target plot with plots formed using permutations of the same data, providing a randomization test protocol for graphics. As a result, lineups are a useful and innovative tool for evaluating charts; on an individual level, they can also be used to evaluate a specific participant's perceptual reasoning ability in the context of statistical graphics.

4.2.2 Measures of visuospatial ability

Participants are asked to complete several cognitive tests designed to measure spatial and reasoning ability. Tasks are timed such that participants are under pressure to complete; participants are not expected to finish all of the problems in each section. This allows for a better discrimination between scores and prevents score compression at the top of the range.

The **visual searching task** (VST), shown in Figure 4.2, is designed to test a participant's ability to find a target stimulus in a field of distractors, thus making the visual search task similar in concept to lineups. Historically, visual search has been used as a measure of brain damage (Goldstein et al., 1973; DeMita et al., 1981; Moerland et al., 1986); however, similar tasks have been used to measure cognitive performance in a variety of situations, for example under the influence of drugs in (Anderson and Revelle, 1983). The similarity to the lineup protocol as well as the simplicity of the test and its' lack of color justify the slight deviation from forms of visual search tasks typically used in normal populations.

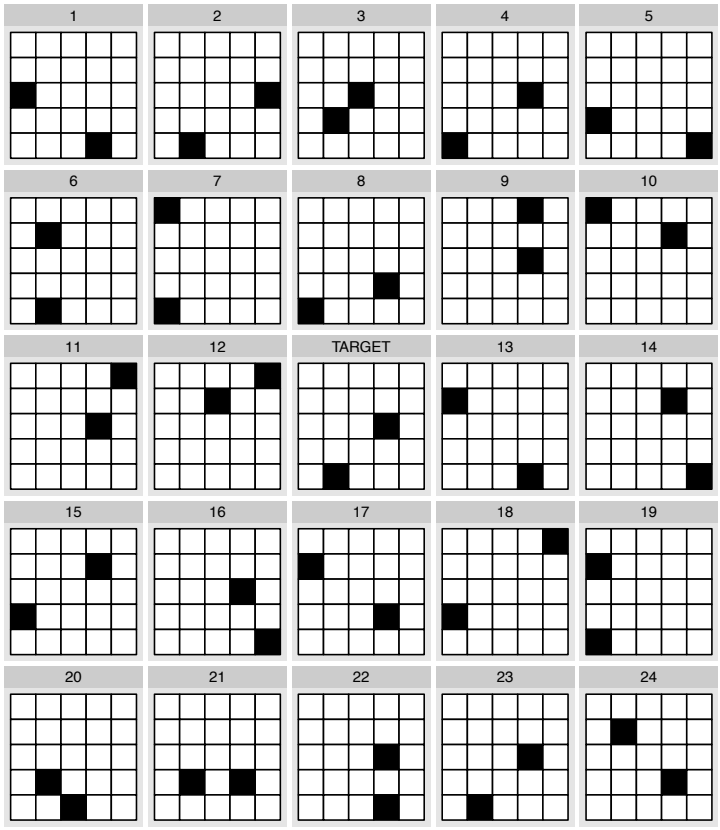
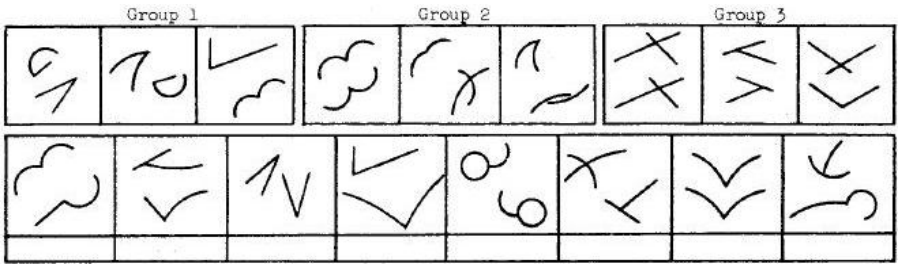


Figure 4.2: Visual Search Task (VST). Participants are instructed to find the plot numbered 1-24 which matches the plot labeled "Target". Participants will complete up to 25 of these tasks in 5 minutes.

The **figure classification task** tests a participant's ability to extrapolate rules from provided figures. This task is associated with inductive reasoning abilities (factor I in Ekstrom et al. (1976)). An example is shown in Figure 4.3a.

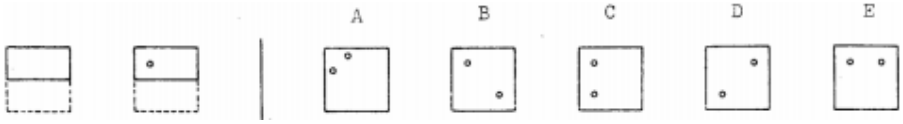
The figure classification test requires the same type of reasoning as the lineups: participants

must determine the rules from the provided classes, and extrapolate from those rules to classify new figures. In lineups, participants must determine the rules based on the panels appearing in the lineup; they must then identify the plot which does not conform. As such, the figure classification test has content validity in relation to lineup performance: it is measuring similar underlying criteria.



(a) Figure Classification Task. Participants classify each figure in the second row as belonging to one of the groups above. Participants complete up to 14 of these tasks (each consisting of 8 figures to classify) in 8 minutes.

(b) Card Rotation Task. Participants mark each figure on the right hand side as either the same as or different than the figure on the left hand side of the dividing line. Participants complete up to 20 of these tasks (each consisting of 8 figures) in 6 minutes.



(c) Paper Folding Task. Participants are instructed to pick the figure matching the sequence of steps shown on the left-hand side. Participants complete up to 20 of these tasks in 6 minutes.

Figure 4.3: Tests of spatial ability from Ekstrom et al. (1976).

The **card rotation test** measures a participant's ability to rotate objects in two dimensions in order to distinguish between left-hand and right-hand versions of the same figure. It tests mental rotation skills, and is classified as a test of spatial orientation in (Ekstrom et al., 1976), though it does require that participants have both mental rotation ability and short-term visual memory. An example is shown in Figure 4.3b. The card rotation test is often used in studies investigating the effect of visual ability on the use of visual aids (Mayer and Sims, 1994) and statistical graphs (Lowrie and Diezmann, 2007) in education.

Two-dimensional comparisons are an important component of lineup performance. In some

lineup situations, these comparisons sometimes involve translation, but in other lineups, rotation is required. Lineups also require visual short-term memory, so the additional factor measured implicitly by this test does not reduce its potential relevance to lineup performance.

The **paper folding test** measures participants' ability to visualize and mentally manipulate figures in three dimensions. A sample question from the test is shown in Figure 4.3c. It is classified as part of the visualization factor in (Ekstrom et al., 1976), which differs from the spatial orientation factor because it requires participants to visualize, manipulate, and transform the figure mentally, which makes it a more complex and demanding task than simple rotation. The paper folding test is associated with the ability to extrapolate symmetry and reflection over multiple steps. Lineups require similar manipulations in two-dimensional space, and also require the ability to perform complex spatial manipulations mentally; for instance, comparing the interquartile range of two boxplots as well as their relative alignment to a similar set of two boxplots in another panel.

Between cognitive tasks, participants were also asked to complete three blocks of 20 lineups each, assembled from previous studies (Hofmann et al., 2012; Majumder et al., 2013). Participants have 5 minutes to complete each block of 20 lineups. Figure 4.1 shows a sample lineup of box plots.

In addition to these tests, participants were asked to complete a questionnaire which includes questions about colorblindness, mathematical background, self-perceived verbal, mathematical, and artistic skills, time spent playing video games, and undergraduate major. These questions are designed to assess different factors which may influence a participant's skill at reading graphs and performing spatial tasks.

4.2.3 Test Scoring

All test results were scored so that random guessing produces an expected value of 0; therefore each question answered correctly contributes to the score by 1, while a wrong answer is scored by $-1/(k-1)$, where k is the total number of possible answers to the question. Thus, for a test consisting of multiple choice questions with k suggested answers with a single correct

answer each, the score is calculated as

$$\#\text{correct answers } - 1/(k-1) \cdot \#\text{wrong answers.} \tag{4.2}$$

This allows us to compare each participant's score in light of how many problems were attempted as well as the number of correct responses. Combining accuracy and speed into a single number does not only make a comparison of test scores easier, this scoring mechanism is also used on many standardized tests, such as the SAT and the battery of psychological tests (Diamond and Evans, 1973; Ekstrom et al., 1976) from which parts of this test are drawn. The advantage of using tests from the Kit of Factor Referenced Cognitive tests (Ekstrom et al., 1976) is that the tests are extremely well studied (including an extensive meta-analysis in (Voyer et al., 1995) of the spatial tests we are using in this study) and comparison data are available from the validation of these factors (Schaie et al., 1998; Hampson, 1990; Mayer and Sims, 1994) and previous versions of the kit (French et al., 1963).

4.3 Results

Results are based on an evaluation of 38 undergraduate students at Iowa State University. 61% of the participants were in STEM fields, the others were distributed relatively evenly between agriculture, business, and the social sciences. Students were evenly distributed by gender, and were between 18 and 24 years of age with only one exception. This is reasonably representative[1] of the university as a whole; in the fall 2014 semester, 26% of students were associated with the college of engineering, 24% were associated with the college of liberal arts and sciences, 15% were associated with the college of human sciences, 7% with the college of design, 13% with the business school, and 15% with the school of agriculture.

4.3.1 Comparison of Spatial Tests with Previously Validated Results

The card rotation, paper folding, and figure classification tests have been validated using different populations, many of which are demographically similar to Iowa State students (naval

[1] http://www.registrar.iastate.edu/sites/default/files/uploads/stats/university/F14summary.pdf

recruits, college students, late high-school students, and 9th grade students). We compare Iowa State students' unscaled scores in Table 4.1, adjusting data from other populations to account for subpopulation structure and test length.

	Card Rotation	Paper Folding	Figure Classification	Visual Search
ISU Students	83.4 (24.1)	12.4 (3.7)	57.0 (23.8)[1]	21.9 (2.3)
Scaled Scores	88.0 (34.8)	13.8 (4.5)	58.7 (14.4)[2]	–
Unscaled Scores	44.0 (24.6)[3]	13.8 (4.5)	M: 120.0 (30.0) F: 114.9 (27.8)	–
Population	approx. 550 male naval recruits	46 college students (1963 version)	suburban 11th & 12th grade students (288-300 males, 317-329 females)	

Table 4.1: Comparison of scores from Iowa State students and scores reported in (Ekstrom et al., 1976). Scaled scores are calculated based on information reported in the manual, scaled to account for differences in the number of questions answered during this experiment. Data shown are from the population most similar to ISU students, out of the data available. The visual search task (Goldstein et al., 1973; DeMita et al., 1981; Moerland et al., 1986) is not part of the Kit of Factor Referenced Cognitive Test data, and thus we do not have comparison data for the form used in this experiment.

Table 4.1 shows mean scores and standard deviation for ISU students and other populations. Values have been adjusted to accommodate for differences in test procedures and sub-population structure; for instance, some data is reported for a single part of a two-part test, or results are reported for each gender separately.

Scaling Scores To calculate "scaled" comparison scores between tests which included different numbers of test sections (as shown in Table 4.1), we scaled the mean in direct proportion to the number of questions (thus, if there were two sections of equivalent size, and the reference score included only one of those sections, we multiplied the reported mean score by two). The variance calculation is a bit more complicated: In the case described in the main text, where the reference section contained half of the questions, the variance is multiplied by two, causing the standard deviation to be multiplied by approximately 1.41.

This scaling gets slightly more complicated for scores which have two sub-groups, as with the figure classification test, which separately sumarizes male and female participants' scores.

[1] ISU students took only Part I due to time constraints.
[2] Averages calculated assuming 294 males and 323 females.
[3] Data from Part I only.

To get a single unified score with standard deviation, we completed the following calculations:

$$\mu_{\text{all}} = (N_F\mu_F + N_M\mu_M)/(N_F + N_M) \tag{4.3}$$

$$\sigma_{\text{all}} = \sqrt{\left(N_F\sigma_F^2 + N_M\sigma_M^2\right)/(N_F + N_M)}. \tag{4.4}$$

Here μ_F and μ_M are the mean scores for females and males, respectively; N_F and N_M are the number of female and male participants, and σ_F^2 and σ_M^2 are the variances in scores for females and males.

Substituting in the provided numbers, we get

$$\begin{aligned}
\mu_{\text{all}} &= \left(323\cdot 114.9 + 294\cdot 120.0\right)/(323+294) \\
&= 58.7 \\
\sigma_{\text{all}} &= \sqrt{\left(323\cdot 27.8^2 + 294\cdot 30^2\right)/(323+294)} \\
&= 14.4.
\end{aligned}$$

Whenever participants in two studies were not exposed to the same number of questions, the resulting scores are not comparable: both overall scores and their standard deviations are different. We can achieve comparability by scaling the scores accordingly. For example, in order to account for the fact that ISU students took only part I of two parts to the figure classification test (and thus completed half of the questions), we adjust the transformation as follows:

$$\begin{aligned}
\mu_{\text{part I}} &= 1/2\cdot\mu_{\text{all}} \\
\sigma_{\text{part I}} &= 1/\sqrt{2}\cdot\sigma_{\text{all}}
\end{aligned}$$

Once these adjustments have been completed, it is evident that Iowa State undergraduates scored at about the same level as other similar demographics. In fact, both means and standard deviations of ISU students' scores are similar to the comparison groups, which were chosen from available demographic groups based on population similarity.

Comparison population data was chosen to most closely match ISU undergraduate population demographics. Thus, if comparison data was available for 9th and 12th grade students,

scores of Iowa State students were compared to scores of 12th grade students, who are closer in age to college students. When data was available from college students and Army enlistees, comparisons of scores were based on other college students, as college students are more likely to have a similar gender distribution to ISU students.

Applying the grading protocol discussed in section 4.2.3, we see that the ranges of lineup and visuospatial test scores do not include zero; this indicates that we do not see random guessing from participants in any task. Figure 4.4 shows the range of possible scores and the observed score distribution. Participants' scores on the VST indicate score compression; that is, both participants with medium and high visual search abilities scored at the extremely high end of the spectrum. In future experiments, participants should be given less time (or more questions) to better differentiate participants with medium and high-ability.

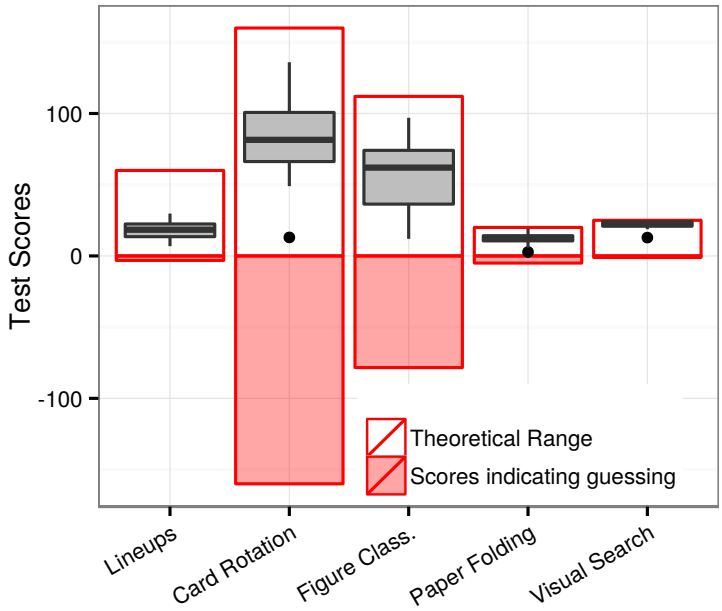


Figure 4.4: Test scores for lineups and visuospatial tests. As none of the participants scored at or below zero, we can conclude that there is little evidence of random guessing. We also note the score compression that occurs on the Visual Search test; this indicates that most participants scored extremely high, and thus, participants' scores are not entirely representative of their ability.

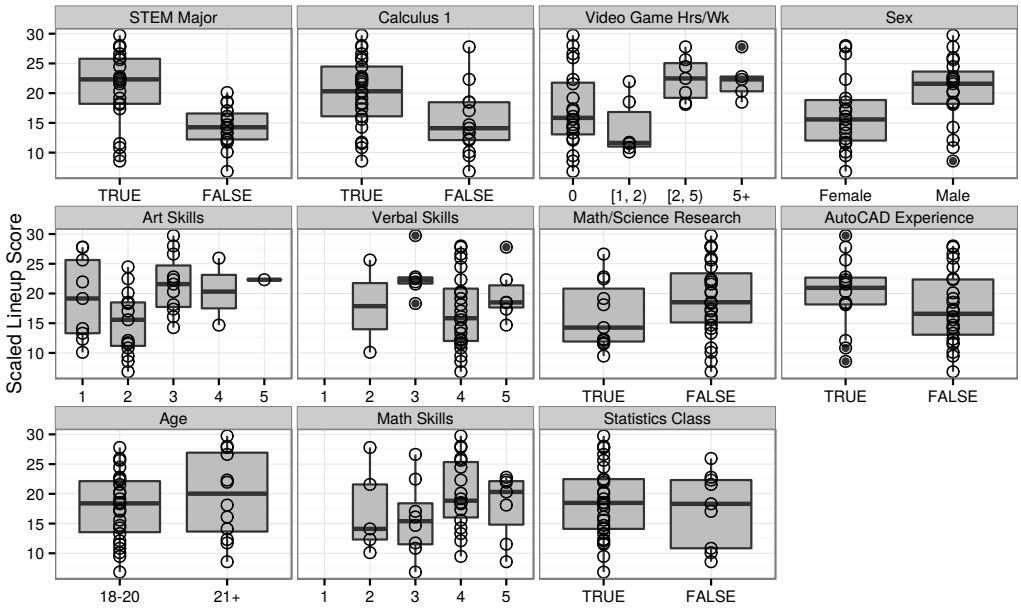


Figure 4.5: Demographic characteristics of participants compared with lineup score. Categories are ordered by effect size; majoring in a STEM field, calculus completion, hours spent playing video games per week, and sex are all associated with a significant difference in lineup score.

4.3.2 Lineup Performance and Demographic Characteristics

Previous work found a relationship between lineup performance and demographic factors such as education level, country of origin, and age (Majumder et al., 2014a); our participant population is very homogeneous, which allows us to explore factors such as educational background and skills on performance in lineup tests.

Figure 4.5 shows participants' lineup scores in relationship to their responses in the questionnaire given at the beginning of the study; this allows us to explore effects of demographic characteristics (major, research experience, etc.) on test performance.

Completion of Calculus I is associated with increased performance on lineups; this may be related to general math education level, or it may be that success in both lineups and calculus requires certain visual skills. This association is consistent with findings in (Shah and Carpenter, 1995), which associated mathematical ability to performance on simple graph description tasks. There is also a significant relationship between hours of video games played per week and score on lineups, however, this association is not monotonic and the groups do not have equal sample size, so the conclusion may be suspect. There is a (nearly) significant

difference between male and female performance on lineups; this is not particularly surprising, as men perform better on many spatial tests (Voyer et al., 1995) and performance on spatial tests is correlated with phase of the menstrual cycle in women (Hausmann et al., 2000). There is no significant difference in lineup performance for participants of different age, self-assessed skills in various domains, previous participation in math or science research, completion of a statistics class, or experience with AutoCAD. These demographic characteristics were chosen to account for life experience and personal skills which may have influenced the results.

Table 4.2 provides the results of a sequence of linear models fit to the lineup data. Each row in the table represents a single model, with one predictor variable (a factor with two or more levels). Due to sample size considerations, multiple testing corrections were not performed; in addition, the independent variables are correlated: in our sample, males are more likely to have completed Calculus 1, but are also more likely to spend time playing video games. As such, a model including two or more of the significant predictor variables shows all included variables to be nonsignificant. To better understand the effects of these variables, a larger study is necessary.

Table 4.2: Participant demographics' impact on lineup score. The table below shows each single demographic variable's association with lineup score. STEM major, completion of Calculus I, time spent playing video games, and gender all show some association with score on statistical lineups.

Variable	DF	MeanSq	F	p.val
STEM Major	1	401.517	14.44	0.001
Calculus 1	1	204.569	6.15	0.018
Video Game hrs	3	108.847	3.44	0.028
Sex	1	140.844	4.02	0.053
Art Skills	4	75.891	2.28	0.082
Verbal Skills	3	60.220	1.68	0.191
STEM Research	1	59.670	1.60	0.214
AutoCAD	1	50.893	1.36	0.252
Age	1	34.434	0.91	0.348
Math Skills	3	37.039	0.98	0.416
Statistics Class	1	9.062	0.23	0.631

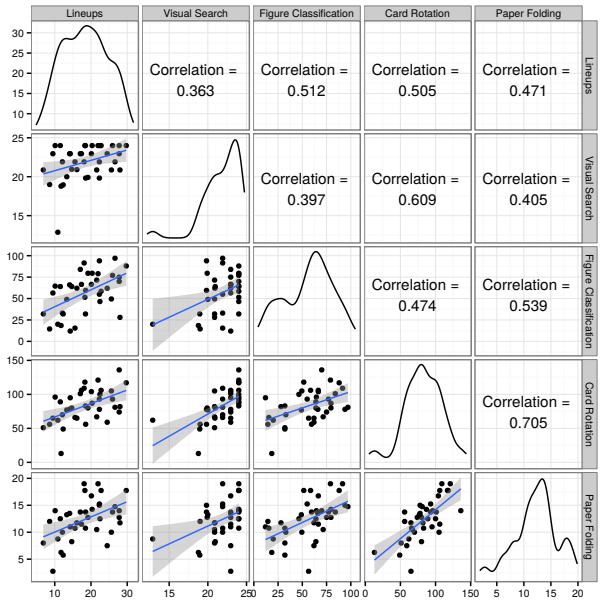


Figure 4.6: Pairwise scatterplots of test scores. Lineup scores are most highly correlated with figure classification scores, and are also highly correlated with card rotation scores. Paper folding and card rotation scores are also highly correlated.

4.3.3 Understanding Visual Abilities and Lineup Performance

Results from the visuospatial tests used in this experiment are highly correlated, as shown in Figure 4.6; this is to be expected given that all of these tests are in some way measuring individuals' visual ability. What is of more interest to us is how other factors, such as e.g. general intelligence, mental processing speed, cognitive resources, motivation, and attention affect performance. In order to assess factors contributing to lineup performance, we first examine the separate dimensions measured by the battery of cognitive tests (other than lineups) using principal components analysis on the scaled test scores, then we examine all five tests using the same procedure.

4.3.3.1 Principal Component Analysis of the Four Visuospatial Tests

A principal component analysis (PCA) of the four established visuo-spatial tests reveals that they all share a very strong first component, which explains about 64% of the total variability. Principal components (PC-) are ordered by importance (how much variability in the data they contain), and each principal component is uncorrelated with every other principal component.

PC1 is essentially an average across all tests representing a general "visual intelligence" factor. The other principal components span another two dimensions, while the last dimension is weak (at 6%). PC2 differentiates the figure classification test from the visual searching test, whereas PC3 differentiates these two tests from the paper folding test.

Table 4.3: Importance of principal components in an analysis of four tests of spatial ability: figure classification, paper folding, card rotation, and visual search.

	PC1	PC2	PC3	PC4
Standard deviation	1.61	0.81	0.73	0.49
Proportion of Variance	0.64	0.16	0.13	0.06
Cumulative Proportion	0.64	0.81	0.94	1.00

Table 4.3 contains the proportion of the variance in the four cognitive tasks represented by each principal component. PC1 accounts for about 60% of the variance; Figure 4.7 and Table 4.4 confirm that PC1 is a measure of the similarity between all 4 tests; that is, a participant's general (or visual) aptitude. PC2 differentiates the figure classification test from the visual searching test, while PC3 differentiates these two from the paper folding test. PC4 is not particularly significant (it accounts for 5.9% of the variance), but it differentiates the card rotation task from the paper folding task.

Table 4.4: Rotation matrix for principal component analysis of the four cognitive tests (visual search, paper folding, card rotation, figure classification).

	PC1	PC2	PC3	PC4
card.rot	0.55	-0.19	-0.38	0.72
fig.class	0.46	0.58	0.66	0.14
folding	0.52	0.33	-0.53	-0.59
vis.search	0.46	-0.72	0.38	-0.34

Figure 4.7 shows that the first PC does not differentiate between any of the tasks; it might be best understood as a general aptitude factor. All of the remaining principal components distinguish between the cognitive tasks; PC2 and PC3 separate paper folding from visual search and from the lineup and figure classification tasks, while PC4 and PC5 mainly separate lineups from card rotation and figure classification. This separation allows us to compare the tasks which are similar from among the principal components. According to Table 4.4, the first three principal components account for 94.1% of the variance.

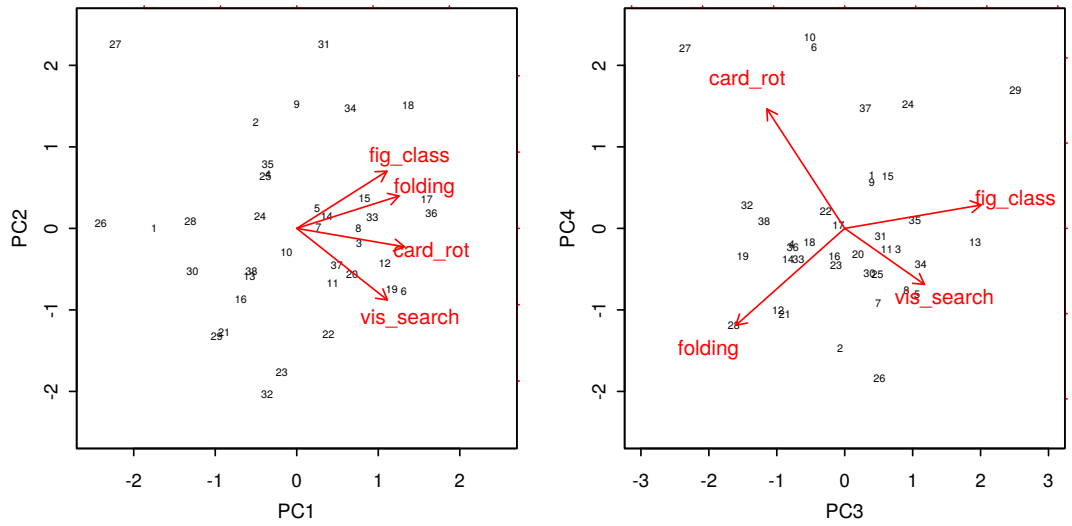


Figure 4.7: Biplots of principal components 1-4 with observations. Principal component analysis was performed on the four cognitive tests used to understand the association between the cognitive skills required for these tests and the skills required for the lineup protocol.

4.3.3.2 Principal Component Analysis of Cognitive Tests and the Lineup Task

Incorporating the lineup task into the principal component analysis, we find the principal components to be fairly similar to the four-component analysis. Table 4.5 shows the importance of each principal component. From the distribution of the variance components, we see that the lineup test spans an additional dimension within the space of the four established tests.

Table 4.5: Importance of principal components, analyzing all five tests.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.73	0.84	0.75	0.70	0.48
Proportion of Variance	0.60	0.14	0.11	0.10	0.05
Cumulative Proportion	0.60	0.74	0.85	0.95	1.00

Table 4.6: PCA Rotation matrix for all five tests.

	PC1	PC2	PC3	PC4	PC5
lineup	0.42	0.49	-0.46	0.60	-0.10
card.rot	0.50	-0.30	0.28	0.23	0.73
fig.class	0.43	0.45	-0.15	-0.75	0.18
folding	0.47	0.07	0.68	0.04	-0.56
vis.search	0.41	-0.69	-0.48	-0.15	-0.33

From the rotation matrix (see Table 4.6) we see that the first principal component, PC1, is

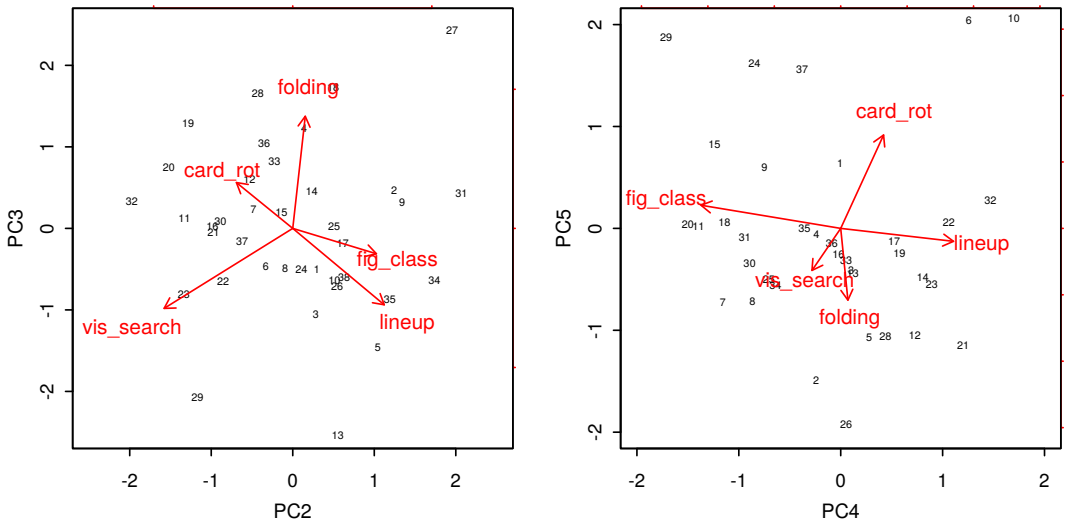


Figure 4.8: Biplots of principal components 2-5 with observations. The lineup task appears to be most similar to the figure classification task, based on the plot of PC2 vs. PC3.

again essentially an average across all tests and accounts for 60.1% of the variance in the data.

Figure classification is strongly related to lineups (PC2, PC3). Performance on the visual search task is also related to lineup performance (PC3). These two components highlight the shared demands of the lineup task and the figure classification task: participants must establish categories from provided stimuli and then classify the stimuli accordingly.

The visual search task is also clearly important to lineup performance: PC3 captures the similarity between the visual search and lineup performance, and aspects of these tasks are negatively correlated with aspects of the paper folding and card rotation tasks within PC3. Paper folding does not seem to be strongly associated with lineup performance outside of the first principal component; card rotation is only positively associated with lineup performance in PC4.

PC4 captures the similarity between lineups and the card rotation task and separates this similarity from the figure classification task; this similarity does not account for much extra variance (10%), but it may be that only some lineups require spatial rotation skills. PC5 contains only 5% of the remaining variance, and is thus not of much interest, however, it seems to capture the relationship between the card rotation task and the paper folding and visual search tasks.

Figure classification is strongly related to lineups, and as in the four-component PCA, figure classification is strongly represented in the first two principal components. While lineups do span a separate dimension, the PCA suggests that they are most closely related to the figure classification task, and least related to the visual searching task.

This emphasizes the underpinnings of lineups: the test utilizes a visual medium, but is ultimately are a classification task presented in a graphical manner. Using lineups as a proxy for statistical significance tests is similar to using a classifier on pictoral data: while the data is presented "graphically", the participant is actually classifying the data based on underlying summary statistics.

4.3.4 Linear model of demographic factors

Note that all of the demographic variables in the survey are highy correlated, for example there is a high correlation between STEM majors and taking calculus. Similarly, the correlation between having taken a statistics class and having been involved in mathematics or statistics research is high. Only one student is doing research who has not taken a statistics course.

A principal component of the five math/stats questions splits the variables into two main areas: the first principal component is an average of math skills, calculus 1 and STEM, while the second principal component is an average of having taken a statistics class and doing research. We therefore decided to use sums of these variables to come up with a separate math and a stats score. Note, that the correlation between the math and the stats score is almost zero.

We fit a linear model of lineup scores in the thus modified demographic variables and the test scores from the visuo-spatial tests, selecting the best model using AIC and stepwise backwards selection. The result is shown in Table 4.7. Only two covariates stay in the model: PC1 and MATH, reflecting two dimensions of what affects lineup scores. We can think of PC1 as a measure of innate visual or intellectual ability, while the MATH score is a matter of both ability and training. The remaining principal components were not sufficiently associated with lineup score to be included in the model.

Table 4.7: Estimates for a linear model of lineup scores.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.1192	1.9149	7.37	0.0000
PC1	1.7672	0.5230	3.38	0.0018
MATH	2.1246	0.8732	2.43	0.0202

4.3.5 Lineup Types

Each of the three sets of 20 lineups was taken from previous studies on different designs to investigate which plot type most effectively conveyed important characteristics of the underlying data set.

4.3.5.1 Example Lineups

4.3.6 Lineup Set 1

The experiment in the first lineup section examined the use of boxplots, density plots, histograms, and dotplots to compare two groups which vary in mean and sample size. The experiment was originally designed to explore the use of lineups to test plots of competing design(Hofmann et al., 2012). This set of lineups consists of 20 plots selected from the plots used in the full experiment; each set of data is displayed with each of the four plot types.

4.3.7 Lineup Set 2

The second lineup section also explored two groups of data, this time comparing boxplots, bee swarm boxplots, boxplots with overlaid jittered data, and violin plots. Participants were much more accurate in this experiment than in the experiment described previously, because of the types of plots compared as well as the underlying data distributions.

4.3.8 Lineup Set 3

The final lineup section explored QQ-plots from various model simulations, using reference lines, acceptance bands, and rotation to determine which plots allowed participants to most effectively identify violations of normality. Rotated QQ-plots showed lower performance because participants were able to more accurately compare acceptance bands to residuals, and thus

could identify that the reference bands were too liberal. As a result, performance was somewhat lower for rotated plots, even though participants were more accurate when comparing the residuals to the reference bands.

4.3.8.1 Cognitive Test Scores and Lineup Performance by Lineup Task

Figure 4.12 shows the correlations between the three lineup tasks and the figure classification, card rotation, and paper folding tasks. The visual search task is only slightly correlated with the three lineup tasks and is therefore omitted from this figure. Performance on lineup tasks 1 and 2, which dealt with the distributuion of two groups of numerical data, is most strongly correlated with the performance on the figure classification task, which measures general reasoning ability. Performance on lineup task 3, which investigated the potential to visually identify nonnormality in residual QQ-plots, is more associated with the card rotation and paper folding tests, which measure visuospatial ability. This suggests that certain lineup tasks may require more visual ability than others; in the case of the QQ-lineups a successful evaluation needed participants to mentally rotate plots to compare vertical distances, requiring more mental manipulation than the first two lineup tasks.

In order to examine which lineup tasks are most closely associated with visual abilities tested in the aptitude portions of an experiment, we employ principal component analysis on participant scores averaged across each block of lineups.

Principal component analysis separates multivariate data into orthogonal components using a rotation matrix to transform correlated input data into an orthogonal space. The importance of a principal component is also evaluated to assess the proportion of the overall variance in the data contained within the component (Table 4.5 shows the importance of each PC for the analysis of the aggregate lineup score and the four aptitude tests).

For variables X_i, $i = 1, ..., K$ a principal component analysis results in a set of K principal components PC_j, $j = 1, ..., K$, given as the rows of the rotation matrix, M (which has dimension $K \times K$, indexed by i and j respectively).

The importance I_j of each component is determined by the amount of variability the data exhibits along each of the principal axis.

The influence of variable X_i on component PC_j is then defined as:

$$\text{Influence of variable} X_i \text{ on PC}_j = M_{ij} \times I_j$$

Influence is therefore a measure of the contribution of an input variable to the principal component, scaled by the importance of that principal component to the overall variability in the data. Small (absolute) values indicate that there is little influence of X_i on PC_j; negative values show the direction of the influence (to maintain the separation of variables as part of principal components analysis).

Figure 4.13 shows the influence of each input variable on each principal component for a principal component analysis including each lineup task as a separate input variable. The input variables are shown on the y axis, with the influence of the variable on each PC shown on the x axis. This allows us to consider the rotation matrix visually (while accounting for the importance of each principal component); for instance, we see that again, the first PC accounts for most of the variance and seems to represent general visual aptitude.

PC2 emphasizes the overlapping variation in performance on lineup test 1 and the figure classification test. PC3 emphasizes the additional variation in performance on lineup test 3 and the visual search test, while PC4 emphasizes the extra variability in performance on lineup test 2 and the paper folding test. All three lineups, plus the figure classification, paper folding, and visual search tests contribute to PC5. PC6 and PC7 jointly account for less than 10% of the variance in the data and do not display any distinct patterns in the loadings.

While lineups constitute a distinct principal component when aggregated into a single score, this PCA of the separate lineup types and the cognitive tests indicates that different lineup experiments exist in different principal component loadings. Overall, there is an additional principal component gained from separating the lineup blocks by experiment type. As lineup tasks 1 and 2 contained similar plot types, it is possible that those two tasks overlap in the component space while lineup task 3 is distinct.

The relationship between participant performance on different types of lineups (and different types of plots) and performance on tests of spatial ability bears further investigation; this study suggests that there may be an effect, but there is simply not enough variation to make

definitive conclusions about the relationship between different measures of visual aptitude and performance on specific lineup tasks.

A larger study might not only address the question of which lineup tasks require certain visual skills, but also the use of different types of plots from a perceptual perspective. Preliminary results of performance on different types of plots are shown below, but a larger study is needed for definitive results. The advantage of the lineup protocol is that it allows us to not only consider individual performance but also to compare aggregate performance on different types of plots. Integrating information about the visual skills required for each type of plot provides information about the underlying perceptual skills and experience required to read different types of plots.

4.3.9 Lineup Plot Types

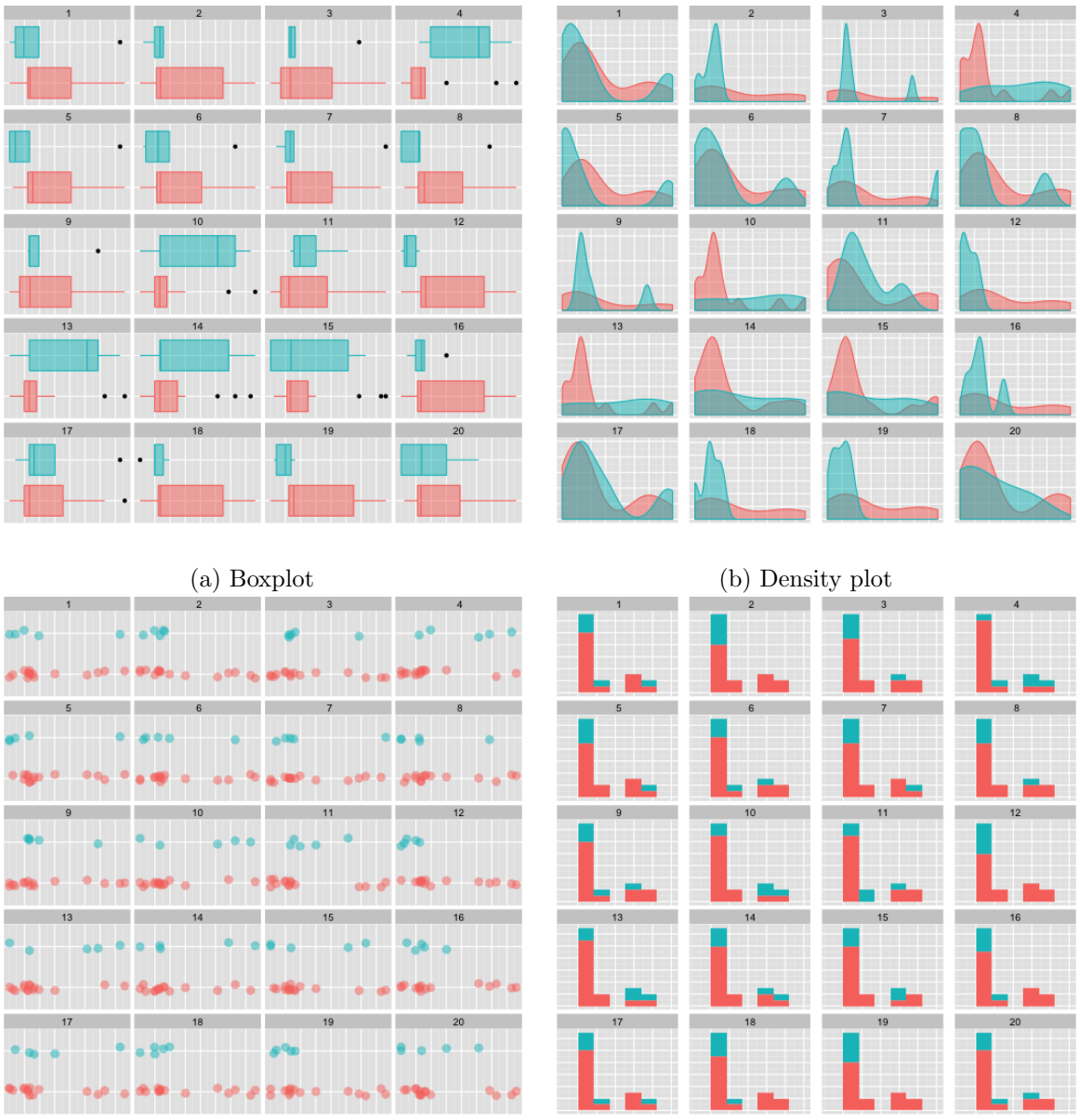
We can also compare participants' performance on specific types of lineup plots compared with their scores on the visual aptitude tests, for instance, accuracy on lineups which require mental rotation may be related to performance on the card rotation task.

Figure 4.14 compares performance on each different type of plot. The x axis shows scaled score, the y axis shows the density of participant scores. As two different lineup tasks utilized boxplots to test different qualities of the distribution of data (outliers vs. difference in medians), different tasks are shown as different colors, so that accuracy on tasks which are shown in blue can be compared to other blue density curves.

Figure 4.15 shows the association between scaled score on each type of lineup and score on the visual reasoning tests. Sample size for each plot type is fairly small - between 5 and 10 plots per individual, so there is low power for systematic inference, but we can establish that the card rotation task is much more significantly associated with the QQ-plots tasks compared to the other tasks. In addition, rotated QQ-plots seem to be much more asssociated with the paper folding task scores than other QQ-plot tasks; this may be because they require more visual manipulation than other QQ-plots.

For comparison, the correlation between general lineup score (non-subdivided) and the card rotation test score was 0.505, the correlation between general lineup score and the figure

classification test was 0.512, and the correlation between lineup score and the paper folding test was 0.471.While we can compare the correlation strength between tasks, it is clear that the correlation between the score on any single lineup type and a particular visual aptitude score is lower than the overall relationship that we attribute to visual ability. Additional data is imperative to understand the reasoning required for specific types of plots - it is likely that the 5-10 trials per participant presented in each chart in Figure 4.15 are simply not sufficient to uncover any specific relationship between reasoning ability and lineup task.



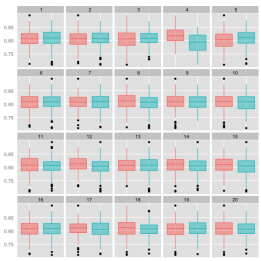
(a) Boxplot

(b) Density plot

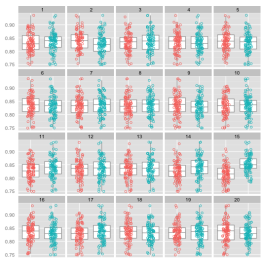
(c) Dotplot

(d) Histogram

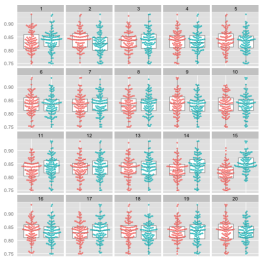
Figure 4.9: Types of plots in the first set of lineups. These plots are used to compare two distributions.



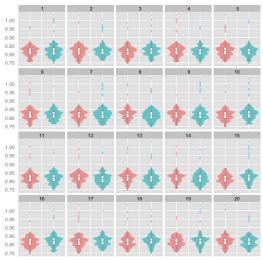
(a) Boxplots.



(b) Boxplots with jittered points.

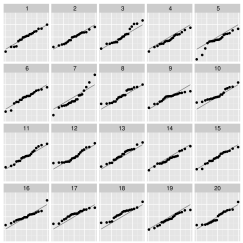


(c) Bee swarm boxplots.

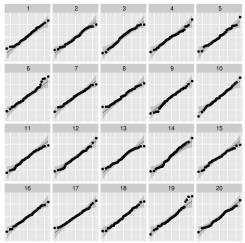


(d) Violin plots.

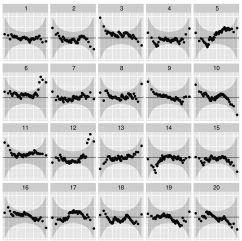
Figure 4.10: Types of plots in the second set of lineups. These plots are used to compare two distributions.



(a) QQ-plot with guide line.



(b) QQ-plot with acceptance bands.



(c) QQ-plot rotated 45 degrees.

Figure 4.11: Types of plots in the third set of lineups. These plots are used to assess residual normality.

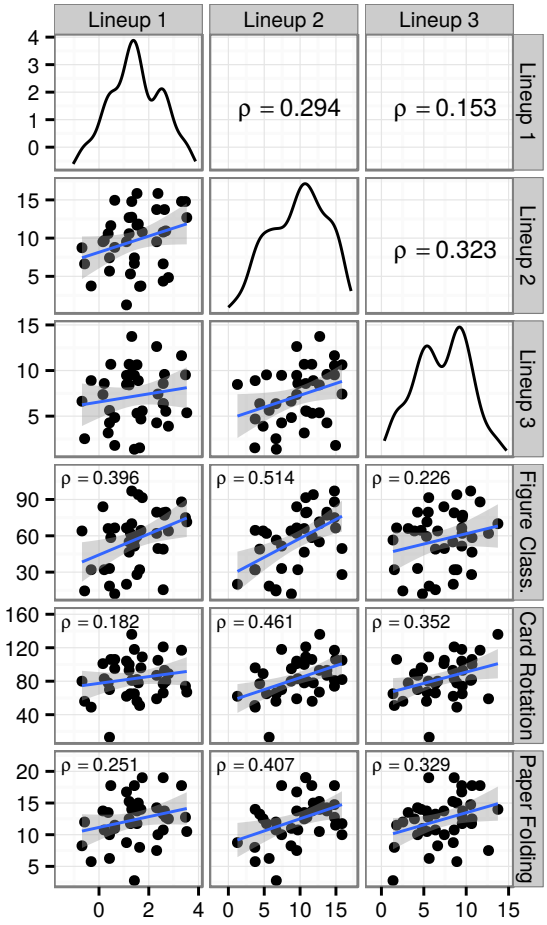


Figure 4.12: Pairwise scatterplots of test scores, with lineups separated into the three lineup tasks. All lineup tasks are moderately correlated with the figure classification task, and while tasks 1 and 2 are most strongly correlated with figure classification, lineup task 3 is most strongly correlated with the card rotation and paper folding tasks.

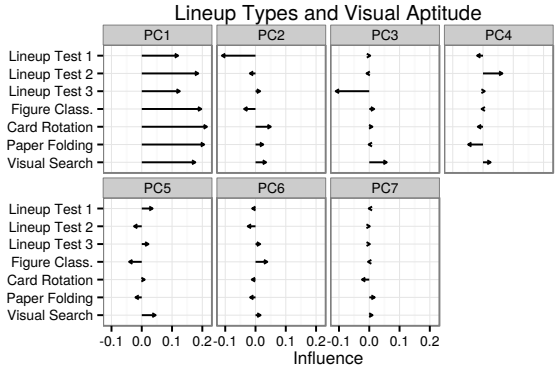


Figure 4.13: Principal Component influence plot, showing each input variable's contribution to the principal component, scaled by the proportion of variability in the data contained in the principal component.

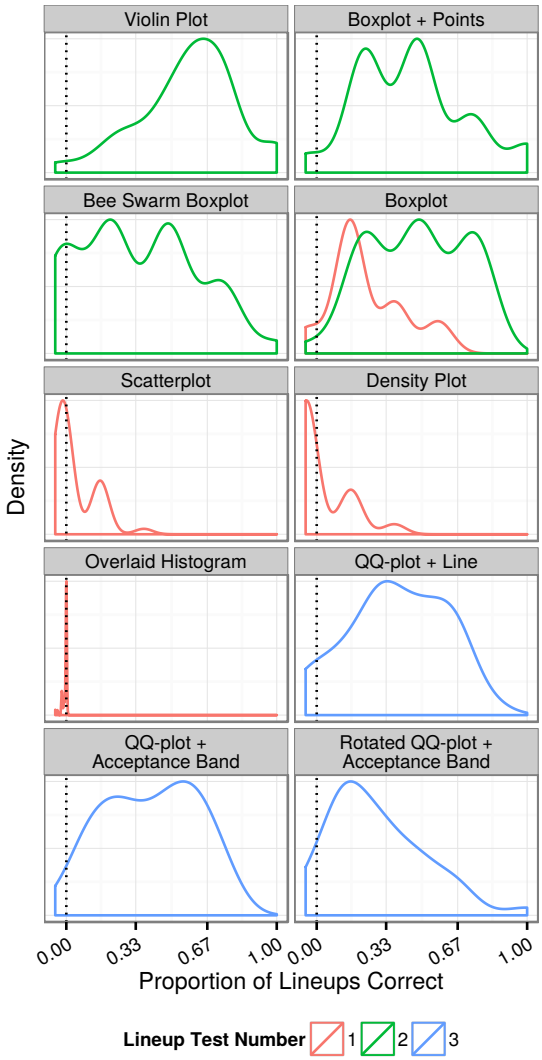


Figure 4.14: Density plots of scaled scores for different types of lineups. For the same experiment (shown by line color), certain types of plots are more difficult to read and are associated with lower participant scores.

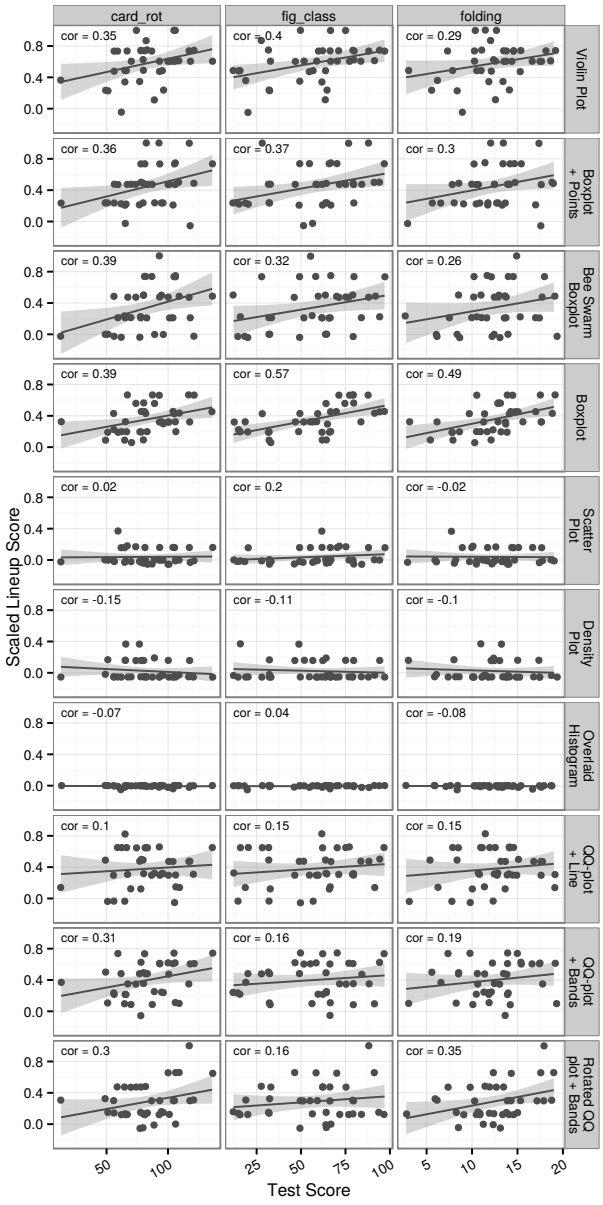


Figure 4.15: Scatterplots of scaled lineup scores by aptitude test scores. There is some indication that different types of lineup tasks may utilize different visual skills; for instance, QQ-plots with confidence bands may require more skill at mental rotation than QQ-plots without the bands.

4.4 Discussion and Conclusions

Performance on lineups is strongly related to performance on tests of visual ability; however, this relationship is mediated by demographic factors such as major (STEM or not) and completion of calculus I. In addition to these demographic factors, many facets of intelligence are highly correlated; participants who score higher on general aptitude tests may score higher on tests of visual ability (and may also score higher on lineup tests).

Despite these caveats, we have demonstrated that the general lineup task is most closely related to a classification task, rather than tests of spatial ability. This is an important verification of a tool that is useful for examining statistical graphics, as it emphasizes the idea that while the testing medium is graphical in nature, the task is in fact a classification task, where the viewer must determine the most important features of each plot and then identify which plot is different.

When lineup tasks with different goals are viewed separately, there is some indication that different tasks are associated with different visual abilities. Lineup tasks 1 and 2 are quite similar, and are more associated with the figure classification task; lineup task 3, while still moderately correlated with the figure classification task, is also moderately correlated with the visuospatial ability tests (paper folding, card rotation). Future studies testing larger sets of lineups may be useful to understand which types of plots require additional visuospatial skills, as plots which appeal to a wider audience may be more successful when conveying important information.

In addition to this theoretical information, the figure classification test may be useful for pre-screening participants in future online lineup studies. Such studies often suffer from participants who do not take the task seriously, and internal verification questions, as well as pre-qualification tasks are often used to reduce extraneous variability. While it is impractical to require participants to score well on several different tests, it is reasonable to ask participants to pre-qualify for a task by completing a figure classification test. As the figure classification test is different from the lineup task, this will not bias participants' scores on the domain of interest while ensuring that the participant pool is sufficiently motivated to complete the lineup

questions.

The demographic results from this study indicate that in future lineup studies, it may be important to record information about participants' mathematical training, so that studies can be compared across participant pools with more reliability.

All results and data shown here were collected and analyzed in accordance with IRB # 13-581.

CHAPTER 5. STATISTICAL GRAPHICS AND FEATURE HIERARCHY

Intended for submission to the Journal of Computational and Graphical Statistics

5.1 Introduction and background

Numerical information can be difficult to communicate effectively in raw form, due to limits on attention span, short term memory, and information storage mechanisms within the human brain. Graphics are much more effective for communicating numerical information, as (well-designed) graphics order the numerical information spatially and utilize the higher-bandwidth visual system. Visual data displays serve as a form of external cognition (Zhang, 1997; Scaife and Rogers, 1996), ordering and visually summarizing data which would be hopelessly confusing in tabular format. One fantastic example of this phenomenon is the Hertzsprung-Russell (HR) diagram, which was described as "one of the greatest observational syntheses in astronomy and astrophysics" because it allowed astronomers to clearly relate the absolute magnitude of a star to its' spectral classification; facilitating greater understanding of stellar evolution (Spence and Garrison, 1993). The data it displayed was previously available in several different tables; when plotted on the same chart, information that was invisible in a tabular representation became immediately clear (Lewandowsky and Spence, 1989b). Graphical displays more efficiently utilize cognitive resources by reducing the burden of storing, ordering, and summarizing raw data; this frees bandwidth for higher levels of information synthesis, allowing observers to note outliers, understand relationships between variables, and form new hypotheses.

Graphical displays are powerful because they efficiently and effectively convey numerical information, but there exists relatively sparse empirical information about how the human

perceptual system processes these displays. Our understanding of the perception of statistical graphics is informed by general psychological and psychophysics research as well as more specific research into the perception of data displays (Cleveland and McGill, 1984).

One relevant focus of psychological research is pre-attentive perception, that is, perception which occurs automatically in the first 200 ms of exposure to a visual stimulus (Treisman, 1985).

Research into preattentive perception provides us with some information about the temporal hierarchy of graphical feature processing. Color, line orientation, and shape are processed preattentively; that is, within 200 ms, it is possible to identify a single target in a field of distractors, if the target differs with respect to color or shape (Goldstein, 2009a). Research by Healey and Enns (1999) extends this work, demonstrating that certain features of three-dimensional data displays are also processed preattentively. However, neither target identification nor three-dimensional data processing always translate into faster or more accurate inference about the data displayed, particularly when participants have to integrate several preattentive features to understand the data.

Feature detection at the attentive stage of perception has also been examined in the context of statistical graphics; researchers have evaluated the perceptual implications of utilizing color, fill, shapes, and letters to denote categorical or stratified data in scatterplots. Cleveland and McGill (1984) ranked the optimality of these plot aesthetics based on response accuracy, preferring colors, amount of fill, shapes, and finally letters to indicate category membership. Lewandowsky and Spence (1989a) examined both accuracy and response time, finding that color is faster and more accurately perceived (except by individuals with color deficiency). Shape, fill, and discriminable letters (letters which do not share visual features, such as HQX) were identified as less accurate than color, while confusable letters (such as HEF) result in significantly decreased accuracy.

Another area of psychological research, Gestalt psychology, examines perception as a holistic experience, establishing and evaluating mental heuristics used to transform visual stimuli into useful, coherent information. Gestalt rules of perception can be easily applied to statistical graphics, as they describe the way we organize visual input, focusing on the holistic experience

rather than the individual perceptual features.

For example, rather than perceiving four legs, a tail, two eyes, two ears, and a nose, we perceive a dog. This is due to certain perceptual heuristics, which provide a "top-down" method of understanding visual stimuli by taking into account past experience. The rules of perceptual grouping or organization, as stated in Goldstein (2009a) are:

- **Proximity**: two elements which are close together are more likely to belong to a single unit.
- **Similarity**: the more similar two elements are, the more likely they belong to a single unit.
- **Common fate**: two elements moving together likely belong to a single unit.
- **Continuity**: two elements which blend together smoothly likely belong to one unit.
- **Closure**: elements which can be assembled into closed or convex objects likely belong together.
- **Common region**: elements contained within a common region likely belong together.
- **Connectedness**: elements physically connected to each other are more likely to belong together.

The plots in figure 5.1 demonstrate several of the gestalt principles which combine to order our perceptual experience from the top down. These laws help to order our perception of charts as well: points which are colored or shaped the same are perceived as belonging to a group (similarity), points within a bounding interval or ellipse are perceived as belonging to the same group (common region), and regression lines with confidence intervals are perceived as single units (continuity, connectedness, closure, and/or common region depending on the rendering of the intervals).

The processing of visual stimuli utilizes low-level feature detection, which occurs automatically in the preattentive perceptual phase, and higher-level mental heuristics which are informed

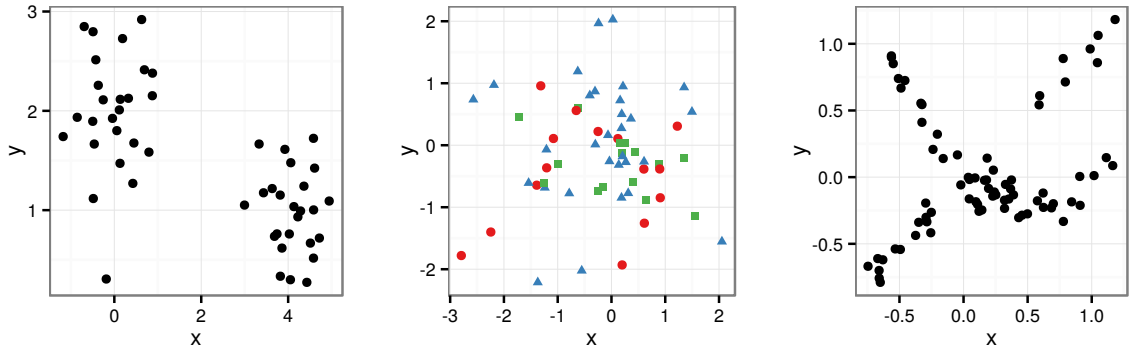


Figure 5.1: Proximity renders the fifty points of the first scatterplot as two distinct (and equal-sized) groups. Shapes and colors create different groups of points in the middle scatterplot, invoking the Gestalt principle of Similarity. Continuity renders the points in the scatterplot on the right hand side into two groups of points on curves: one a straight line with an upward slope, the other a curve that initially decreases and at the end of the range shows an uptick.

by experience. Both types of mental processes utilize physical location, color, and shape in order to organize our perception and direct attention to graphical features which stand alone. Research on preattentive perception is important because features that are perceived preattentively do not require as much mental effort to process from raw visual stimuli; subsequent top-down gestalt heuristics can be applied to the categorized features in order to make sense of the visual scene once the attentive stage of perception is reached.

This paper describes the results of a user study designed to explore the hierarchy of gestalt principles in perception of statistical graphics. We utilize information from previous studies (Demiralp et al., 2014; Robinson, 2003) concerning the hierarchy of preattentive feature perception in order to maximize the effect of preattentive feature differences.

Statistical graphics can be difficult to examine experimentally; qualitative studies rely on descriptions of the plot by participants who may not be able to articulate their observations precisely, while quantitative studies may only be able to examine whether the viewer can accurately read numerical information from the chart, instead of exploring the overall utility of the data display. Statistical lineups, described in the next section, are an important experimental tool for evaluating the perceptual utility of graphical displays. Lineups fuse commonly used psychological tests (target identification, visual search) 4 with statistical hypothesis tests to facilitate formal experimental evaluation of statistical graphics.

Statistical Lineups

Lineups are an experimental tool designed to serve as a visual hypothesis test, separating "significant" visual effects from those that would be expected under a null hypothesis (Buja et al., 2009; Majumder et al., 2013; Hofmann et al., 2012; Wickham et al., 2010). A statistical lineup consists of (usually) 20 sub-plots, arranged in a grid (examples are shown in figure 5.8). Of these plots, one plot is the "target plot", generated from either real data or an alternate model (equivalent to H_A in hypothesis testing); the other 19 plots are generated either using bootstrap samples of the real data or by generating "true null" plots from the null distribution H_0. If participants can identify the target plot from the field of distractors, then the visual display is deemed significant in the same sense that a numerical test with $p < 0.05$ is significant.

Apart from the hypothesis testing construct, the use of statistical lineups to test statistical graphics conforms nicely to psychological testing constructs such as visual search (DeMita et al., 1981; Treisman and Gelade, 1980), where a single target is embedded in a field of distractors and response time, accuracy, or both are used to measure the complexity of the underlying psychological processes leading to identification.

In this study, we modify the lineup protocol by introducing a second target to each lineup. The two targets represent two different, competing signals; the participant's choice then demonstrates empirically which signal is more salient. If both targets exhibit similar signal, participants may identify both targets, removing any forced-choice scenario which might skew results (few participants exercised this option).

By tracking the proportion of observers choosing either target plot (a measure of overall lineup difficulty) as well as which proportion of observers choose one target over the other target, we can determine the relative strength of the two competing signals amid a field of distractors. At this level, signal strength is determined by the experimental data and the generating model; we are measuring the "power" (in a statistical sense) of the human perceptual system, rather than raw numerical signal.

Using this testing framework, we apply different aesthetics, such as color and shape, as well as plot objects which display statistical calculations, such as trend lines and bounding

ellipses. These additional plot layers, discussed in more detail in the next section, are designed to emphasize one of the two competing targets and affect the overall visual signal of the target plot relative to the null plots. We expect that in a situation similar to the third plot of figure 5.1, the addition of two trend lines would emphasize the continuity of points in the plot, producing a stronger visual signal, even though the underlying data has not changed. Similarly, the grouping effect in the first plot in the figure would be enhanced if the points in each group were colored differently, as the proximity heuristic would be supplemented by similarity. In plots that are ambiguous, containing some clustering of points as well as a linear relationship between x and y, additional aesthetic cues may "tip the balance" in favor of recognizing one type of signal.

This study is designed to inform our understanding of the perceptual implications of these additional aesthetics, in order to provide guidelines for the creation of data displays which provide visual cues consistent with gestalt heuristics and preattentive perceptual preferences.

The next section discusses the particulars of the experimental design, including the data generation model, plot aesthetics, selection of color and shape palettes, and other important considerations. Experimental results are presented in section 5.3, and implications and conclusions are discussed in section 5.4.

5.2 Experimental Setup and Design

In this section, we discuss the generating data models for the two types of signal plots and the null plots, the selection of plot aesthetic combinations and aesthetic values, and the design and execution of the experiment.

5.2.1 Data Generation

Lineups require a single "target" data set (which we are expanding to two competing "target" data sets), and a method for generating null plots. When utilizing real data for target plots, null plots are often generated through permutations.

Here, it is possible to generate true null plots, which are generated from the null model and do not depend on the data used in the target plot. This experiment will measure two

competing gestalt heuristics, proximity and continuity, using two data-generating models: M_C, which generates data with K clusters, and M_T, which generates data with a positive correlation between x and y. True null datasets are created using a mixture model M_0 which combines M_C and M_T. Both M_C and M_T generate data in the same range of values. Additionally, M_C generates clustered data with linear correlations that are within $\rho = (0.25, 0.75)$, similar to the linear relationship between datasets generated by M_0, and M_T generates data with clustering similar to M_0. These constraints provide some assurance that participants who select a plot with data generated from M_T are doing so because of visual cues indicating a linear trend (rather than a lack of clustering compared to plots with data generated from M_0), and participants who select a plot with data generated from M_C are doing so because of visual cues indicating clustering, rather than a lack of a linear relationship relative to plots with data generated from M_0.

5.2.1.1 Regression Model M_T

This model has the parameter σ_T to describe the amount of scatter around the trend line. It generates N points $(x_i, y_i), i = 1, ..., N$ where x and y have a positive linear relationship. The data generation mechanism is as follows:

Algorithm 5.2.1

Input Parameters: sample size N, σ_T standard deviation around the line

Output: N points, in form of vectors x and y.

1. *Generate $\tilde{x}_i$, $i = 1, ..., N$, as a sequence of evenly spaced points from $[-1, 1]$.*
2. *Jitter $\tilde{x}_i$ by adding small uniformly distributed perturbations to each of the values: $x_i = \tilde{x}_i + \eta_i$, where $\eta_i \sim Unif(-z, z)$, $z = \frac{2}{5(N-1)}$.*
3. *Generate y_i as a linear regressand of x_i: $y_i = x_i + e_i$, $e_i \sim N(0, \sigma_T^2)$.*
4. *Center and scale x_i, y_i.*

We compute the coefficient of determination for all of the plots to assess the amount of linearity in each panel, computed as

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{5.1}$$

where TSS is the total sum of squares, $TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{N} e_i^2$, the residual sum of squares. The expected value of the coefficient of determination $E\left[R^2\right]$ in this scenario is

$$E\left[R^2\right] = \frac{1}{1 + 3\sigma_T^2},$$

because $E[RSS] = N\sigma_T^2$ and $E[TSS] = \sum_{i=1}^{N} E\left[y_i^2\right]$ (as $E[Y] = 0$), where

$$E\left[y_i^2\right] = E\left[x_i^2 + e_i^2 + 2x_i e_i\right] = \frac{1}{3} + \sigma_T^2.$$

The use of R^2 to assess the strength of the linear relationship (rather than the correlation) is indicated because human perception of correlation strength more closely aligns with R^2 (Bobko and Karren, 1979; Lewandowsky and Spence, 1989b).

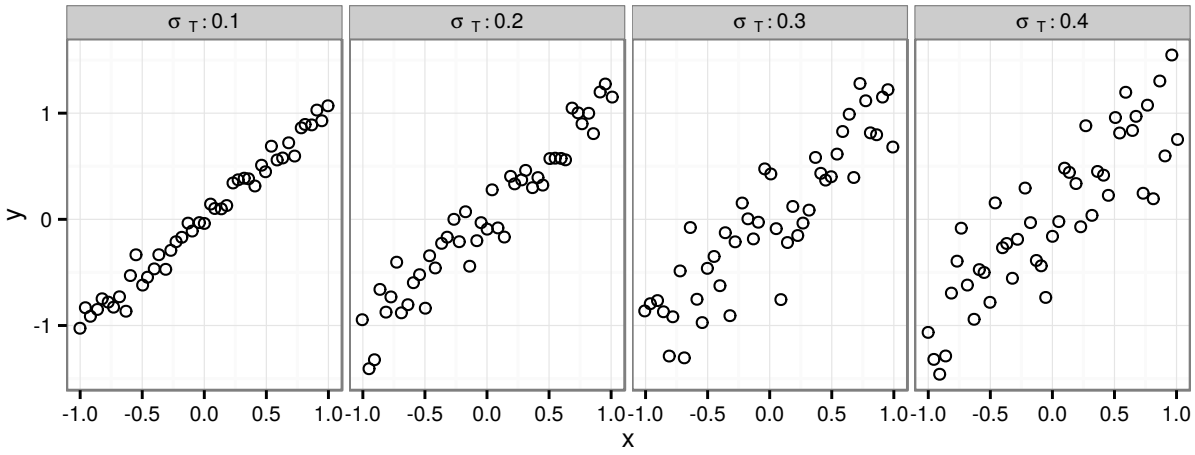


Figure 5.2: Set of scatterplots showing one draw each from the trend model M_T for parameter values of $\sigma_T \in \{0.1, 0.2, 0.3, 0.4\}$.

5.2.1.2 Cluster Model M_C

We begin by generating K cluster centers on a $K \times K$ grid, then we generate points around selected cluster centers. Parameters K and σ_C describe the number of clusters and the variability around cluster centers, respectively.

Algorithm 5.2.2

Input Parameters: N points, K clusters, σ_C cluster standard deviation

Output: N points, in form of vectors x and y.

1. *Generate cluster centers (c_i^x, c_i^y) for each of the K clusters, $i = 1, ..., K$:*

   (a) *in form of two vectors c^x and c^y of permutations of $\{1, ..., K\}$, such that*

   (b) *the correlation between cluster centers $cor(c^x, c^y)$ falls into a range of $[.25, .75]$.*

2. *Center and standardize cluster centers (c^x, c^y):*

$$\tilde{c}_i^x = \frac{c_i^x - \bar{c}}{s_c} \quad \text{and} \quad \tilde{c}_i^y = \frac{c_i^y - \bar{c}}{s_c},$$

   where $\bar{c} = (K+1)/2$ and $s_c^2 = \frac{K(K+1)}{12}$ for all $i = 1, ..., K$.

3. *For the K clusters, we want to have nearly equal sized groups, but allow some variability. Group sizes $g = (g_1, ..., g_K)$ with $N = \sum_{i=1}^K g_i$, for clusters $1, ..., K$ are therefore determined as a draw from a multinomial distribution:*

$$g \sim \text{Multinomial } (K, p) \text{ where } p = \tilde{p} / \sum_{i=1}^{K} \tilde{p}_i, \text{ for } \tilde{p} \sim N\left(\frac{1}{K}, \frac{1}{2K^2}\right).$$

4. *Generate points around cluster centers by adding small normal perturbations:*

$$\begin{aligned} x_i &= \tilde{c}_{g_i}^x + e_i^x, \text{ where } e_i^x \sim N(0, \sigma_C^2), \\ y_i &= \tilde{c}_{g_i}^y + e_i^y, \text{ where } e_i^y \sim N(0, \sigma_C^2). \end{aligned}$$

5. *Center and scale x_i, y_i.*

As a measure of clustering we use a coefficient to assess the amount of variability within groups, compared to total variability. Note that for the purpose of clustering, variability is measured as the variability in both x and y from a common mean, i.e. we implicitly assume that the values in x and y are on the same scale (which we achieve by scaling in the final step of the generation algorithm).

For two numeric variables x and y and grouping variable g with $g_i \in \{1, ..., K\}, i = 1, ..., n$, we compute the *cluster index* C^2 as follows: let $j(i)$ be the function that maps index $i = 1, ..., n$

to one of the clusters $1, ..., K$ given by the grouping variable g. Then for each level of g, we find a cluster center as $\bar{x}_{j(i)}$ and $\bar{y}_{j(i)}$, and we determine the strength of the clustering by comparing the within cluster variability with the overall variability:

$$\begin{aligned} C^2 &= 1 - \frac{CSS}{TSS}, \\ CSS &= \sum_{i=1}^{n} \left(x_{j(i)} - \overline{x}_{j(i)}\right)^2 + \left(y_{j(i)} - \overline{y}_{j(i)}\right)^2, \\ TSS &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + (y_i - \bar{y})^2 . \end{aligned} \tag{5.2}$$

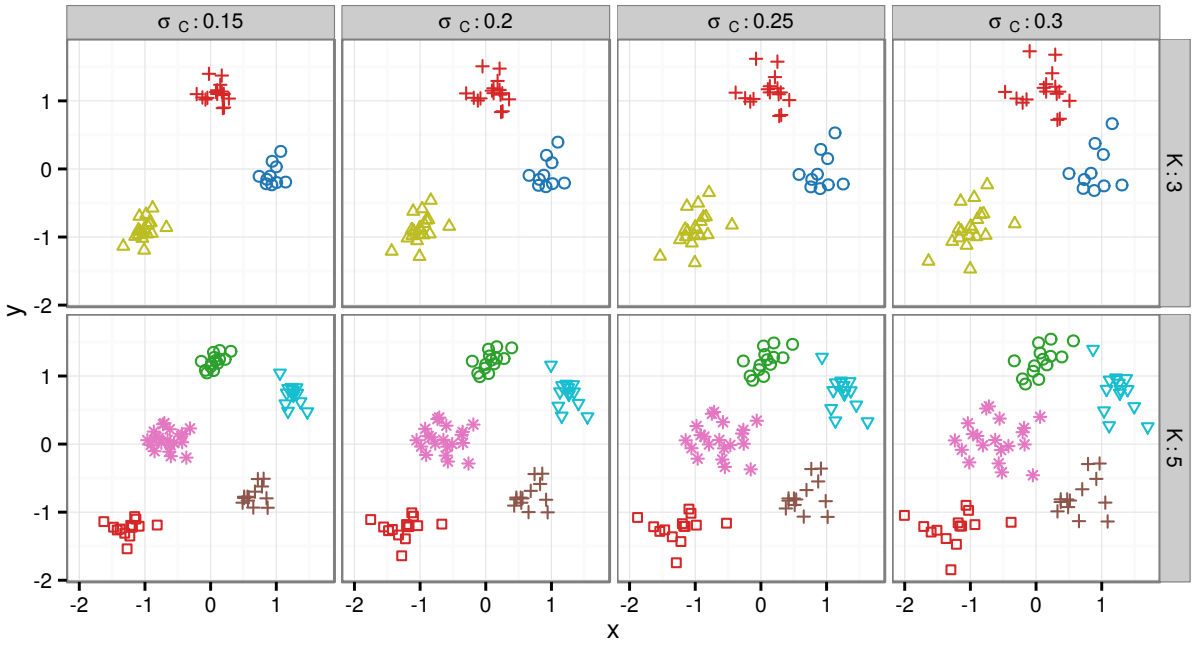


Figure 5.3: Scatterplots of clustering output for different inner cluster spread σ_C (left to right) and different number of clusters K (top and bottom), generated using the same random seed at each parameter setting. The colors and shapes shown are those used in the lineups for $K = 3$ and $K = 5$.

5.2.1.3 Null Model M_0

The generative model for null data is a mixture model M_0 that draws $n_c \sim \text{Binomial}(N, \lambda)$ observations from the cluster model, and $n_T = N - n_c$ from the regression model M_T. Observations are assigned groups using hierarchical clustering, which creates groups consistent

with any structure present in the generated data. This provides a plausible grouping for use in aesthetic and statistics requiring categorical data (color, shape, bounding ellipses).

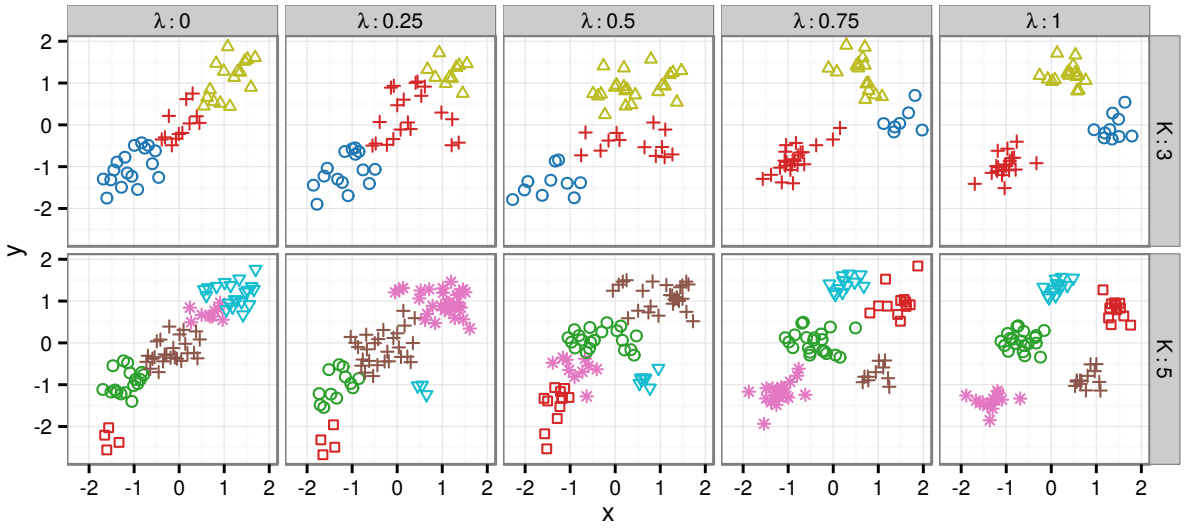


Figure 5.4: Scatterplots of data generated from M_0 using different values of λ, generated using the same random seed at each λ value.

Null data in this experiment is generated using $\lambda = 0.5$, that is, each point in a null data set is equally likely to have been generated from M_C and M_T.

5.2.1.4 Parameters used in Data Generation

Models M_C, M_T, and M_0 provide the foundation for this experiment; by manipulating cluster standard deviation σ_C and regression standard deviation σ_T (directly related to correlation strength) for varying numbers of clusters $K = 3, 5$, we can systematically control the statistical signal present in the target plots and generate corresponding null plots that are mixtures of the two distributions. For each parameter set $\{K, N, \sigma_C, \sigma_T\}$, as described in table 5.1, we generate a lineup dataset consisting of one set drawn from M_C, one set drawn from M_T, and 18 sets drawn from M_0.

The parameter values were chosen after examining the parameter space through simulation of 1000 lineup datasets for each combination of $\sigma_T \in \{0.2, 0.25, ..., 0.5\}$, $\sigma_C \in \{0.1, 0.15, ..., 0.4\}$, and $K \in \{3, 5\}$. Each lineup dataset consists of one trend target dataset generated using M_T, one cluster target dataset generated from M_C, and 18 null datasets generated from M_0; for

Parameter	Description	Choices
K	# Clusters	3, 5
N	# Points	$15 \cdot K$
σ_T	Scatter around trend line	.25, .35, .45
σ_C	Scatter around cluster centers	.25, .30, .35 ($K = 3$) .20, .25, .30 ($K = 5$)

Table 5.1: Parameter settings for generation of lineup datasets.

each of these sub-lineup datasets generated, the previously described statistics for trend and cluster strength were computed. We compared the statistics for the relevant target plot to the most extreme value for the 18 null plots, using both density plots (as shown in figure 5.5) and higher-level summaries as shown in figures 5.6 and 5.7.

These distributions allow us to objectively assess the difficulty of detecting the target datasets computationally (without relying on human perception). A target plot with $R^2 = 0.95$ is very easy to identify when surrounded by null plots with $R^2 = 0.5$, while null plots with $R^2 = 0.9$ make the target plot more difficult to identify. This approach is similar to that taken in Roy Chowdhury et al. (2014).

Figure 5.5 shows densities of each measure computed from the maximum of 18 null plots compared to the measure in the signal plot for one combination of parameters. There is some overlap in the distribution of R^2 for the null plots compared to the target plot displaying data drawn from M_T. As a result, the distribution of the cluster statistic values is more easily separated from the null data sets than the distribution of the trend statistic, that is, $\sigma_C = 0.20$ is producing cluster target data sets that are a bit easier to identify numerically than trend targets with a parameter value of $\sigma_T = 0.25$.

Figures 5.6 and 5.7 show the 25th and 75th percentiles of the distribution of R^2 and cluster strength summary statistics for each set of parameter values. These plots guided our selection of parameter values for trend and cluster models.

Additionally, we note that there is an interaction between σ_C and σ_T: the distinction between target and null on a fixed setting of clustering becomes increasingly difficult as the standard deviation for the linear trend is increased, and vice versa. There may additionally be

Figure 5.5: Density of test statistics measuring trend strength and cluster strength for target distributions and null plots based on 1,000 draws of lineup data with $\sigma_T = 0.25, \sigma_C = 0.20$ and $K = 3$.

a three-way interaction between σ_C, σ_T, and K: the size of the blue intervals (bottom figure) changes in size between different levels of K, it changes for different levels of σ_C and σ_T. These interactions suggest that in order to examine differences in aesthetics, we must block by parameter settings (this can be accomplished using dataset-level blocks). Each dataset is non-deterministic, because we have a random process generating from different parameter settings, not a deterministic run as in an engineering setting. It is thus important to use replicates of each parameter setting to ensure that we can separate data-level effects from parameter-level effects.

Using information from the simulation, we identified values of σ_T and σ_C corresponding to "easy", "medium" and "hard" numerical comparisons between corresponding target data sets and null data sets. It is important to note that the numerical measures we have described in equations (5.1) and (5.2) only provide information on the numerical discriminability of the target datasets from the null datasets; the simulation cannot provide us with information on the perceptual discriminability, and it has been established that human perception of scatterplots does not replicate statistical measures exactly (Bobko and Karren, 1979; Mosteller et al., 1981; Lewandowsky and Spence, 1989b).

Each of the generated datasets is then plotted as a lineup, where we apply plot aesthetics, such as color and trend lines, which emphasize clusters and/or linear relationships respectively.

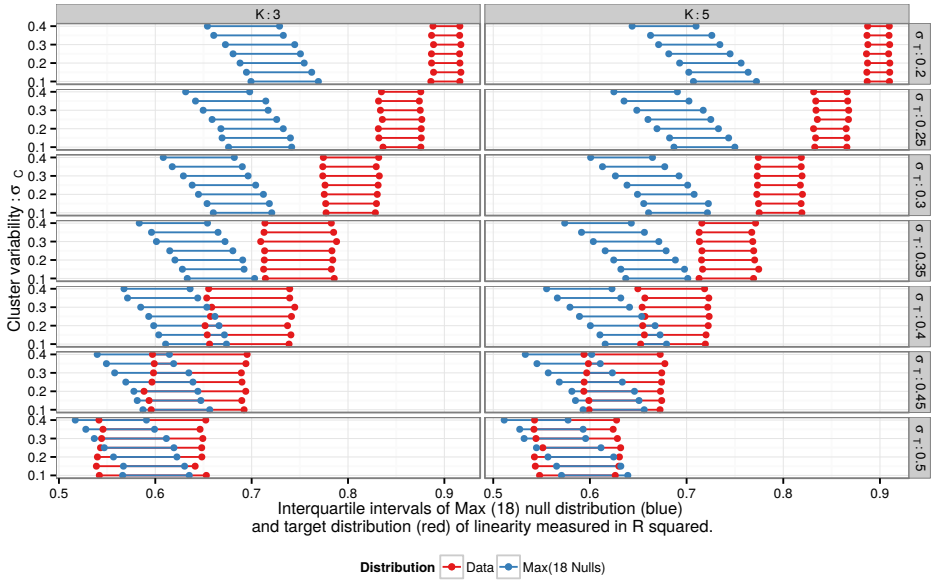


Figure 5.6: Simulated interquartile range of R^2 values for target and null data distributions.

Our goal is to experimentally determine how these aesthetics change participants' ability to identify each target plot. The next section describes the aesthetic combinations and their anticipated effect on participant responses.

5.2.2 Lineup Rendering

5.2.2.1 Plot Aesthetics

Gestalt perceptual theory suggests that perceptual features such as shape, color, trend lines, and boundary regions modify the perception of ambiguous graphs, emphasizing clustering in the data (in the case of shape, color, and bounding ellipses) or linear relationships (in the case of trend lines and prediction intervals), as demonstrated in figure 5.1. For each dataset we examine the effect of plot aesthetics (color, shape) and statistical layers (trend line, boundary ellipses, prediction intervals) shown in table 5.2 on target identification. Examples of these plot aesthetics are shown in figure 5.8.

We expect that relative to a plot with no extra aesthetics or statistical layers, the addition of color, shape, and 95% boundary ellipses increases the probability of a participant selecting the target plot with data generated from M_C, the cluster model, and that the addition of

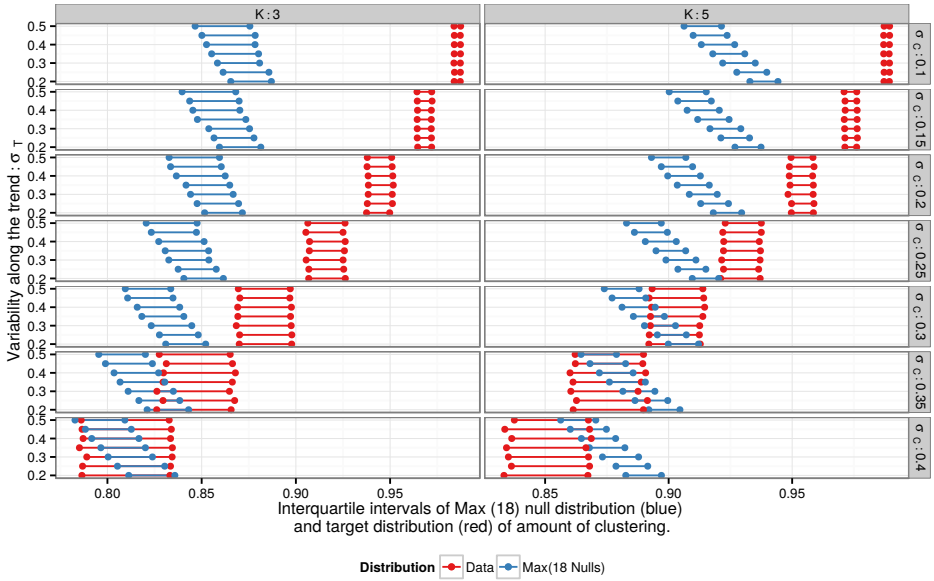


Figure 5.7: Simulated interquartile range of C^2 (cluster cohesion) values for target and null data distributions.

these aesthetics decreases the probability of a participant selecting the target plot with data generated from M_T, the trend model.

Similarly, we expect that relative to a plot with no extra aesthetics or statistical layers, the addition of a trend line and prediction interval increases the probability of a participant selecting the target plot with data generated from M_T, the trend model, and decreases the probability of a participant selecting the target plot with data generated from M_C, the cluster model.

5.2.2.2 Color and Shape Palettes

Colors and shapes used in this study were selected in order to maximize preattentive feature differentiation. Demiralp et al. (2014) provide sets of 10 colors and 10 shapes, with corresponding kernels determined by user studies which indicate perceptual distance. Using these perceptual kernels for shape and color, we identified sets of 3 and 5 colors and shapes which maximize the sum of pairwise differences, subject to certain constraints imposed by software and accessibility concerns.

The color palette used in Demiralp et al. (2014) and shown in figure 5.9 is derived from colors

		Line Emphasis		
	Strength	0	1	2
Cluster Emphasis	0	None	Line	Line + Prediction
	1	Color Shape	Color + Line	
	2	Color + Shape Color + Ellipse		Color + Ellipse + Line + Prediction
	3	Color + Shape + Ellipse		

Table 5.2: Plot aesthetics and statistical layers which impact perception of statistical plots, according to gestalt theory.

available in Tableau visualization software(Hanrahan, 2003). In order to produce experimental stimuli accessible to the approximately 4% of the population with red-green color deficiency (Gegenfurtner and Sharpe, 2001), we removed the gray hue from the palette. This modification produced maximally different color combinations which did not include red-green combinations, while also removing a color (gray) which is difficult to distinguish for those with color deficiency.

Software compatibility issues led us to exclude two shapes used in Demiralp et al. (2014) and shown in figure 5.10. The left and right triangle shapes (available only in unicode within R) were excluded due to size differences between unicode and non-unicode shapes. After optimization over the sum of all pairwise distances, the maximally different shape sequences for the 3 and 5 cluster datasets also conform to the guidelines in Robinson (2003): for $K = 3$ the shapes are from Robinson's group 1, 2, and 9, for $K = 5$ the shapes are from groups 1, 2, 3, 9, and 10. Robinson's groups are designed so that shapes in different groups show differences in preattentive properties; that is, they are easily distinguishable. In addition, all shapes are non-filled shapes, which means that they are consistent with one of the simplest solutions to overplotting of points in the tradition of Tukey (1977); Cleveland (1994) and Few (2009). For this reason we abstained from the additional use of alpha-blending of points to diminish the effect of overplotting in the plots.

5.2.3 Experimental Design

The study is designed hierarchically, as a factorial experiment for combinations of σ_C, σ_T, and K, with three replicates at each parameter combination. The parameters are used to generate lineup datasets which serve as blocks for the plot aesthetic level of the experiment; each dataset is rendered with every combination of aesthetics described in table 5.2. Participants are assigned to evaluate the pre-generated plots according to an augmented balanced incomplete block scheme: each participant is asked to evaluate 10 plots, which consist of one plot at each combination of σ_C and σ_T, randomized across levels of K, with one additional plot providing replication of one level of $\sigma_C \times \sigma_T$. Each of a participant's 10 plots will present a different aesthetic combination; as a result, no participant will see the same dataset twice.

5.2.4 Hypotheses

The primary purpose of this study is to understand how visual aesthetics affect signal detection in the presence of competing signals. We expect that plot modifications which emphasize similarity and proximity, such as color, shape, and 95% bounding ellipses, will increase the probability of detecting the clustering relationship, while plot modifications which emphasize continuity, such as trend lines and prediction intervals, will increase the probability of detecting the linear relationship.

A secondary purpose of the study is to relate signal strength (as determined by dataset parameters σ_C, σ_T, and K) to signal detection in a visualization by a human observer.

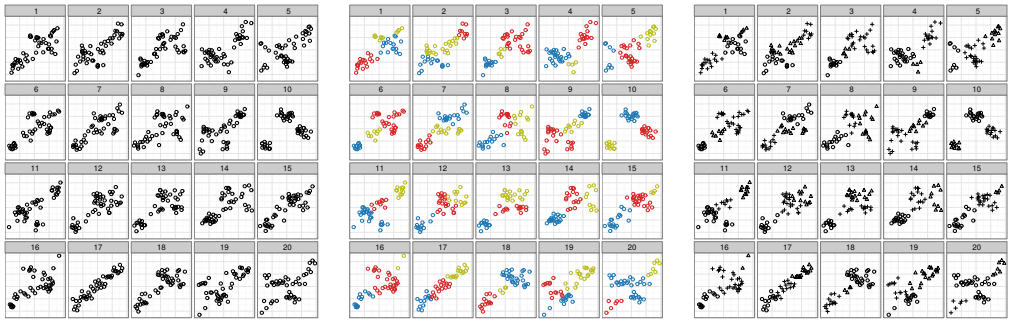
5.2.5 Participant Recruitment

Participants were recruited using Amazon's Mechanical Turk service(Amazon, 2010), which connects interested workers with "Human Intelligence Tasks" (HITs), which are (typically) short tasks which cannot be easily automated. Only workers with at least 100 previous HITs at a 95% successful completion rate were allowed to sign up for completing this task. These restrictions reduce the amount of data cleaning required by ensuring that participants have experience with the Mechanical Turk system and reliably complete accepted HITs.

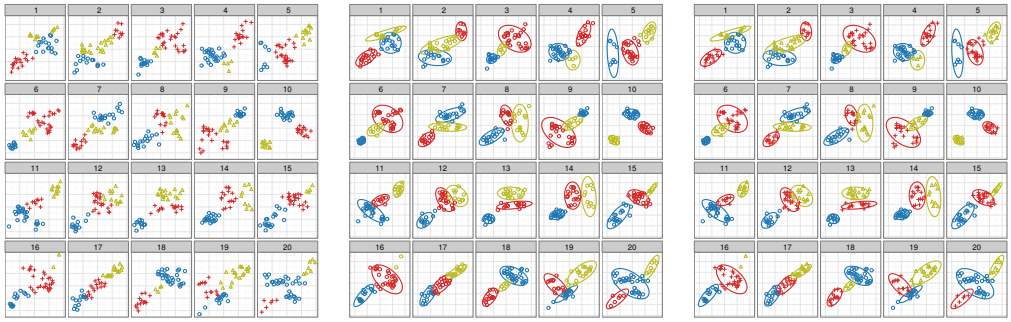
Participants were asked to complete an example task similar to the task in the experiment before deciding whether or not to complete the HIT. The lineups used as examples contained only one target (5 trend and 5 cluster trials were provided), and participants had to correctly identify target plots in at least two lineups before accepting the HIT and proceeding to the experimental phase. The webpage used to collect data from Amazon Turk participants is available at http://www.mlcape.com:8080/mahbub/turk16/index.html. No data was recorded from the example task because participants had not yet provided informed consent.

Once participants completed the example task and provided informed consent, they could accept the HIT through Amazon and were directed to the main experimental task. Participants were required to complete 10 lineups, answering "Which plot is the most different from the others?". Participants were asked to provide a short reason for their choice, such as "Strong linear trend" or "Groups of points", and to rate their confidence in their selection from 1 (least confident) to 5 (most confident). After the first question, basic demographic information was collected: age range, gender, and highest level of education.

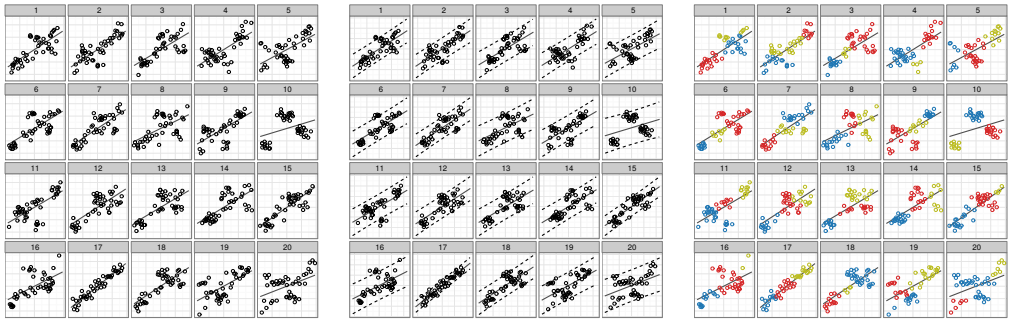
(a) Plain (b) Color (c) Shape



(d) Shape + Color (e) Color + Ellipse (f) Shape + Ellipse



(g) Trend (h) Trend + Error (i) Trend + Color



(j) Trend + Color + Ellipse

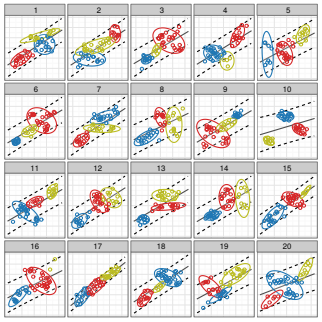


Figure 5.8: Each of the 10 plot feature combinations tested in this study, with $K = 3$, $\sigma_T = 0.25$ and $\sigma_C = 0.20$.

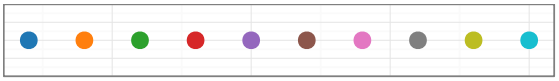


Figure 5.9: Colors in Demiralp et al. (2014). This study removed gray from the palette to make the experiment more inclusive of participants with colorblindness.

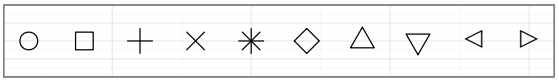


Figure 5.10: Shapes in Demiralp et al. (2014). In order to control for varying point size due to Unicode vs. non-Unicode characters, the last two shapes were removed.

5.3 Results

5.3.1 General results

Data collection was conducted over a 24 hour period, during which time 1356 individuals completed 13519 unique lineup evaluations. Participants who completed fewer than 10 lineups were removed from the study (159 participants, 1060 evaluations), and lineup evaluations in excess of 10 for each participant were also removed from the study (421 evaluations). After these data filtration steps, our data consist of 12010 trials completed by 1201 participants.

Of the participants who completed at least 10 lineup evaluations, 61% were male, relatively younger than the US population and relatively well educated (see figure 5.11). Each plot was evaluated by between 11 and 37 individuals (Mean: 22.24, SD= 4.62). 82.7% of the participant evaluations identified at least one of the two target plots successfully (Trend: 26.6%, Cluster: 56.7%).

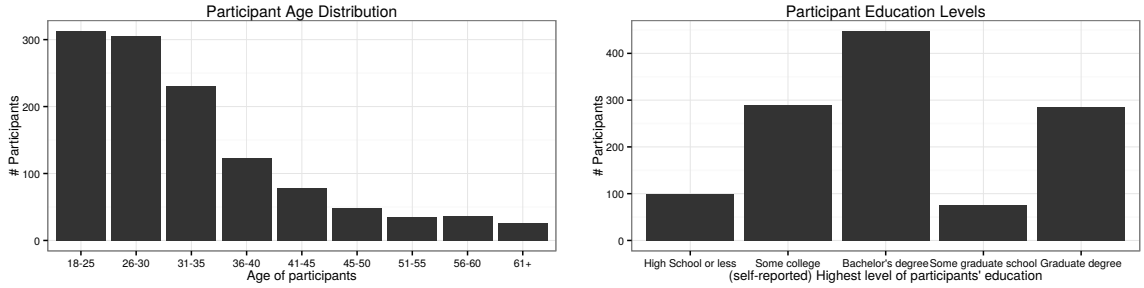


Figure 5.11: Basic demographics of participants.

From figure 5.12 we see that users identified more cluster targets than trend targets; this may be a result of the number of plot types which emphasize clusters over trend: 5 plot types emphasize clustering (according to our hypothesis) and only two emphasize linear trends. Individuals also do not primarily identify one target type over another target type, but generally pick both types over the course of ten lineups.

We first consider the effect of plot aesthetics on target selection for each target type (separately), and then compare the probability of selecting the cluster target compared with the trend target as a function of plot aesthetics. Finally, we will consider data on response times for each type of plot.

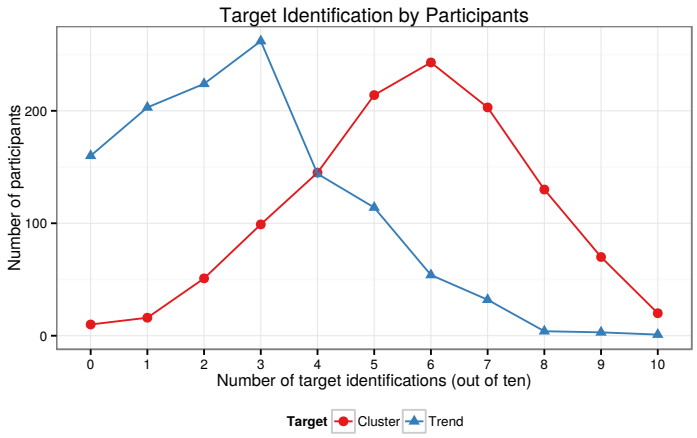


Figure 5.12: Target identifications by users. Users are not generally primed for one target over the other target.

5.3.2 Single Target Plot Models

We model the probability of selecting the target plot using a logistic regression with plot type as a fixed effect, and random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level).

For plot type i, displaying dataset $j = 1, ..., 54$ and participant $k = 1, ..., P$, we model

$$\text{logit } P(\text{target identification}) = \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta + \epsilon, \tag{5.3}$$

where β_i describe the effect of specific plot types

$\gamma_j \overset{iid}{\sim} N\left(0, \sigma^2_{\text{data}}\right)$, the random effect for dataset specific characteristics

$\eta_k \overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right)$, the random effect for participant characteristics

and $\epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2_e\right)$, the error associated with a single trial evaluation

Random effects for dataset and participant are assumed to be orthogonal to one another, as participants only see an individual dataset one time. We note that any variance due to parameters K, σ_T, and σ_C is contained within σ^2_{data} and can be examined using a subsequent model.

5.3.2.1 Trend Target Model

Using equation 5.3, we define success as "the participant correctly identified the trend target plot generated by M_T". Figure 5.13 shows the fixed effects of the resulting model fit. Color, Shape, and Ellipse aesthetics (and combinations thereof) decrease participant recognition of the trend target plot, while the Trend + Error combination increases participant recognition of the trend target plot.

These results are consistent with our hypothesis that aesthetics which emphasize the gestalt similarity heuristic decrease recognition of the trend target plot. The aesthetic combinations of color + shape + ellipse and color + ellipse, which recruit gestalt heuristics for similarity and common region, strongly reduce the probability of detecting the trend target plot. Aesthetic combinations which only activate the gestalt similarity heuristic, such as color, shape, and color+shape, have somewhat less of an effect. As would be predicted by previous studies, such as Lewandowsky and Spence (1989a), color (or color + shape) more strongly detracts from trend target recognition than shape alone.

The trend line aesthetic does not significantly increase trend target plot recognition, either alone or in the conflict condition color + trend. This may be because the gestalt heuristic recruited in this case is continuity ("elements which blend together smoothly likely belong to one unit"), the same heuristic recruited by the points alone. Thus, the trend line may provide only slight additional visual emphasis from the gestalt perspective.

Once error bands are added to the plot, many other heuristics may be applied: closure (perception of a closed object even if it is incomplete), common region, and connectedness (depending on the style of the error bands). For the 95% intervals in the stimuli plots, closure and common region are the most likely heuristics recruited with the addition of error bands (a ribbon-style interval would recruit connectedness as well). The results are consistent with this idea: only the combined closure and continuity emphasis of trend + error bands significantly increases the probability that participants will identify the target plot generated under M_T.

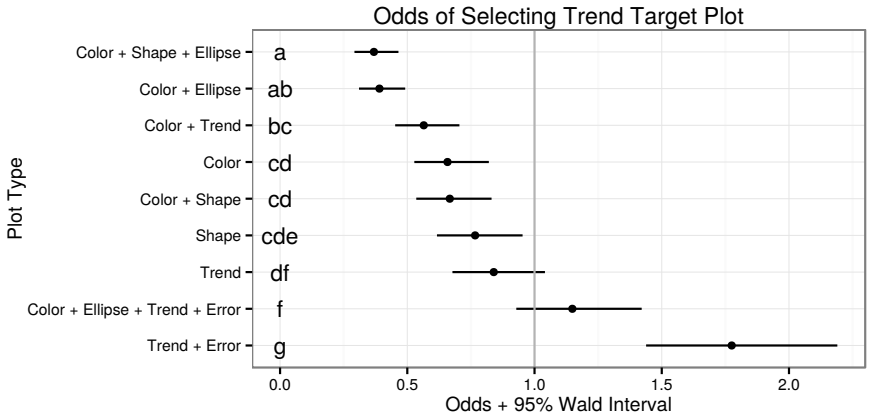


Figure 5.13: Odds of detecting the trend target plot for each aesthetic. Only the combination of Trend + Error significantly increases the odds of trend target plot detection relative to the control plot (plain scatterplot).

5.3.2.2 Cluster Target Selection

We now examine the probability of selecting the cluster target plot as a function of plot type, with random effects for dataset (which encompasses parameter effects) and participant (accounting for variation in individual skill level). The model fit here is the same as that shown in equation (5.3), except that success in this model is defined as identification of the cluster target plot.

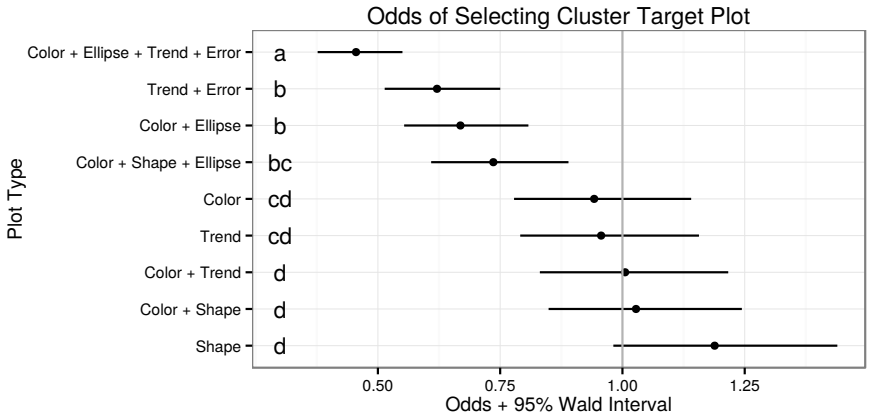


Figure 5.14: Odds of detecting the cluster target plot for each aesthetic, relative to a plain scatterplot. The presence of error lines or bounding ellipses significantly decreases the probability of correct target detection, and no aesthetic successfully increases the probability of correct target detection. This may be due to differences in group size for null plots, with data generated under M_0 compared with the cluster target plot displaying data generated under M_C.

Figure 5.14 contains odds and 95% Wald intervals of the estimated fixed effects obtained by fitting equation 5.3 to a binary indicator of successful cluster target identification. According to the model results, no plot aesthetics significantly increase the likelihood of selecting the cluster target plot; however, several aesthetic combinations decrease this likelihood. Consistent with our hypothesis, Color + Ellipse + Trend + Error and Trend + Error plot aesthetic combinations significantly decrease the detection of the cluster target plot. S However, the implication that Color + Ellipse and Color + Shape + Ellipse also decrease cluster target detection is not consistent with our hypotheses. Examination of participants' reasons for selecting specific target plots provides at least some explanation; participants cited reasons such as "There is no circle highlighting the yellow symbols in this plot" and "Lack of a circle around the red symbols".

This suggests that our group allocation for null target plots may have produced unintentional results; rather than providing unambiguous gestalt cues which reinforced group separation, instead, our null plots provided mixed cues which varied the number of groups and the presence of the additional similarity cue.

In order to confirm this hypothesis numerically, we used simulation (as described in section 5.2.1.4 to examine the distribution of group size (as measured by gini impurity) in order to establish whether there were any systematic differences in group size inequality between data generated from M_0 (null data) and data generated from M_C (cluster data). Figure 5.15 demonstrates that the cluster plots have lower group size differences (e.g. are more equally sized) than null plots at all parameter combinations. It is therefore possible that some participants will identify extraordinarily unequal group sizes present in null plots as significantly different from the other lineup plots, ignoring any cluster signal. Future studies should more tightly control group size in order to reduce this effect.

Numerically, these null data sets did have uneven group allocation; bounding ellipse estimation failed for groups with fewer than 3 points and in these cases, ellipses were not drawn. Visually, the conspicuous absence of an ellipse will lead participants to select null plots with that feature (see section 5.3.6 for a more detailed look at participants' responses).

This effect actually provides some additional information as to the hierarchy of gestalt

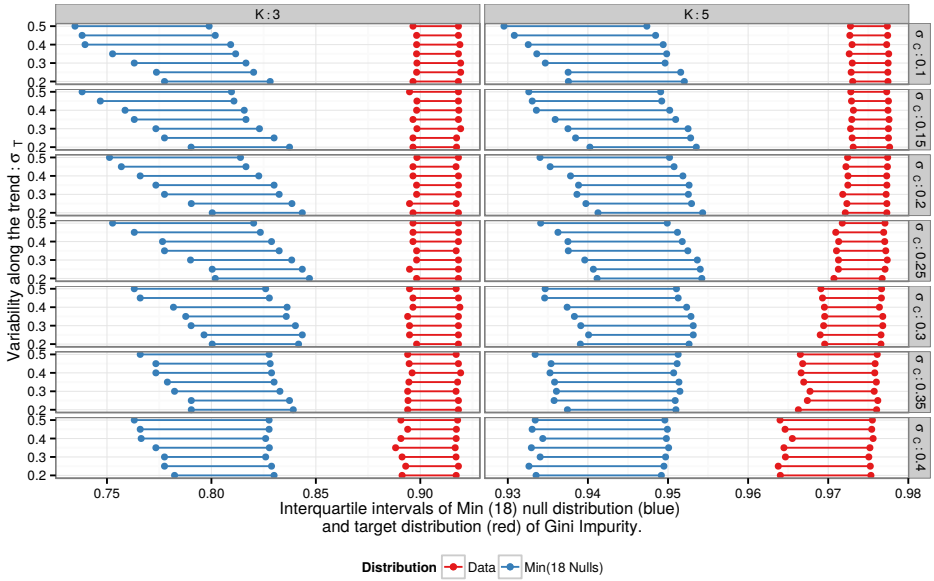


Figure 5.15: Simulated interquartile range of group size inequality statistic values for cluster and null data distributions.

features: for plots displaying the same data (including at least one plot with group size < 3), participants were more likely to identify the cluster target plot under the Color and Shape aesthetics than under Color + Ellipse or Color + Shape + Ellipse conditions. The presence of the ellipse (and the gestalt common region heuristic) dominated the effect of point similarity (albeit not in the way the authors originally intended). In future experiments, it will be advantageous to control the variability in group size in order to remove the conflicting visual influence of gestalt common region heuristics with the greater similarity and proximity present in the target plot.

5.3.3 Face-Off: Trend versus Cluster

Next, we consider only the subset of trials in which participants identified one of the two target plots (9936 trials). For these trials, we compare the probability of selecting the cluster target generated by M_C compared with the probability of selecting the trend target generated by M_T. Overall, participants favored clusters to trends at a ratio of about 2:1. We remove this effect using an intercept, and model cluster vs. trend decisions using a logistic regression with a random effect for each dataset to account for different difficulty levels in the generated data.

The estimated odds of a decision in favor of cluster over trend target are shown in figure 5.16. From left to right the odds of selecting the cluster target over the trend target increase. As hypothesized, the strongest signal for identifying groups, is color + shape + ellipse, while trend + error results in the strongest signal in favor of trends. Most of the effects are not significantly different (see the letter values Piepho (2004) based on Tukey's Post Hoc difference tests on the left hand side of the figure, representing pairwise comparisons of all of the designs, adjusted for multiple comparison). Trend + error plots and color + ellipse + trend + error plots are significantly different from all of the other designs.

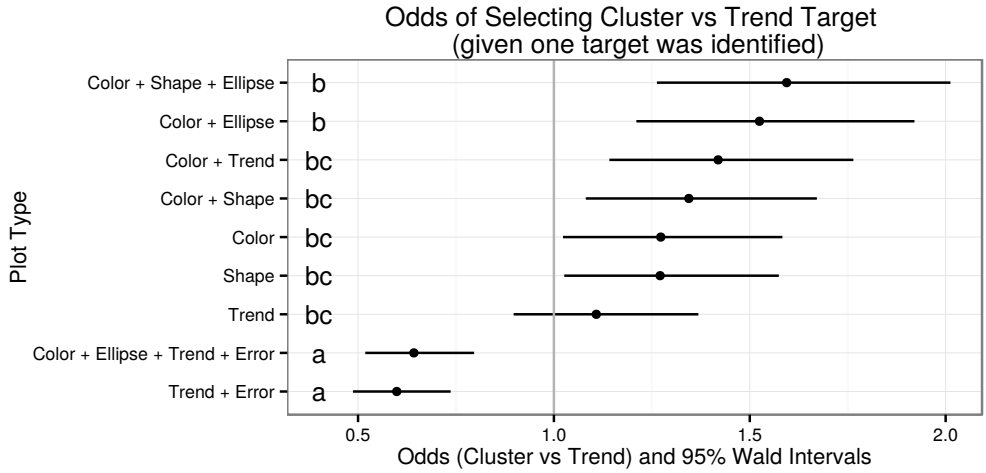


Figure 5.16: Estimated odds of decision for cluster versus trend target based on evaluations that resulted in the identification of one of these targets. Plot types are significantly different if they do not share a letter as given on the left hand side of the plot.

Examining the model results from the perspective of Gestalt heuristics, it is clear that the similarity/proximity effect, as indicated by spatial clustering and aesthetics such as color and shape, dominates the equation, including dominating the color + trend (similarity vs. continuity) condition.

When trend line and error are present in the same plot, additional Gestalt ordering principles are present: common region and possibly closure (due to the enclosed space between the two error lines), in addition to the continuity heuristic present due to the trend line and the linear relationship between x and y. The interaction between these three heuristics dominates the perceptual experience, decreasing the probability that a participant will select the cluster

target plot (and increasing the probability that the trend target will be selected).

This interaction effect explains the different outcomes seen by the two conditions with conflicting aesthetics: the color+trend condition is more likely to result in cluster plot selection, while the color + ellipse + trend + error condition is more likely to result in trend plot selection, because the combined effect of the gestalt heuristics present in the trend + error elements is stronger than the effect of color + ellipse elements, which only invoke Gestalt heuristics of similarity and common region.

5.3.4 Response Time

As data collection was conducted entirely online, we cannot measure responses in the millisecond range characteristic of many psychometric studies, however, the data server does record the time between initial lineup presentation (trial start) and answer submission (trial end). Examining differences in average response times across trials provides us with an additional measure of trial difficulty or perceptual complexity. We can also explore whether participants spent more time on certain types of plots and how additional time is realted to accurate target identification.

We model log-transformed reaction time as a function of evaluation outcome (neither target identified, cluster or trend target identified, or both targets identified) and plot type. In order to remove the "novelty" effect of an unfamiliar task, we also include an indicator variable for the first trial an individual completed (Majumder et al., 2014b). A random effect for participant and dataset is included to account for the experimental design. Figure 5.17 displays the model results for each outcome of the lineup evaluation; simple plots (color, trend, shape) take less time to evaluate (across all conditions) than plots with more than one aesthetic. Additionally, participants who identified the cluster target took less time (in most cases) than participants identifying the trend target.

In a second model, we fit log response time as a function of the plot type, first trial, and the interaction between the outcome and data-generation parameters σ_C, σ_T, and K. In order to model the task as designed, we have coded σ_C and σ_T according to difficulty level - easy, medium, and hard, rather than modeling the numerical parameters themselves; this allows us

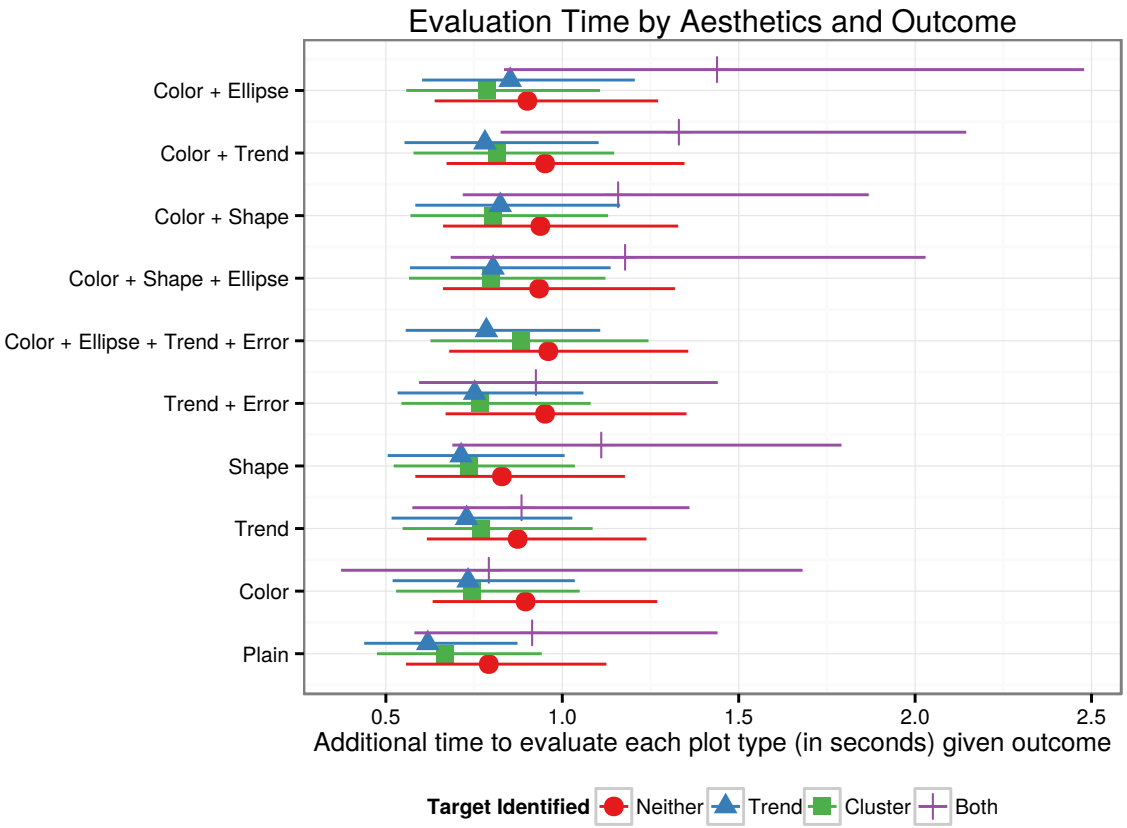


Figure 5.17: Results of a model describing log evaluation time by evaluation outcome and plot type. Participants take less time to evaluate plots with a single aesthetic compared with more complicated plots.

to describe the psychological task rather than the numerical task (and also simplifies the model slightly). Figure 5.18 shows the estimated difference in time as a function of difficulty level and trial outcome, and table 5.3 shows the additional fitted effects and 95% intervals which are not shown graphically. The time to evaluate each plot increases slightly with trend difficulty and cluster difficulty (across trials), conditional on outcome, but the "medium" difficulty trials seem to be somewhat discordant in many cases; in some cases, time to evaluate increases and in others, it decreases. This may be due to a conflict between the trend and cluster target plots: when there is no clear signal numerically, evaluation time increases while participants waver between potential targets.

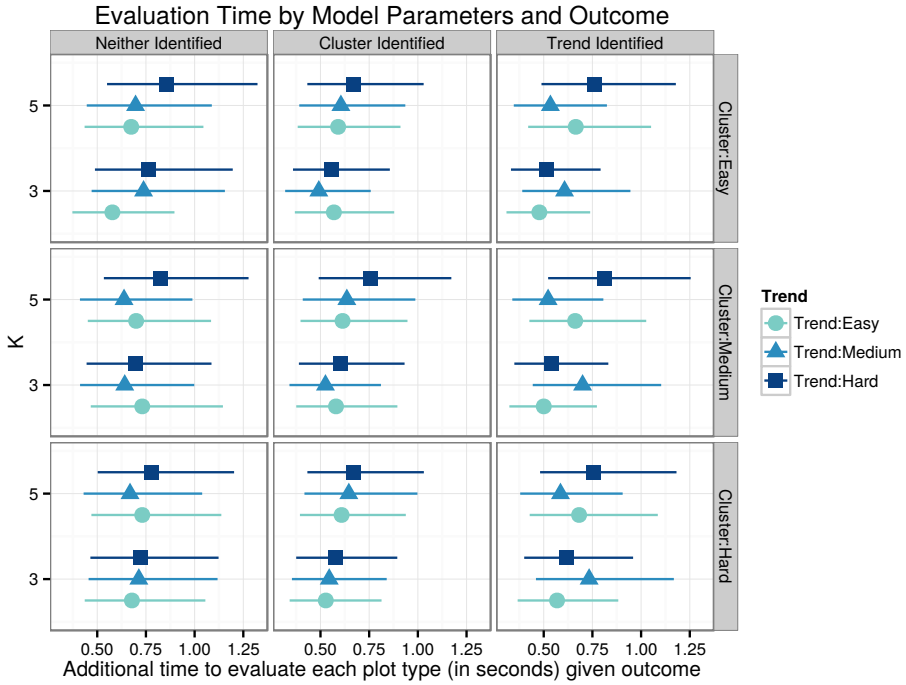


Figure 5.18: Results of a model describing log evaluation time by evaluation outcome and plot type.

5.3.5 Participant Confidence

In addition to participant identification of target plots, we also asked participants to rate their confidence in their answer. Figure 5.19 shows aggregate participant confidence rating as a function of trial outcome. Participants who did not identify either target plot were less likely to be "extremely confident" in their answer, while participants who identified either the trend or the cluster target correctly were highly confident that their answer was correct. Overall, though, participants seem to have some degree of confidence in their answer, regardless of whether the answer was correct.

5.3.6 Participant Reasoning

As part of each trial, participants were asked to provide a short justification of their plot choice. Figure 5.20 gives an overview of summaries of participants' reasoning in form of word clouds. What can be seen is a strong focus in terms of the reasoning depending on the outcome. If the participant chose one of the targets, the reasoning reflects this choice. When neither of

Effect	Estimate	95% Confidence Interval
Intercept	41.617	(27.00, 64.16)
First Trial	1.264	(1.23, 1.30)
Plot Type: Trend	1.153	(1.11, 1.20)
Plot Type: Color	1.139	(1.10, 1.18)
Plot Type: Shape	1.109	(1.07, 1.15)
Plot Type: Color + Shape	1.232	(1.19, 1.28)
Plot Type: Color + Ellipse	1.207	(1.16, 1.25)
Plot Type: Color + Trend	1.227	(1.18, 1.27)
Plot Type: Trend + Error	1.167	(1.12, 1.21)
Plot Type: Color + Shape + Ellipse	1.215	(1.17, 1.26)
Plot Type: Color + Ellipse + Trend + Error	1.274	(1.23, 1.32)

Table 5.3: Exponentiated estimates for parameters describing plot type and first trial effects. We model log evaluation time as a function of plot type, first trial, and data generation parameters, with a random effect for participant. Additional parameters and intervals are shown in figure 5.18.

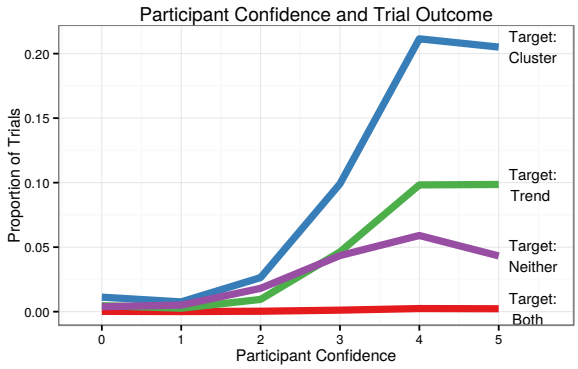


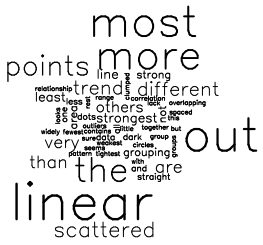
Figure 5.19: Participant confidence levels compared with trial results.

the targets is chosen, there is less focus in the response. The word clouds look surprisingly similar independently of plot type - with the exception of the Ellipse + Color plots: here, the mentioning of specific colors is indicative of participants' distraction from the intended target towards an imbalance of the color/group distribution.

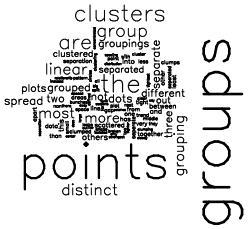
For a more quantitative analysis, responses were categorized based on keywords such as "line(ar)", "correlation", "group", "cluster", "clump", as well as the presence of negation words (non, not, less, etc.). In addition to linear, nonlinear, and group sentiment, many responses focused on the presence of outliers or the amount of variability present in the chosen plot.

The results of this analysis, shown in figure 5.21 and supported by figure 5.20, indicate that

for the most part participants were making decisions based on the criteria we manipulated; rather than alternate visual cues such as group size. In future studies, however, group size should be more tightly controlled to reduce the presence of distractor aesthetics in null plots.

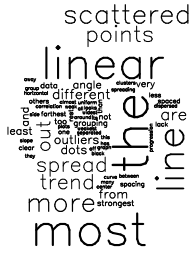


(a) Plain, neither target

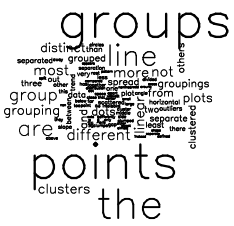


(b) Plain, cluster target

(c) Plain, trend target

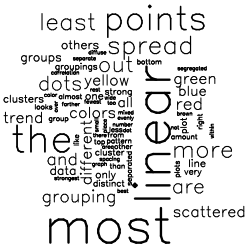


(d) Trend, neither target

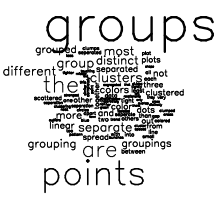


(e) Trend, cluster target

(f) Trend, trend target



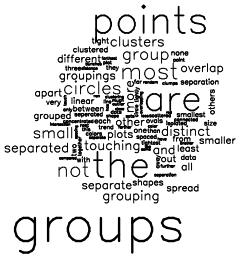
(g) Color, neither target



(h) Color, cluster target

(i) Color, trend target

(j) Color + Ellipse, neither target



(k) Color + Ellipse, cluster target

(l) Color + Ellipse, trend target

Figure 5.20: Wordclouds of participants' reasoning by outcome for a selected number of plot types. Mostly, the reasoning and the choice of the target are highly associated. For the Color + Ellipse plot, participants were distracted from either target by an imbalance in the group/color distribution, as can be seen from the reasoning in the bottom left wordcloud.

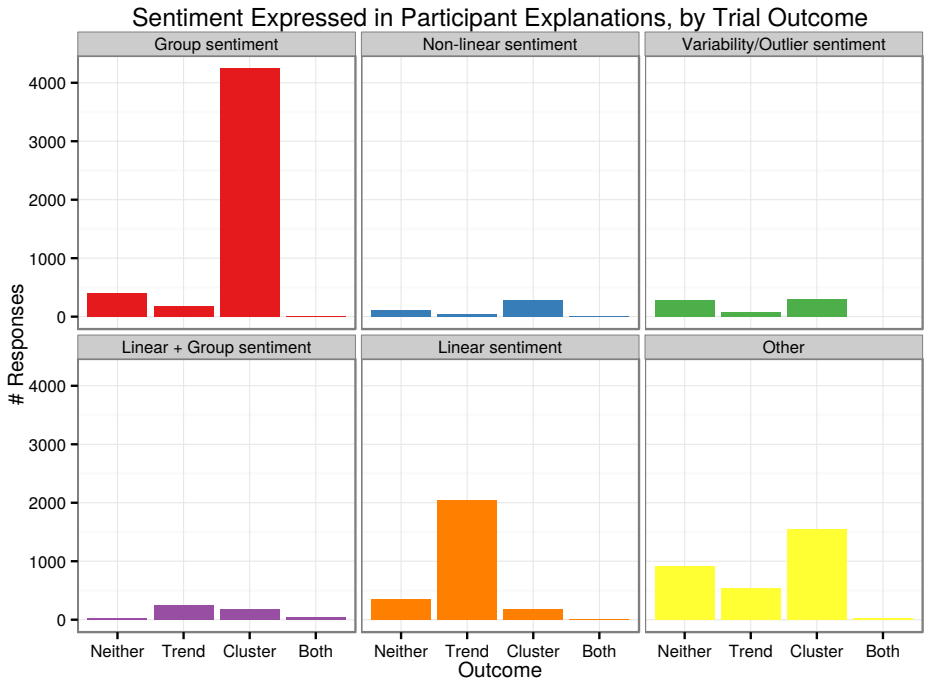


Figure 5.21: Lexical analysis of participants' justification of plot selection. Group sentiment in the reasoning is highly associated with selection of the cluster target plot; linear sentiment is highly associated with selection of the trend target plot.

5.4 Discussion and Conclusions

Taken together, the results presented suggest that plot aesthetics influence the perception of the dominant effect in ambiguous data displays. This effect is not simply additive (otherwise, the two conflicting aesthetic conditions would result in similarly neutral effects); rather, the effect is consistent with layering of gestalt perceptual heuristics. Plot layers which add additional heuristics show larger effects than plot layers which duplicate heuristics which are already in play. For example, adding ellipses to a plot which has color aesthetics increases group recognition by recruiting the closure heuristic in addition to the point similarity heuristic recruited by color; adding shape to a plot which has color aesthetics may increases group recognition slightly, but does not add additional gestalt heuristics (though point similarity is emphasized through two different mechanisms).

In order to explicitly rank aesthetics given this nonadditive mechanism, it would be necessary to test ellipse and error band aesthetics alone; in this study, we have only examined those aesthetics in combination with color and regression line plot layers, as the bounding aesthetics are seldom seen alone.

The results of this study also demonstrate the strength of the lineup protocol as a tool for evaluating data displays. The combination of empirical results and participants' written responses allows researchers to examine the manipulated variablesas well as any alternative hypotheses participants may have utilized, such as group size inequality instead of cluster cohesion.

While further studies are necessary to control for the effects of group size as well as to explore the gestalt heuristics applicable to other types of plots, these results demonstrate the importance of carefully constructing graphs in order to consistently convey the most important aspects of the displayed data.

BIBLIOGRAPHY

Ahluwalia, A. (1978). An intra-cultural investigation of susceptibility to "perspective" and "non-perspective" spatial illusions. British Journal of Psychology, 69:233–241.

Amazon (2010). Mechanical Turk. https://www.mturk.com/mturk/welcome.

Amer, T. (2005). Bias due to visual illusion in the graphical presentation of accounting information. Journal of Information Systems, 19(1):1–18.

Anderson, K. J. and Revelle, W. (1983). The interactive effects of caffeine, impulsivity and task demands on a visual search task. Personality and Individual Differences, 4(2):127–134.

Babcock, J. S. and Pelz, J. B. (2004). Building a lightweight eyetracking headgear. In Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, ETRA '04, pages 109–114, New York, NY, USA. ACM.

Baddeley, A. D. and Hitch, G. (1974). Working memory. Psychology of learning and motivation, 8:47–89.

Bobko, P. and Karren, R. (1979). The perception of pearson product moment correlations from bivariate scatterplots. Personnel Psychology, 32(2):313–325.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4361–4383.

Byron, L. and Wattenberg, M. (2008). Stacked graphs – geometry and aesthetics. IEEE Trans. Vis. Comput. Graph., 14(6):1245–1252.

Carpenter, P. A. and Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. Journal of Experimental Psychology: Applied, 4(2):75.

Carswell, M. C. and Wickens, C. D. (1987). Information integration and the object display an interaction of task demands and display superiority. Ergonomics, 30(3):511–527.

Chowdhury, N. R., Cook, D., Hofmann, H., Majumder, M., and Zhao, Y. (2014). Utilizing distance metrics on lineups to examine what people read from data plots.

Cleveland, W. S. (1994). The Elements of Graphing Data. Hobart Press, 1 edition.

Cleveland, W. S., McGill, M. E., and Robert, M. (1988). The Shape Parameter of a Two-Variable Graph. Journal of the American Statistical Association, 83(402):289–300.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387):pp. 531–554.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. Science, 229(4716):828–833.

Conti, G., Ahamad, M., and Stasko, J. (2005). Attacking information visualization system usability overloading and deceiving the human. In Proceedings of the 2005 symposium on Usable privacy and security, pages 89–100. ACM.

Coren, S. and Porac, C. (1983). Subjective contours and apparent depth: A direct test. Attention, Perception, & Psychophysics, 33(2):197–200.

Day, R. H. and Stecher, E. J. (1991). Sine of an illusion. Perception, 20:49–55.

Demiralp, c., Bernstein, M., and Heer, J. (2014). Learning perceptual kernels for visualization design. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis).

DeMita, M. A., Johnson, J. H., and Hansen, K. E. (1981). The validity of a computerized visual searching task as an indicator of brain damage. Behavior Research Methods & Instrumentation, 13(4):592–594.

Diamond, J. and Evans, W. (1973). The correction for guessing. Review of Educational Research, pages 181–191.

EIA (2014a). Gasoline and diesel fuel update. http://www.eia.gov/petroleum/gasdiesel/reformulated_map.cfm. [Online; accessed 28-Feb-2014].

EIA (2014b). Weekly retail gasoline and diesel prices. http://www.eia.gov/dnav/pet/pet_pri_gnd_dcus_nus_w.htm. [Online; accessed 28-Feb-2014].

Ekstrom, R. B., French, J. W., Harman, H. H., and Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Environmental Protection Agency (2011). Air quality data. http://www.epa.gov/airdata/ad_data_daily.html. [Online; accessed 25-Oct-2013].

Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press, Burlingame, CA, 1 edition.

Fisher, G. H. (1970). An experimental and theoretical appraisal of the perspective and size-constancy theories of illusions. The Quarterly journal of experimental psychology, 22(4):631–652.

French, J. W., Ekstrom, R. B., and Price, L. A. (1963). Kit of reference tests for cognitive factors. Educational Testing Service, Princeton, NJ.

Gattis, M. and Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(1):231.

Gegenfurtner, K. R. and Sharpe, L. T. (2001). Color vision: From genes to perception. Cambridge University Press.

Goldstein, E. B. (2009a). Encyclopedia of perception. Sage Publications.

Goldstein, E. B. (2009b). Sensation and Perception. Thomson Wadsworth, Belmont, CA, eighth edition.

Goldstein, G., Welch, R. B., Rennick, P. M., and Shelly, C. H. (1973). The validity of a visual searching task as an indicator of brain damage. Journal of consulting and clinical psychology, 41(3):434.

Green, R. and Hoyle, E. (1963). The poggendorff illusion as a constancy phenomenon. Nature, 200:611–612.

Gregory, R. (1968). Perceptual illusions and brain models. Proc. Roy. Soc. B, 171:279–296.

Gregory, R. L. (1963). Distortion of visual space as inappropriate constancy scaling. Nature, 199(678-91):1.

Gregory, R. L. (1997). Knowledge in perception and illusion. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 352(1358):1121–1127.

Hampson, E. (1990). Variations in sex-related cognitive abilities across the menstrual cycle. Brain and cognition, 14(1):26–43.

Hanrahan, P. (2003). Tableau software white paper - visual thinking for business intelligence. Tableau Software, Seattle, WA.

Hausmann, M., Slabbekoorn, D., Van Goozen, S. H., Cohen-Kettenis, P. T., and Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. Behavioral neuroscience, 114(6):1245.

Healey, C. G. and Enns, J. T. (1999). Large datasets at a glance: Combining textures and colors in scientific visualization. Visualization and Computer Graphics, IEEE Transactions on, 5(2):145–167.

Healey, C. G. and Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. Visualization and Computer Graphics, IEEE Transactions on, 18(7):1170–1188.

Helander, M. G., Landauer, T. K., and Prabhu, P. V. (1997). Handbook of human-computer interaction. Elsevier.

Henson, D. B. and Williams, D. E. (1980). Depth perception in strabismus. British Journal of Ophthalmology, 64(5):349–353.

Hering, E. (1861). Beiträge zur Physiologie. Engelmann, Leipzig.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012). Graphical tests for power comparison of competing designs. Visualization and Computer Graphics, IEEE Transactions on, 18(12):2441–2448.

Hofmann, H. and Vendettuoli, M. (2013). Common Angle Plots as perception-true visualizations of categorical associations. IEEE Transactions on Visualization and Computer Graphics, 19(12):2297–2305.

Hollands, J. and Spence, I. (1998). Judging proportion with graphs: The summation model. Applied Cognitive Psychology, 12(2):173–190.

Holopigian, K., Blake, R., and Greenwald, M. J. (1986). Selective losses in binocular vision in anisometropic amblyopes. Vision research, 26(4):621–630.

Howe, C. Q. and Purves, D. (2005). Natural-scene geometry predicts the perception of angles and line orientation. Proceedings of the National Academy of Sciences of the United States of America, 102(4):1228–1233.

Hubel, D. H. and Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. The Journal of physiology, 206(2):419–436.

Kosara, R. and Ziemkiewicz, C. (2010). Do mechanical turks dream of square pie charts? In Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV), pages 373–382. ACM Press.

Kosslyn, S. M. (1994). Elements of Graph Design. W.H. Freeman and Company.

Kosslyn, S. M., Ball, T. M., and Reiser, B. J. (1978). Visual images preserve metric spatial information: evidence from studies of image scanning. Journal of experimental psychology: Human perception and performance, 4(1):47.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive science, 11(1):65–100.

Lewandowsky, S., Herrmann, D. J., Behrens, J. T., Li, S.-C., Pickle, L., and Jobe, J. B. (1993). Perception of clusters in statistical maps. Applied Cognitive Psychology, 7(6):533–551.

Lewandowsky, S. and Spence, I. (1989a). Discriminating strata in scatterplots. Journal of the American Statistical Association, 84(407):682–688.

Lewandowsky, S. and Spence, I. (1989b). The perception of statistical graphs. Sociological Methods & Research, 18(2-3):200–242.

Light, A. and Bartlein, P. J. (2004). The end of the rainbow? color schemes for improved data graphics. Eos, Transactions American Geophysical Union, 85(40):385–391.

Lowrie, T. and Diezmann, C. M. (2007). Solving graphics problems: Student performance in junior grades. The Journal of Educational Research, 100(6):369–378.

Loy, A., Follett, L., and Hofmann, H. (2015). Variations of q-q plots – the power of our eyes! The American Statistician, tentatively accepted:???–???

Majumder, M., Hofmann, H., and Cook, D. (2013). Validation of visual statistical inference, applied to linear models. Journal of the American Statistical Association, 108(503):942–956.

Majumder, M., Hofmann, H., and Cook, D. (2014a). Human factors influencing visual statistical inference.

Majumder, M., Hofmann, H., and Cook, D. (2014b). Human factors influencing visual statistical inference. arXiv.org.

Matlin, M. W. (2005). Cognition. Wiley, 6th edition.

Mayer, R. E. and Sims, V. K. (1994). For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning. Journal of educational psychology, 86(3):389.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2):81.

Moerland, M., Aldenkamp, A., and Alpherts, W. (1986). A neuropsychological test battery for the apple ll-e. International journal of man-machine studies, 25(4):453–467.

Morgan, M. (1999). The poggendorff illusion: a bias in the estimation of the orientation of virtual lines by second-stage filters. Vision Research, 39(14):2361 – 2380.

Morgan, M., Adam, A., and Mollon, J. (1992). Dichromats detect colour-camouflaged objects that are not detected by trichromats. Proceedings of the Royal Society of London. Series B: Biological Sciences, 248(1323):291–295.

Mosteller, F., Siegel, A. F., Trapido, E., and Youtz, C. (1981). Eye fitting straight lines. The American Statistician, 35(3):150–152.

National Climate Data Center (2011). Quality controlled local climatological data. http://cdo.ncdc.noaa.gov/qclcd_ascii/. [Online; accessed 25-Oct-2013].

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. Nature Reviews Neuroscience, 8(5):379–391.

Penrose, L. S. and Penrose, R. (1958). Impossible objects: A special type of visual illusion. British Journal of Psychology, 49(1):31–33.

Piepho, H.-P. (2004). An algorithm for a letter-based representation of all-pairwise comparisons. Journal of Computational and Graphical Statistics, 13(2):456–466.

Playfair, W. (1786). Commercial and Political Atlas. Cambridge, London.

Playfair, W., Wainer, H., and Spence, I. (2005). Playfair's Commercial and Political Atlas and Statistical Breviary. Cambridge University Press.

Poulton, E. (1985). Geometric illusions in reading graphs. Perception & psychophysics, 37(6):543–548.

Ratwani, R. M., Trafton, J. G., and Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. Journal of Experimental Psychology: Applied, 14(1):36.

Robbins, N. (2005). Creating More Effective Graphs. Wiley.

Robinson, H. (2003). Usability of scatter plot symbols. ASA Statistical Computing & Graphics Newsletter, 14(1):9–14.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., and Zhao, Y. (2014). Utilizing distance metrics on lineups to examine what people read from data plots. arXiv.org.

RStudio Inc. (2013). shiny: Web Application Framework for R. R package version 0.6.0.99.

Scaife, M. and Rogers, Y. (1996). External cognition: how do graphical representations work? International journal of human-computer studies, 45(2):185–213.

Schaie, K. W., Maitland, S. B., Willis, S. L., and Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. Psychology and aging, 13(1):8.

Schonlau, M. (2003). Visualizing categorical data arising in the health sciences using hammock plots. In Proceedings of the Section on Statistical Graphics (JSM '03). American Statistical Association.

Seckel, A. (2007). Masters of Deception: Escher, Dal & the Artists of Optical Illusion. Sterling.

Shah, P. and Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. Journal of Experimental Psychology: General, 124(1):43.

Shah, P. and Miyake, A. (2005). The Cambridge handbook of visuospatial thinking. Cambridge University Press.

Shepard, S. and Metzler, D. (1988). Mental rotation: Effects of dimensionality of objects and type of task. Journal of Experimental Psychology: Human Perception and Performance, 14(1):3.

Silva, S., Sousa Santos, B., and Madeira, J. (2011). Using color in visualization: A survey. Computers & Graphics, 35(2):320–333.

Simkin, D. and Hastie, R. (1987). An information-processing analysis of graph perception. Journal of the American Statistical Association, 82(398):454–465.

Spence, I. (1990). Visual psychophysics of simple graphical elements. Journal of Experimental Psychology: Human Perception and Performance, 16(4):683–692.

Spence, I. and Garrison, R. F. (1993). A remarkable scatterplot. The American Statistician, 47(1):12–19.

Sun, J. Z., Wang, G. I., Goyal, V. K., and Varshney, L. R. (2012). A framework for bayesian optimality of psychophysical laws. Journal of Mathematical Psychology, 56(6):495–501.

Tan, J. K. (1994). Human processing of two-dimensional graphics: Information-volume concepts and effects in graph-task fit anchoring frameworks. International Journal of Human-Computer Interaction, 6(4):414–456.

Treinish, L. A. and Rogowitz, B. E. (2009). Why should engineers and scientists be worried about color?

Treisman, A. (1985). Preattentive processing in vision. Computer Vision, Graphics, and Image Processing, 31(2):156 – 177.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1):97–136.

Trickett, S. B. and Trafton, J. G. (2006). Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In Diagrammatic representation and inference, pages 286–300. Springer.

Tufte, E. (1991). The Visual Display of Quantitative Information. Graphics Press, USA, 2 edition.

Tukey, J. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.

Tversky, B. and Schiano, D. J. (1989). Perceptual and conceptual factors in distortions in memory for graphs and maps. Journal of Experimental Psychology: General, 118(4):387.

VanderPlas, S. and Hofmann, H. (2014). Signs of the sine illusion - why we need to care. Journal of Computational and Graphical Statistics.

Vanderplas, S. and Hofmann, H. (in press). Spatial reasoning and data displays. IEEE Transactions on Visualization and Computer Graphics.

Varshney, L. R. and Sun, J. Z. (2013). Why do we perceive logarithmically? Significance, 10(1):28–31.

Vekiri, I. (2002). What is the value of graphical displays in learning? Educational Psychology Review, 14(3):261–312.

Voyer, D., Voyer, S., and Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological bulletin, 117(2):250.

Wainer, H. (2000). Visual Revelations. Psychology Press.

Ward, L. M., Porac, C., Coren, S., and Girgus, J. S. (1977). The case of misapplied constancy scaling: Depth associations elicited by illusion configurations. American Journal of Psychology, 90(4):609–620.

Weintraub, D. J., Krantz, D. H., and Olson, T. P. (1980). The poggendorf illusion: Consider all the angles. Journal of Experimental Psychology: Human Perception and Performance, 6:718–725.

Westheimer, G. (2007). Irradiation, border location, and the shifted-chessboard pattern. Perception, 36(4):483.

Westheimer, G. (2008). Illusions in the spatial sense of the eye: Geometrical-optical illusions and the neural representation of space. Vision Research, 48(20):2128–2142.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.

Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1):3–28.

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

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical inference for infovis. Visualization and Computer Graphics, IEEE Transactions on, 16(6):973–979.

Wilkinson, L., Wills, D., Rope, D., Norton, A., and Dubbs, R. (2006). The grammar of graphics. Springer.

Wolfe, J., Kluender, K., and Levi, D. (2012). Sensation and Perception. Sinauer Associates, Incorporated, 3 edition.

Zhang, J. (1997). The nature of external representations in problem solving. Cognitive science, 21(2):179–217.

Zhang, J. and Norman, D. A. (1994). Representations in distributed cognitive tasks. Cognitive science, 18(1):87–122.

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., and Chowdhury, N. R. (2013). Mind reading: Using an eye-tracker to see how people are looking at lineups. International Journal of Intelligent Technologies & Applied Statistics, 6(4).