NORTHWESTERN UNIVERSITY

Perceptual and Cognitive Affordances of Data Visualizations

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By

Cindy Ya Yang Xiong

EVANSTON, ILLINOIS

June 2021

ABSTRACT

Perceptual and Cognitive Affordances of Data Visualizations

Cindy Ya Yang Xiong

Visualization is a powerful tool to help people better communicate and understand data. But the perception and interpretation of data is riddled with bias. Data visualizations are ambiguous objects such that different people viewing the same visualization can come to different conclusions. The ambiguity is likely associated with people's perceptual biases when viewing visualizations, and that their viewing behaviour can be heavily influenced by their background knowledge and experiences. In addition to the ambiguous nature of visualizations, designing a visualization is a difficult task. This process involves multiple design decisions, and each design decision can dictate viewer takeaways from the visualization. In my dissertation, I share several empirical studies investigating how visualization design could influence viewer perception and interpretation of the same data, referencing methods and insights from psychology and computer sciences. From these studies, I extract usable design guidelines that could help future researchers and practitioners design more effective visualizations to communicate data.

Acknowledgements

I would like to recognize everyone who has helped me throughout this journey. First and foremost, I would like to thank my partner and my biggest supporter, Akira Wada. He was there when I submitted my first paper and struggled to format the references in my Word document. He was finishing his MS in Materials Engineering in California at that time, and he flew over hoping to spend quality time with me. I'm not sure if 'helping me label all my references for a VIS submission' counted as spending quality time, but I deeply appreciate him staying positive and encouraging throughout the whole process. Even today,even after Iâve upgraded myself to 'VIS veteran' status, he is still here, helping me put out multiple fires. Looking ahead, moving to yet another new city will have its challenges, but I'm confident that I can do anything with Akira by my side. He is the best teammate one could ever ask for.

I also want to thank all my mentors who have taught me valuable lessons. My advisor, Professor Steve Franconeri, thank you for your constant support, encouragement, and for generously sharing your knowledge, expertise, and network with me. I vividly remember how nervous I was before giving my first conference presentation, and Steve sat right next to me at the conference hotel lobby in Arizona providing me one last set of feedback on my slides, despite just having gotten off a long, exhausting flight. I also want to thank my dissertation committee members, Professor Satoru Suzuki and Professor David Rapp. Satoru's office sits right upstairs to my lab. One of my favourite things to do when we still hang out in person was to run upstairs (often uninvited) and talk to Satoru to hear his thoughts on my analysis or project ideas. I don't think I've ever walked out of his lab space without feeling motivated or inspired. Satoru is also a stats expert/magician. I remember him very patiently walking me through my first statistical analysis in my first paper — it was a Wilcoxon signed-rank test. I've come a long way since then, but I'll always remember that moment and I strive to be a mentor to my future students with the same wisdom and patience. David's office is a little further away, so I couldn't just barge into his office uninvited (lucky him). Going to his office to pick his brain has always been a treat. Not only because David is extremely smart and our brainstorming sessions always end up with us taking a novel perspective on the project, but also because sometimes I get to walk away with a fun board game. Now to think about it, before I even met him, I've heard people mentioning that he has the third largest board game collection in all of Illinois. David is also a great writer, and his comments and edits on my papers always elevates my writing.

Although not on my dissertation committee, I would also like to thank Professor Jessica Hullman, Professor Casimir Ludwig, and Professor Neal Roese for their time, their support, and their valuable insights for when I was on the job market. Jessica and Cas, I've learned so much from you both in our research collaborations and I say the experience has made me a better scientist. I am especially in awe of how critically you both think about experimental design and how you meticulously build up an intriguing narrative around the project. Moving forward, I'm going to carry on with the experimental rigor and writing clarity I inherited from you both, and hopefully pass it to my own students as well.

At this point I've started to realize the acknowledgement is getting a little long. So I'm going to quickly thank all my lab mates, Cristina Ceja, Christie Nothelfer, Dian Yu, Evan Anderson, Nicole Jardine, Caitlyn McColeman, Elsie Lee (and Taylor Robbins), Steve Haroz, Miriam Novack, Hauke Meyerhoff, and Mady Awad, as well as all the fantastic research assistants that have helped me with my projects, Chase Stokes, Jun Lee, Kylie Lin, Madi Trycha, Meghna Jain, Rachel Shapiro, Gabriel Belkind-Gerson, Emily Manheimer, and Chase Bednarz. I also want to thank my cohort, and especially the Chili's gang, Phoebe Lam, Teresa Vargas, Juston Osborne (and Emily Willroth), James Wages, Cassie Brandes, Makeda Austin, and Alissa Baker-Oglesbee. And I'll never forget the red lobster incident. My buddies in the department, Larry Cheng and Cat Han (I hope I didn't scare you too much by fainting in class that day), Nikita Salovich, Arielle Elliott, Melisa Menceloglu, and Max Smith. Oh let's not forget the honorary department members Emily Hittner and Nicole Mirea. Also thank you to my design cluster cohort who are just absolutely amazing, Kristine Lu, Eureka Foong, Ryan Louie, Eureka Foong, Diego Gómez-Zará, and Garrett Hedman. Not to mention the friends Cristina and I made on that journey to Couches (fearing for our kidney), my collaborators, and everyone who has hosted me when I visited their universities/labs to give talks when I was still a rookie (well, I still am, but a more mature rookie now). Last but not least, I want to thank my friends and mentors at UCLA, my friends from Burnaby (Nice shoes, talking about you), my team from when I interned at the MathWorks, my future colleagues at UMass Amherst, my family, and the bubble tea store owners and staff that have helped fuel me through some rough evenings.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Across science, education, healthcare, and public discourse, we rely heavily on data to understand, communicate, and make decisions. By tapping the power of our visual processing system, which can crunch vast arrays of numbers at a glance and provide us with critical values, statistics, and patterns needed to interpret the world around us, visualized data — maps, infographics, flow charts, word clouds, and network diagrams — can massively enhance our ability to analyze and understand patterns in order to make data-driven decisions.

Well-designed data visualizations can lead to powerful and intuitive processing by a viewer, both for visual analytics and data storytelling. Designing a good visualization is difficult. It requires multiple forms of expertise, weeks of training, and years of practice. Even after this, designers still require ideation and several critique cycles before creating an effective visualization. When badly chosen, that visualization leaves important patterns opaque, misunderstood, or misrepresented.

How can we create well-designed data visualizations to more effectively communicate data? I have taken a ‘reverse engineering’ approach where I examine what people tend to see and take-away from a visualization through observations, interviews, and controlled lab studies. Referencing methods from perceptual psychology, behavioural economics, and computer science, I have established a research program that focuses on creating a mapping between some visualization design elements and the type of interpretations they tend to elicit. This mapping help us understand how viewers perceive, interpret, and make decisions from visualizations, and it can help designers and researchers more effectively communicate their data.

Understanding how people perceive a visualization and use it to make decisions can benefit researchers and practitioners from a multiple disciplines. For example, researchers in public policy communication or political science could use this knowledge to create trustworthy and persuasive visualizations to aid decision making (Nyhan and Reifler, 2019). Education researchers might use this information to help students form better mental representations of difficult topics, such as teaching students water cycles with visual diagrams (Márquez et al., 2006). In healthcare, this knowledge can help doctors better communicate the risks of medical procedures to patients (Ancker et al., 2006), and in politics, journalists can better communicate uncertainties in election outcomes or hurricane paths (Ruginski et al., 2016).

I've been exploring the perceptual and cognitive affordances of visualized data through empirical studies. Affordances are the relationships between the properties of an object that convey potential interactions and a user's capabilities (Norman, 2013). An example would be that a door with a handle can be pulled opened, whereas a door with just a metal plate surface can *only* be pushed open. Similar to physical objects, visual data representations also appear to hold similar affordances, via common tasks done with those designs, or conceptual associations driven by their underlying metaphors (Tversky, 2014).

In the past five years, I realize that designing a good visualization is difficult because, unlike physical objects, people's interpretations of visualized data deeply depend on its design. Creating a visualization involves a series of design decisions, ranging from the type of visualization chosen, how data is aggregated, the color scheme chosen for the visualization, to how visual marks that the visualization is composed of is spatially arranged. Each of these design decisions may lead to a different interpretation from the viewer.

I present the work I've done to date exploring the connections between the design decisions and the viewer interpretation. In Chapter 1, I demonstrate that visualized data is an intriguing but difficult target of study as its interpretation is ambiguous, unlike non-visualizations where the data is presented as raw numbers. Different people looking at the same visualization can come to different conclusions. In Chapter 2 and 3, I demonstrate that the ambiguous nature of visualizations is likely associated with the fact that people are perceptually biased when they view visualizations, and how we look at visualizations is deeply influenced by our background knowledge and expertise. These three chapters justify why it's critical for visualization researchers to investigate how the design of visualizations impacts what viewers see in a visualization, which data values they compare, how they reason with the information available, and what decisions they make with it. In Chapter 4 and 5, I share some initial investigations into how visualization designs could influence viewer interpretations and generate design guidelines that might help future designers create more effective visualizations. Chapter 4 examines how people think about correlation and causation with data and demonstrate that aggregated bar charts tend to elicit incorrect causal thinking. Chapter 5 looks into how spatial arrangement and color mapping of data values can impact viewer takeaways.

This space investigating how visualization design can impact viewer interpretations and decision making is incredibly rich. It's an interdisciplinary effort with practical implications. These design guidelines can be incorporated into visualization tools or be utilized by data journalists or data scientists in the future to generate visualizations that best communicate the designers' intended message and help ensure the viewer sees the 'right' story.

CHAPTER 2

Same Data, Different Decisions

You'd hope that people would interpret data objectively, but that process is riddled with biases (Kahan et al., 2017). Ambiguous figures like the duck-rabbit and Necker cube illusions reveal that our brain can lock into a single view of a multi-stable percept Attneave (1971). The duck-rabbit ambiguity is rare in the real, visual world, but it is ubiquitous in the artificial world of information and affects how we interpret representations of information. In this first Chapter, we demonstrate that visualizations can be similar to the duck-rabbit illusion such that two people looking at the same visualization can "see" different patterns – one the duck and the other the rabbit, and make different conclusions. Imagine reading a story in the New York Times Upshot with a visualization depicting the unemployment rate during the Obama administration from â09 to â12, as shown in Figure 2.1. Democrats might focus on the steadily decreasing trend in the figure and see it as a validating accomplishment of the administration. Republicans might instead focus on the large difference between the 8% goal level and the actual unemployment rate, leading to frustration with the administration (Bostock et al., 2012).

When reading a visualization, viewers need to exercise top-down attentional control to extract a series of relationships and patterns from the data values (Szafir et al., 2016; Egeth et al., 2010; Michal et al., 2016; Michal and Franconeri, 2017). A dataset can contain many patterns to perceive. Extracting patterns and relations from a dataset is more similar to reading sentences in a paragraph, where the reader has to sequentially process each sentence, rather than looking at a picture or scene, in which elements are processed simultaneously to create a gist (Shah

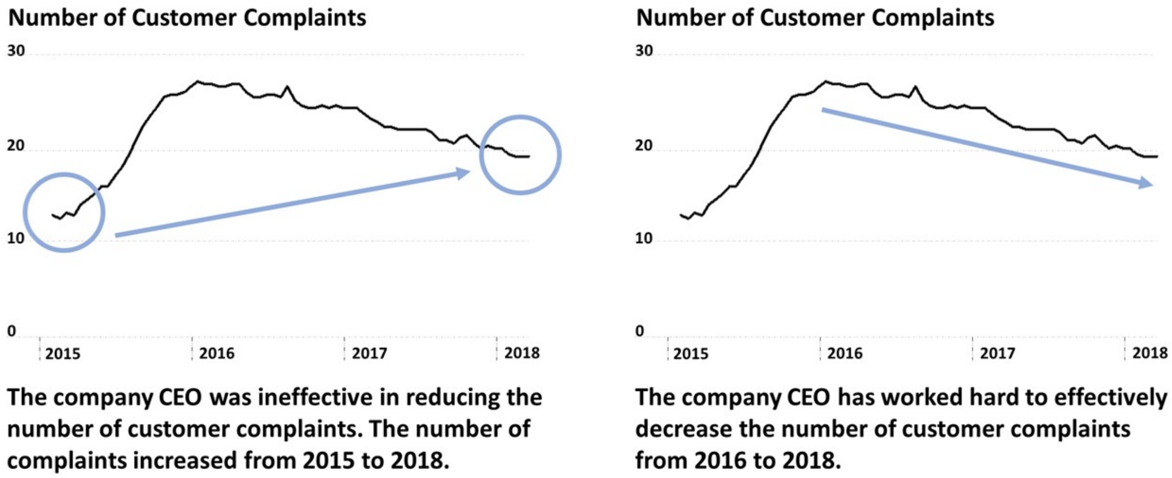
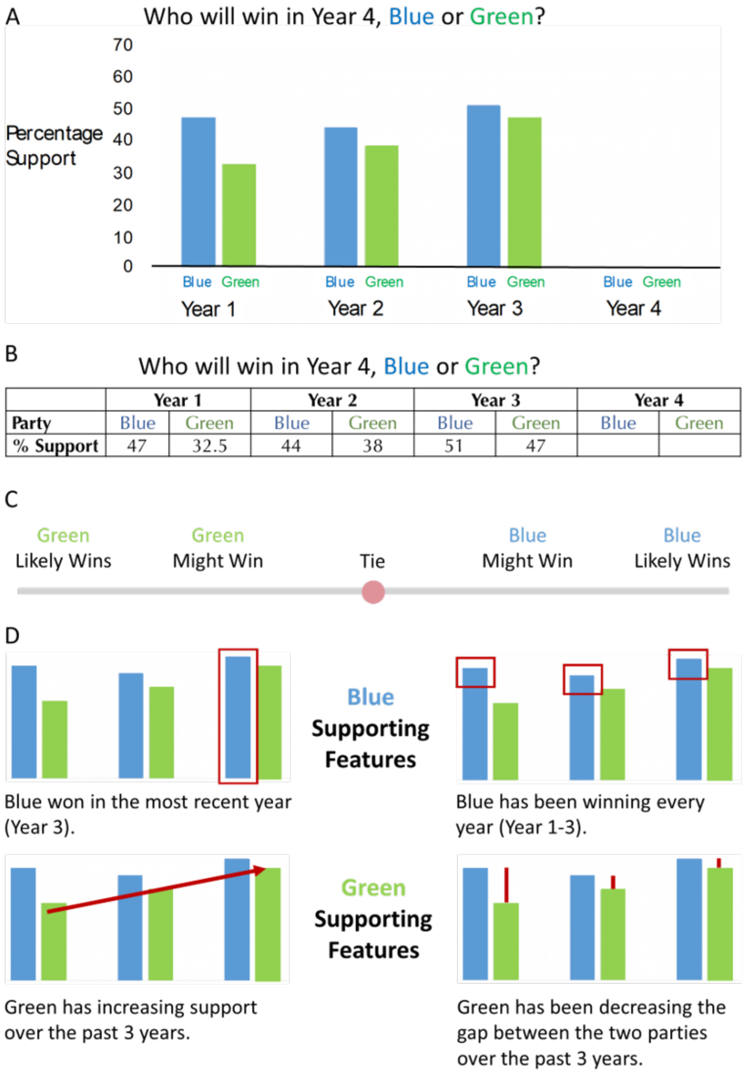


Figure 2.1. An adaptation of an example from data journalists (Bostock et al., 2012).

and Freedman, 2011). For example, when a scientist reads a typical bar chart depicting the results of a two factorial design, they could extract main-effect comparisons like "overall, people performed better in condition A than condition B" or interactions such as "group X performed better in condition A but group Y performed better in condition B" (Shah and Freedman, 2011). This diversity of percepts extractable from a graph might lead to two people seeing different patterns in the same dataset - one viewer might see the main effects (the duck) and another viewer might see the interaction (the rabbit). We know that people can diverge to see different patterns in visualizations and make different decisions due to their motivation, belief, or expertise (Parsons, 2018), but it is still unclear whether this could still happen without the influence of belief.

In the Chapter, I first examine whether the data presentation format impacts viewer decisions. We present some data depicting a neutral topic, either as a table or a bar chart, and compare what participants identify as a salient pattern in the data and how they make predictions about future trends in the data. I hypothesize that people could see different patterns and make different decisions in a visualization, but not in non-visual formats such as tables.



	Year 1		Year 2		Year 3		Year 4	
Party	Blue	Green	Blue	Green	Blue	Green	Blue	Green
% Support	47	32.5	44	38	51	47		

Blue won in the most recent year (Year 3).

Blue has been winning every year (Year 1-3).

Green has increasing support over the past 3 years.

Green has been decreasing the gap between the two parties over the past 3 years.

Figure 2.2. Visualizations are ambiguous figures. Stimuli used in Experiment.

2.1. Experiment 1

I recruited 125 (M_{age} = 32.84(8.81), 51 female) participants from Amazon's Mechanical Turk. I excluded workers who are not based in the United States, have an approval rate below 95%, failed our attention checks, have seen our experimental stimulus before (including in another study), or entered nonsensical answers. The participants were compensated at the rate of 9 dollars per hour.

Figure 2.2A shows a bar graph visualization that displays the results of a student government election across the past three years, with competition between the blue and the green party. The y-axis shows the percentage support the two parties received. Participants were asked 'Which party will win in Year 4?' The visualization contains competing data patterns that might be used to answer this question. As shown in Figure 2.2D, one could notice that the blue party has won most recently, or that blue has won every year from Year 1 to Year 3, both suggesting that blue is more likely to win. Alternatively, one might notice that the green party has been steadily gaining support over the past three years, or that it has decreased the gap between itself and the blue party over the past three years, and conclude that the green party is more likely to win. The underlying data from this bar graph could also be depicted in a table 2.2B. The table contains identical information in the same arrangement as the bar chart, except that data values were shown as numbers.

In the experiment, participants completed a survey released through Qualtrics (Qualtrics, 2014) on Amazon's Mechanical Turk. They viewed either the bar visualization or the table, as shown in Figure 2.2. They stated who they thought would win in the student government election in Year 4, via both a binary forced choice (blue or green party) and a slider to indicate the likelihood of their predicted victory. The slider ranged from 0 to 50 , where 0 indicated green likely wins, 25 indicated a tie, and 50 indicated blue likely wins, except that the numerical values were hidden from participants. They also briefly explained their reasoning. Afterwards, on a separate page, they matched their explanation to one of four choices, shown in Figure 2.2D (participants in the Table condition saw only the sentences, not the annotated graphs). These choices were collected from a series of pilot studies where participants indicated potential patterns that they saw in the same chart. If the participant couldn't match their reasoning to one of the choices, they were instructed to select "other."

2.1.1. Experiment 1 Results

All statistical analysis were done in R. The data, R-script used for analyses, and other experiment-related files are available at the Open Science Framework website: `https://osf.io/vmnh5/`.

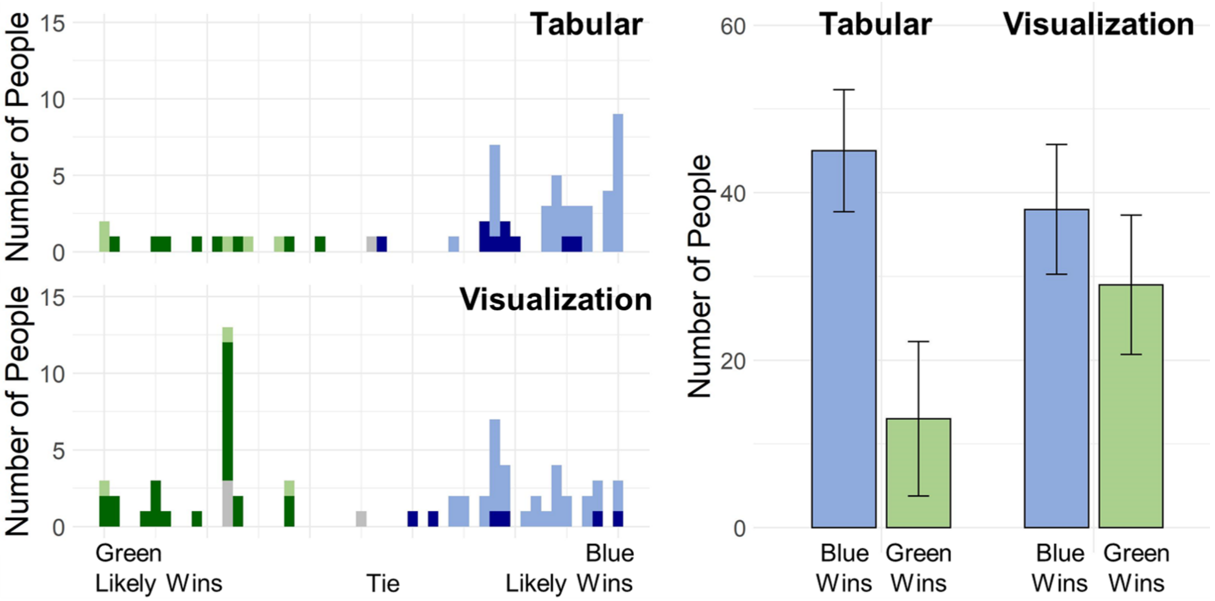


Figure 2.3. Experiment 1 results showing that visualizations are ambiguous figures.

This experiment shows that tables and bar graphs can afford different conclusions, even within the same trivially simple dataset. The left of Figure 2.3 compares the conclusions that people drew from the visualization versus the tabular data for continuous slider response. It shows the distribution of slider response on winning likelihood for blue and green party in visualization and table presentation. Responses in the visualization condition are bimodally distributed while responses in the tabular condition are right-skewed. People that viewed the bar visualization were equally likely to predict blue and green to win while people that viewed the table mostly predicted blue to win. The color represents what the participant indicated as their reasoning for their predictions. with blue representing participants who reasoned with blue supporting features and green representing participants who reasoned with green supporting features (see Figure 2.2D). Overall, the slider response shows a right-skewed distribution for the table, versus a bimodal distribution for the bar graph, suggesting that people's prediction patterns drastically differed depending on how the data was presented to them.

The color-coding in Figure 2.3 depicts the data pattern (see Figure 2.2) that participants picked out to support their prediction. Light blue maps to "blue won in the most recent year" and dark blue maps to "blue has been winning every year." Light green maps to "green has increasing support over the past 3 years." Dark green maps to "green has been decreasing the gap between the two parties over the past 3 years." Grey represents 'others' - where the participant indicated that none of the options captured the pattern that they noticed. Only 4.5% of the responses fell in the 'other' category and examples these responses included "I guessed". From the color distribution in Figure 2.3 left, we see that most people focused on blue-features when they predicted blue to win, and most people focused on green-features when they predicted green to win (χ^2=120, V=1, p<0.001), showing an association between the patterns that people notice, and the conclusions that they draw.

The right of Figure 2.3 shows the count of people that indicated blue or green would win in binary forced choice task fro visualization and text condition. Error bars represent standard error. There is a significant difference between the visual presentation format they saw and the predictions they made (χ^2=6.07, V=0.22, p=0.014), such that participants who saw the tables were more likely to say blue would win, and the participants who saw the bars were equally likely to say blue and green would win, corroborating our findings from the continuous slider shown on the left of Figure 2.3.

2.2. Experiment 2

Experiment 1 asked participants to predict a winner before providing the data pattern that they found salient, showing an association between the salient pattern and the eventual decision. However, putting the decision first might have led participants to pick a pattern that was consistent with their decision, **instead of pattern salience contributing to the decision.** To strengthen the evidence for that direction of this relationship, we reverse the order of these questions in Experiment 2, and tested only the bar graph condition, where responses showed the far stronger bimodality.

We recruited 141 participants for Experiment 2 (M_{age} = 33.40(9.91), 58 female), also from Amazon's Mechanical Turk and followed the same exclusion procedure as that in Experiment 1. We showed participants the bar visualization from Figure 2.2A and asked them to indicate the feature they find the most visually salient. Afterwards, on a separate page, we asked them to indicate which party they predict would win, using the same slider and the binary forced choice from Experiment 1.

When viewing the data as a bar graph, responses were again bimodally distributed across the likelihood of election outcomes. The left of Figure 2.4 shows the distribution of reported election outcome likelihood. Blue colors represent participants who indicated blue-supporting features as visually salient and green colors represent participants who indicated green-supporting features as visually salient, following the same color scheme as in Figure 2.3. Participants who predicted that “Blue likely Wins" were more likely to have indicated blue-supporting features as visually salient prior to making the prediction, and those who predicted that “Green Likely Wins" were more likely to have indicated green-supporting features as salient. No participants indicated the pattern corresponding to dark green to be visually salient. Again, we see that roughly half of the participants predicted that each party would win.

The right of Figure 2.4 shows the summary of binary decisions on which party would win, illustrating a congruence between salient patterns with predictions of which party would win. Error bars represents standard error. This observation is consistent with the slider response measures, as shown in the left of Figure 2.4. A chi-square analysis shows a significant difference between identified salient features for people that made differing predictions (χ^2=18.04, V=0.36, p<0.001).

2.3. Discussion

These two experiment shows that visualizations, unlike tables, are one form of ambiguous figures such that two people looking at the same dataset could come to different conclusion as they focus on different patterns. In the student government election dataset, the increasing

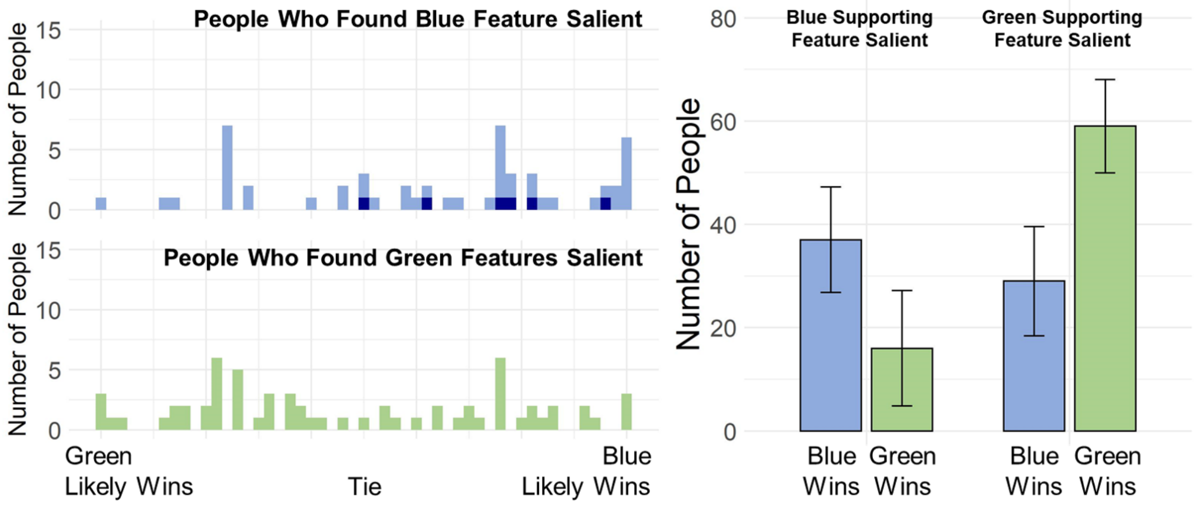


Figure 2.4. Experiment 2 results showing that visualizations are ambiguous figures.

trend of the green party was mostly undetected by participants when the data was presented as a table. Most participant noticed that the blue party has won every year, including the most recent year. When the data was visualized into a bar chart, half of the participant saw the increasing trend in green as visually salient and used that information to predict that green would win the election. This supports our hypothesis that visual representation of data can afford different pattern extraction and thus different decisions with data.

Why were people split in their decisions when the data is shown as a visualization but not when the data is shown as a table? The bar chart visualizes the growing trends for the green party, but this trend is less obvious when the data is presented as a table. We suspect there is a bottom-up process such that visualization emphasizes the pattern of the increasing popularity of the green party, making it more noticeable, and thus increasing the likelihood for people to predict green to win.

There also seem to be a consistent individual differences in perceived saliency of visual features. Future iterations of this work aims to examine the strength of visual saliency in data patterns relative to motivation and belief. We aim to improve data communication to facilitate data-driven, instead of motivation-driven decisions by identifying perceptual components or data visualization techniques that influence cognitive decision making.

These two experiments have limitations that provides promising future research directions. In the present design of the study, participants did two tasks: indicating the features they found salient and predicting who they thought would win the election. While the order of the two tasks has been counterbalanced in Experiment 1 and 2 to show consistent results despite which care first, we don't yet know whether there exists a causal relationship between the two, and in which direction the causal arrow points.

First, what we observed could be entirely perception-driven. Viewers may, through a bottom-up process, find certain features and patterns visually salient, and then form a belief to drive their decisions. For example, imagine a participant looking at the bar chart on student government election. The three blue bars are overall taller and more noticeable. The participant's attention is drawn to these tall blue bars. They didn't even notice that the green bars and them increasing overtime. Their impression on the large blue bars subsequently led them to form a belief that the large blue bars are so large that the blue party will likely win.

On the other hand, viewers could exert their top-down attentional control and pick out patterns that supports a pre-existing belief they have. For example, imagine a participant who believes that the general preferences of voters don't change over time. This participant then looks for patterns in the data that supports this belief, noticing that the blue party has been historically victorious. They then report the pattern of the blue bars beating the green to be very salient. This is also an instance of confirmation bias, as the viewer has already made up their mind and is merely looking for evidences to support their belief.

Both processes could have played a role in our experiment, influencing how viewers perceived and made decisions about visualizations. This may imply that bottom-up process can interact with top-down processes in data interpretation to form a feedback loop, creating a circular causal chain similar to the "which came first: the chicken or the egg?" question. Say an individual sees some data and makes a decision using the data. Did they form a belief first and then seek out corresponding patterns in the data? Or did they found certain patterns visually salient first and then form the belief? Future iterations of this work could further tease these two processes apart, and asses their effect in real-world visual data communication instances such as the one in Figure 2.1.

I encourage future researchers to bridge work in human cognition and data visualization to shed light on the impact of information visualization on data communication and decision-making. As we increasingly rely on data to understand, communicate, and make decisions, we need to further understand how our brains work to extract critical values, statistics, and patterns needed to make decisions about data, to help us design data visualizations that increase data comprehension and effective data communication. As Klein and O'Brien pointed out, people use less information than they think to make decisions (Klein and OâBrien, 2018). Now more than ever before, it is critical to consider implications of visualization design to facilitate data-driven, instead of motivation-driven or saliency-driven decisions with data.

CHAPTER 3

Perceptual Bias in Data Visualizations

“Seeing is believing” implies that vision delivers reality. Yet, experiments in perceptual psychology reveal that perception of a variety of visual information can be systematically influenced by recent history and context. Staring at downwardly moving dots can cause a ‘waterfall illusion’ where people may perceive subsequent, static dots as moving in the opposite direction (Wade, 1994). Similarly, a circle will appear smaller in size in the context of larger, concentric circles than when surrounded by smaller, concentric circles, as in the Ebbinghaus illusion (Roberts et al., 2005).

A small set of systematic biases based on history and context is already on the radar of visualization designers, such as how hue categories can bias the perception and memory of colors, or how background colors can strongly alter the perception of foreground colors (Ware, 2012). But we generally assume that data visualizations are otherwise perceived in an unbiased (albeit, potentially noisy) manner, particularly for more precise visual data encodings such as position (Cleveland and McGill, 1984a).

The perception of visual magnitudes can be biased across multiple feature dimensions. Orientation estimates for a line can be either repulsed or attracted by the orientations of nearby objects, depending on the parameters of the display (Parkes et al., 2001). Perceived brightness can be affected by background brightness (Ware, 2012). Categorical boundaries between hues can exaggerate differences between those that straddle a boundary, compared to hues that do not (Bornstein and Korda, 1984).

Similar patterns of categorical bias are found for object sizes once a set of object size categories have been learned (Kosslyn et al., 1977). When a short ellipse is in the context of a taller ellipse, the short ellipse is perceived as even shorter and the taller ellipse as even taller (Sweeny et al., 2011). After adapting to a vertically tall ellipse that is no longer present, the perceived height of a subsequent circle will appear vertically shorter (Köhler and Wallach, 1944). Memory for the size of a single object in a crowd of other objects can be biased toward the average size of all objects (Ariely, 2001; Chong and Treisman, 2003; Brady and Alvarez, 2011).

Most importantly for the present study, even the perceived positions of objects can be biased, especially when positional changes break a categorical or relational boundary. Positions are encoded somewhat categorically (similar to color), so that changes to a category boundary are easier to detect, compared to equally distant changes that do not cross a category boundary (Kosslyn et al., 1989). Viewers are more accurate at detecting change for a dot position when the dot is moved to the opposite side of a set of crossed gridlines, as opposed to moving to an equally distant position on the same side of the gridline (Kranjec et al., 2014). When detecting changes to the spacing between pairs of circles, performance is higher when the changes affect the categorical position relations for the circles ('touching' to 'not touching'), as opposed to when they do not ('not touching but close' to 'not touching but far') (Kim and Biederman, 2012; Lovett and Franconeri, 2017).

Most relevant for the present work is the susceptibility of position estimates to biases in memory, such that the position memory of a briefly presented object can be biased by context or nearby salient points in the display. Memory for the position of an object can be biased toward the average position of an associated group of objects (Alvarez and Oliva, 2008; Lew and Vul, 2015). The recalled position of a dot inside a circle is pulled toward the center of one of four imaginary quadrants within the circle (Huttenlocher et al., 1991), suggesting that the nearest of these salient category boundaries pulled the position representation toward itself during recall.

Similar effects of irrelevant ‘background’ values can bias cognitive processing, such as magnitude estimates at the level of verbal, numerical reports. In the ‘anchoring effect’ (Tversky and Kahneman, 1974; Furnham and Boo, 2011), uncertain target estimates can be strongly biased by other provided values, even if they are objectively irrelevant. For example, a population estimate of Nova Scotia would be biased by introducing an irrelevant value (e.g. “Is the population of Nova Scotia more or less than 200,000?”), such that a larger primed value (e.g., 200,000) would lead to a larger population estimate, while a smaller primed value (e.g., 20,000) would lead to a smaller estimate (Jacowitz and Kahneman, 1995).

Such higher-level cognitive influences may also affect pattern perception in data visualizations. Socially-derived information signals (e.g., polls) can influence graph perception, such that other individuals’ judgment of graphical information can bias how a single individual perceives and judges the same information (Hullman et al., 2011). The intensity of title-wording can cause graph viewers to overestimate or underestimate the slope of an associated, noisy scatterplot line, such that viewers who saw a high-intensity title recall a steeper slope than those who saw a low-intensity title (Newman et al., 2018). Previously viewed scatterplots can also influence viewer judgments of class separability in novel scatterplots, which could be interpreted similarly as an anchoring effect (Valdez et al., 2018). For example, priming with a clearly separable point cloud can bias perception of an ambiguous point cloud to appear more separable than if primed with a non-separable point cloud.

The present experiments will show that, despite the high precision of position, perception of data from positional encodings can also be biased in systematic ways. Specifically, depending on the visualization, these data values can be significantly underestimated, overestimated, or ‘pulled’ toward other irrelevant position values present in the same graph.

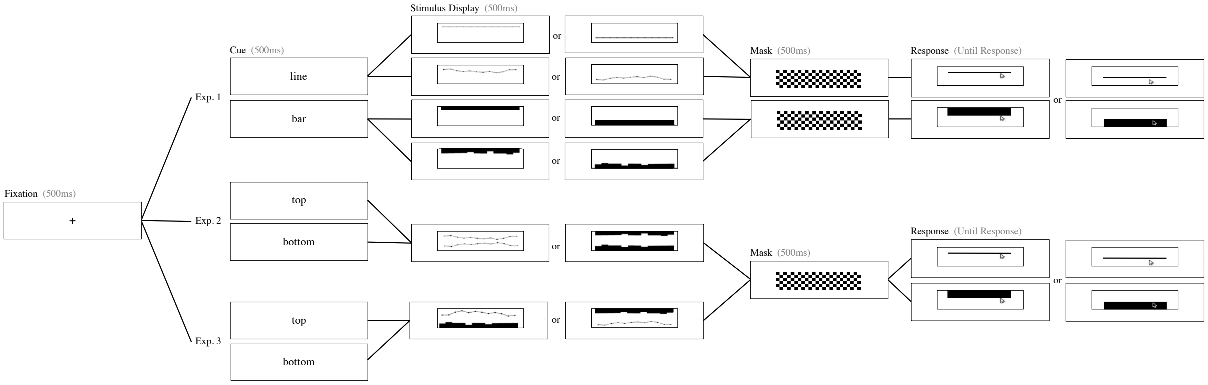


Figure 3.1. Experimental procedure and design for Chapter 2 Experiments.

3.1. Experiment 1

Experiment 1 tests how accurately people can perceive average positions of a single line or single set of bars in a graph. This experiment establishes a baseline to understand how a *single* graphed data series is perceived without the potential influence of other graphed data series. These findings will be later compared to how the presence of an *additional* graphed data series may further bias perception. We investigates whether people report average line and bar positions in graphs in a biased way by comparing participants’ *estimation* of average line or bar positions to the *true average position* of the line or bars in a mixed-model design.

Participants were cued before each trial with the task of either estimating the average position of a line or a set of bars to be shown on a subsequent stimulus display (see Figure 3.1). Depending on that precue, participants then saw a stimulus display that contained a set of bars or a line appearing on the top or bottom half of the display (see Figure 3.1). As shown in Figure 3.2, these data series could be uniform (where all points on the line or on the set of bars are of the same value), or noisy (where all points are of different values).

For a total of 576 trials, each participant completed 288 trials for each line and bar position estimate, with half of trials for each condition displaying a noisy version of the data series and the remaining half of the trials displaying a uniform data series. For analysis, we examined the average position estimations participants made across all these dimensions. We also examined the effect of the initial location of the response probe.

Thirteen undergraduate students from Northwestern University (M_{age} = 18.62 years, SD_{age} = 0.65) participated in exchange for course credit in an introductory psychology class. We excluded one participant who did not complete the experiment from our data analysis.

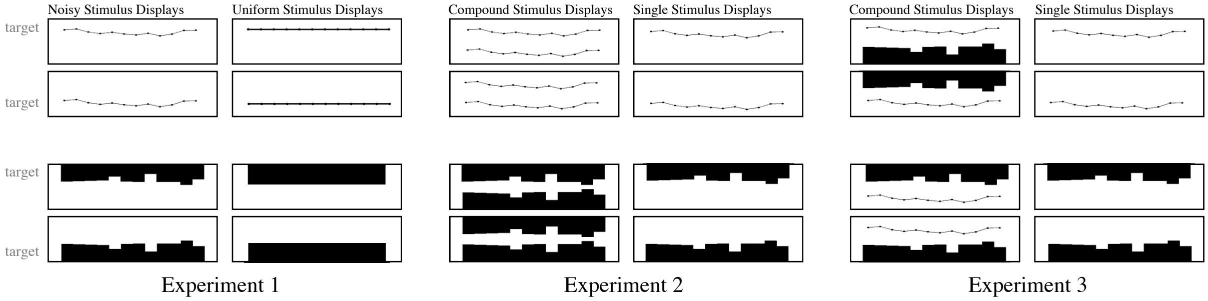


Figure 3.2. Design space for Experiment 1, 2, and 3 in Chapter 3.

3.1.1. Underestimation of Lines

We used a mixed-effect linear model to predict estimated position with fixed effects of the actual display location of the line (i.e., the pixel value and whether it was on the top or bottom half of the display), whether the line was noisy or uniform, the initial location of the response probe, practice effect (as trial number), and participants as random intercept. In this model, we excluded all trials in which participants made an obviously wrong estimate, defined as making an estimate in the *bottom quarter* of the display for a stimulus appearing in the *top half* of the display, or vice versa.

The top two rows of Figure 3.4 shows the results of the line estimates in Experiment 1. The first and fourth columns show example displays representing the 'medium' mean position for noisy and uniform displays, respectively. The second and third columns show the density curves for true average position distributions *(solid black lines for each of the three means: high, medium, low; pixel values given)*, and density curves for the estimated average position distributions for each of the three means *(line estimates: solid red lines; bar estimates: solid blue lines)*. The orientations of the solid red or blue lines (shown between the density curves for the estimated average position distributions) show the differences in position estimates between noisy and uniform conditions.

As shown in Figure 3.4 (top two rows), we observed an overall underestimation of line position, in which participants estimated average line positions to be lower than where the average position of the line actually appeared ($MD_{overall}$=-4.49 pixels, SE=0.11, Est=1.01, $CI_{95\%}$=[0.95, 1.07], p<0.001, $\eta_{partial}$=0.68). This underestimation persisted regardless of whether the line appeared on the top or bottom half of the display screen (MD_{bottom}=-6.50, SE_{bottom}=0.15, MD_{top}=-2.47 pixels, SE_{top}= 0.16, Est=-5.57, $CI_{95\%}$=[-8.45, -2.70], p<0.01, $\eta_{partial}$=0.96), although participants underestimated the average of the bottom line more than the average of the top line. Estimation error did not depend on whether the line was noisy or uniform (Est=0.12, $CI_{95\%}$=[-0.28,0.53], t=0.61, p=0.54, $\eta_{partial}$=0.00). This provides further evidence that the underestimation of average line positions is not an artifact of the noise in the line stimulus, as underestimation occurred for even uniform lines on the display.

There was also a very small interaction between the average pixel position of the line and whether that line was presented in the top half or bottom half of the display (*Est*=-0.094, $CI_{95\%}$=[-0.13, -0.054], *p*<0.001, $\eta_{partial}$=0.006). For lines appearing in the top half of the display, participants underestimated average position *less* when the lines were located closer to the center of the screen. For lines appearing in the bottom half of the display, participants underestimated average position *less* when the lines were located closer to bottom boundary of the display box (further away from center).

We found no practice effect ($\eta_{partial}$=0.00, *p*=0.45), suggesting that average position estimations did not get more or less biased as participants completed more trials. We also found that while initial probe location (where the response probe for position estimations was initially presented in the display) had a significant influence on the estimation error (*Est*=0.62, $CI_{95\%}$=[0.21, 1.02], *p*=0.0022, $\eta_{partial}$=0.003), its effect size was small (see Figure 3.3).

Was this systematic underestimation an artifact of poor average estimation strategies? We considered whether the underestimation in average line position was the result of participants simply choosing the lowest point on the noisy line as their response for the average position of the line. To test this, we compared participants' estimated average line positions with the lowest point on the noisy line using a different mixed-effect linear model, with estimated line locations as the dependent variable (DV) and the position of the lowest point on the noisy line as the independent variable (IV). If participants estimated the average line only relying on the position of the lowest point, the slope of the linear model should be 1. Using a Wald test with confidence intervals at 95%, we found that the test slope of the mixed-effect model was in the range of [0.64, 0.70] for lines on the top of the display, and in the range of [0.57, 0.63] for lines on the bottom of the display. Neither range included the value 1, which suggests that participants did not base their average position estimations only on the lowest point for lines on both the top and bottom half of the display. We did a similar comparison predicting whether estimated average positions depended on the average of the highest and lowest points on the line, and found the slope of this regression line to also not include 1 in its 95% confidence interval, (Est_{top}=[-0.057, 0.12], Est_{bottom}=[-0.10, 0.06]). This suggests that the participants were not simply averaging the highest and lowest points on the line stimulus to make their estimations.

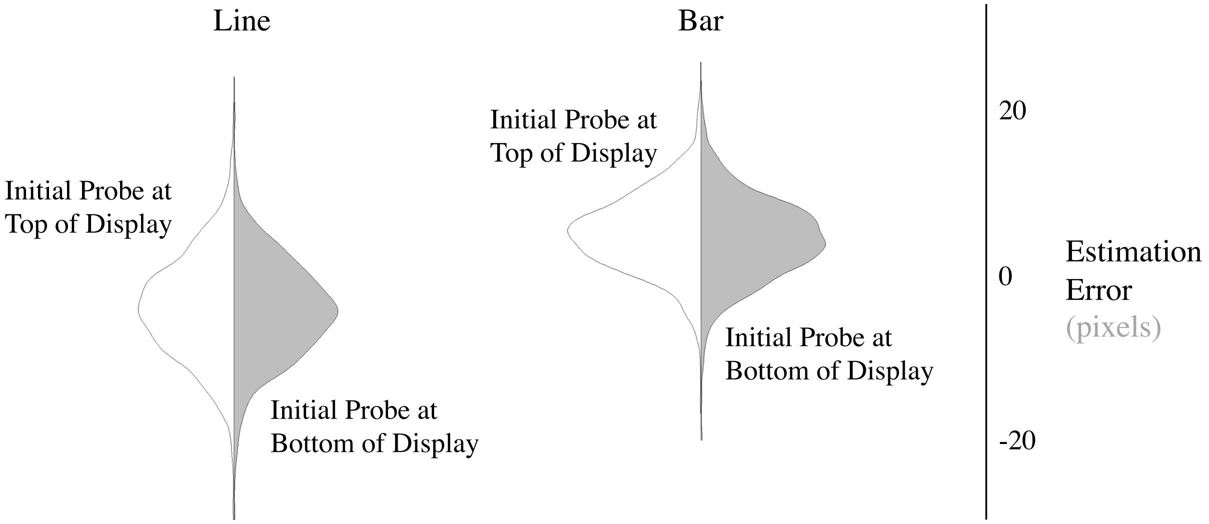


Figure 3.3. Difference between the initial probe location in average line and bar estimation tasks.

3.1.2. Overestimation of Bars

We utilized the same mixed-effect linear model to examine estimated average bar positions. As shown in Figure 3.4 (bottom two rows), we observed an overall overestimation of bar position, where participants systematically estimated average bar positions to be higher than its actual position ($MD_{overall}$=4.19 pixels, SE=0.086, Est=0.72, $CI_{95\%}$=[0.59, 0.84], $\eta_{partial}$=0.33, p<0.001). This overestimation persisted regardless of whether the bar stimulus appeared on the

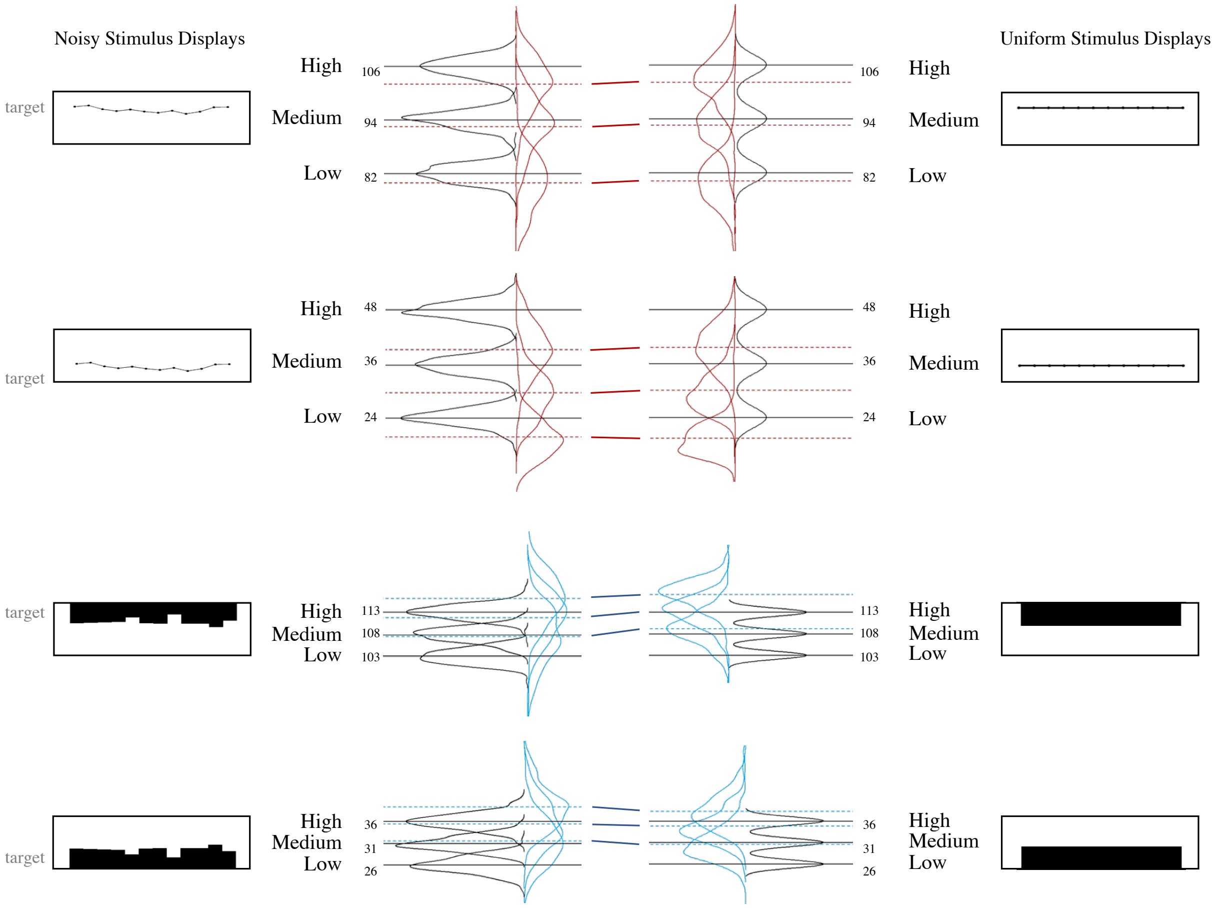


Figure 3.4. Results from Experiment 1 for noisy and uniform simple graphs.

top (MD_{top}=4.43, SE_{top}=0.12) or bottom half of the display screen (MD_{bottom} = 3.95, SE_{bottom} = 0.12, Est=10.26, $CI_{95\%}$=[4.00, 16.51], p<0.001, $\eta_{partial}$=0.98), although participants overestimated the average position of the bottom bars significantly more than the average position of the top bars. There was no effect of whether the bars were noisy or uniform (MD=0.15, SE=1.33, Est=0.12, $CI_{95\%}$=[-0.21, 0.45], p=0.57, $\eta_{partial}$=0.00), suggesting that this systematic bias may not be the result of participants utilizing outliers for position averages. There was no practice effect ($\eta_{partial}$=0.00, p=0.85), but an effect of initial probe location ($\eta^2_{partial}$=0.007, p<0.001) with a small effect size (see Figure 3.3).

We then further investigated whether this overestimation was an artifact of participants basing average bar position estimations on the highest point on the bar graph. We compared participants' estimated average bar positions with the highest point on the noisy set of bars using the same method as before, and found the 95% confidence interval for the test slope of the model to be in the range of [0.29, 0.36] for the bottom set of bars and [0.22, 0.29] for the top set of bars. Since 1 is not included in either confidence interval, the results suggest that participants did not simply base their average position estimations on the highest point on the bars. Similar comparisons between estimated average position and the average of the highest and lowest points on the bars also found that the slope of the regression line did not include 1 in its 95% confidence interval (Est_{top}= [0.01, 0.11], Est_{bottom}=[0.22, 0.29]), suggesting that the participants were not simply averaging the highest and lowest points on the set of bars to generate their average positional estimates.

3.1.3. Discussion

Experiment 1 illustrated a systematic underestimation of average line positions and overestimation of average bar positions. Interestingly, this effect occurred regardless of whether the lines and bars were noisy or uniform. Comparing the variance of the estimations via an F-test, we found that while participants estimated uniform lines and uniform sets of bars with more precision and less variance than with noisy lines and noisy sets of bars ($F(1,3454)$=1.1479, p<0.0001), biases were still prevalent in both noisy and uniform displays. Overall, this experiment showcased that position encoding is not immune to perceptual bias.

3.2. Experiment 2

Experiment 1 provided evidence for biased reports of average position for a simple graph with one data series, but how is this bias affected by the presence of an additional data series within a complex graph? Experiment 2 expands on this investigation of biases by determining whether the perception of average line and bar positions can be further biased by other lines and bars present on the same graph, respectively. We first explore how the position of one line may influence the perceived position of a target line within a display containing two lines (referred to as "compound line-line" displays). We then test the effect of an additional set of bars on the average position estimations of a target set of bars in a display with two sets of bars (referred to as "compound bar-bar" displays).

The methods were similar to those in Experiment 1, with the following changes. Participants were presented with noisy compound line-line or noisy compound bar-bar displays and were precued to report the average position of a line or a set of bars presented at the top or bottom of the display (see Figure 3.1).

For a total of 240 trials, each participant completed 120 trials for each line and bar average position estimation condition. For half of these trials, participants were tasked with judging the position of the top data series and in the other half, the bottom data series. We also included 144 control trials in which each participant estimated the average position of a single noisy line or a single noisy set of bars (referred to as "single-line" or "single-bar" displays; see Figure 3.2), replicating Experiment 1.

Twelve different undergraduate students from Northwestern University (M_{age}=19.31 years, SD_{age}=1.55) participated in exchange for course credit in an introductory psychology class.

3.2.1. Underestimation and Overestimation

Using the same analysis method as Experiment 1, Experiment 2 replicated the results of Experiment 1 in the single-line and single-bar conditions.

In the single-line displays, participants still underestimated average line positions (*MD*=-5.27, *SE*=0.18, *p*<0.001) (see first two rows in Figure 3.6). Further analyses showed no practice effect ($\eta^2_{partial}$=0.00, *p*=0.91) and no effect of initial location of the response probe ($\eta^2_{partial}$ = 0.003, *p* = 0.12).

In single-bar displays, participants overestimated average bar positions (*MD*=3.39, *SE*=0.17, *p*<0.001) (see last two rows in Figure 3.6). We found no practice effect ($\eta^2_{partial}$=0.001, *p*=0.23), but a small effect of the initial response probe location ($\eta^2_{partial}$=0.004, *p*=0.0025).

3.2.2. Perceptual Pull

For compound line-line displays and compound bar-bar displays, we observed an underestimation in average line positions and an overestimation in average bar positions. Additionally, we also found an effect of “perceptual pull”: position estimates for a target data series (a set of bars or a line) were ‘pulled’ toward the irrelevant data series shown on the same graph.

We used another mixed-effect model predicting estimation error with fixed effects of whether the data series was present in the top half of the display or the bottom half, whether the display was compound or single, and trial number, and a random effect of participants. In compound line-line displays, there were no main effects of line location (top or bottom) (*Est*=-2.05, $CI_{95\%}$=[-2.22, -0.44], *p*=0.82, $\eta_{partial}$=0.00) or display type (compound or single) (*Est*=-2.05,

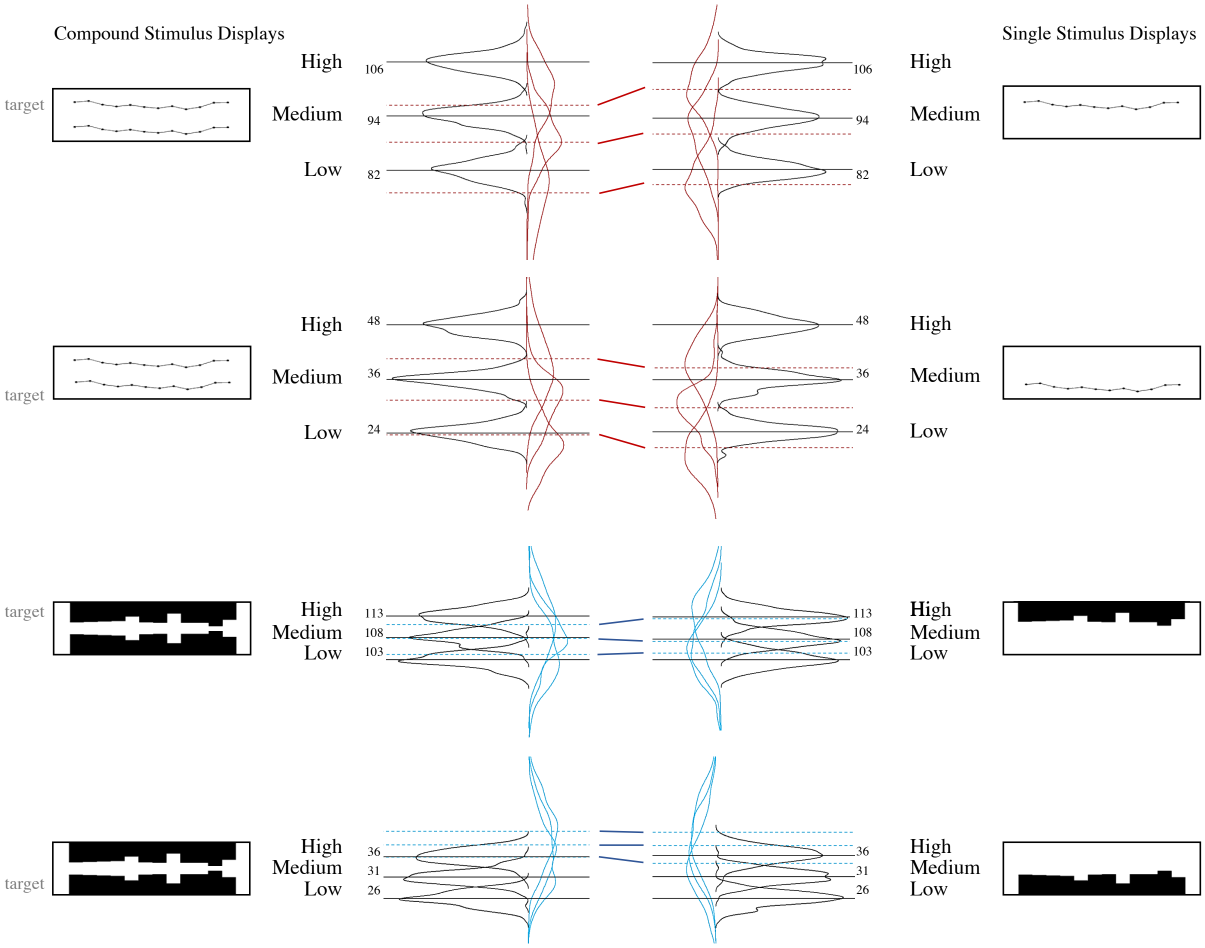


Figure 3.5. Results from Experiment 2 for compound stimulus displays (i.e., one line and one set of bars) and single stimulus displays (i.e., one line or one set of bars).

$CI_{95\%}$=[-3.07, -1.02],p=0.82, $\eta_{partial}$=0.00). However, there was a significant interaction between line location and display type, such that the magnitude of line position underestimation between lines that appeared on the top and bottom half of the display differed (Est=3.75, $CI_{95\%}$=[2.30, 5.19], p<0.001, $\eta_{partial}$=0.011). Underestimation of the top line was *exaggerated*, such that participants underestimated the top line even more compared to single-line displays

($MD_{top,single-compound}$=-1.92, *SE*=0.71). In contrast, underestimation of the bottom line was *reduced* compared to single-line displays ($MD_{bottom,single-compound}$=2.01, *SE*=0.65).

From this interaction between display type and location of the line (top or bottom) on position estimation, we speculate that the top and bottom lines ‘pulled’ position perception toward one another. The presence of the irrelevant top line further biased perception of the average position of the bottom target line, and the presence of the irrelevant bottom line further influenced perception of the average position of the top target line.

We also observed the same perceptual pull in the compound bar-bar displays using an identical mixed model. There was a significant main effect of bar position such that bottom bars were significantly more overestimated than top bars (*Est*=-8.08, $CI_{95\%}$=[-8.81, -7.34], $p<0.01$, $\eta_{partial}$=0.22). There was no main effect of display type (*Est*=-0.61, $CI_{95\%}$=[-1.46, 0.23], p=0.63, $\eta_{partial}$=0.00), but there was a significant interaction between bar location and display type (*Est*=1.51, $CI_{95\%}$=[0.32, 2.71], p=0.013, $\eta_{partial}$=0.003). Overestimation of the top bars was *reduced*[1], such that participants overestimated the top bars less compared to single-bar displays ($MD_{top,single-compound}$=0.97, *SE*=0.44). In contrast, overestimation of the bottom set of bars was *exaggerated* more compared to single-bar displays ($MD_{bottom,single-compound}$=-0.53, *SE*=0.48). This interaction again suggests that, similar to the lines, there exists a perceptual pull between bars, ‘pulling’ the perception of their average positions toward one other.

3.2.3. Discussion

Experiment 2 replicated evidence that perception of average line and bar positions is biased. Furthermore, it showcased a perceptual pull effect, in which the presence of an irrelevant line

[1] Note that this is a decrease in absolute position in the reference frame of the display, meaning participants perceived the top set of bars as vertically longer in the compound bar-bar displays than the single displays.

or set of bars in the same display pulled average position estimations of a target line or set of bars toward the position of this irrelevant data series (see Figure 3.6).

3.3. Experiment 3

Experiment 2 illustrated the existence of perceptual pull, but what determines the extent of the perceptual pull? Is it data series-dependent (in which a specific type of graphed data series could influence the strength of the perceptual pull more than another type)? Is perceptual pull dependent on sufficient perceptual similarity between the two graph elements (i.e., irrelevant bars would exert a larger influence on average bar judgments than on average line judgments)? Experiment 3 expands upon Experiment 2 by diversifying the types of graphed data series present in the display to test the extent to which a line and a set of bars on the same graph can influence each other.

3.3.1. Design and Procedure

The methods used in Experiment 3 were similar to that of Experiment 2. We will refer to displays with a line present in the top half and a set of bars in the bottom half of the display as “compound line-bar” displays, and displays with a set of bars on the top half and a line on the bottom half of the display as “compound bar-line” displays (see Figure 3.2).

Twelve different undergraduate students from Northwestern University (M_{age}=19.00 years, SD_{age}=1.04) participated in this experiment in exchange for course credit in an introductory psychology class.

3.3.2. Underestimation and Overestimation

Experiment 1 and 2 results were replicated. In single-line displays, there was still a systematic underestimation of average line positions as lower than their true positions (*MD*=-6.47, *SE*=0.26, p<0.001), and an overestimation of average bar positions as higher than their true positions (*MD*=3.85, *SE*=0.19, p<0.001) (see far right of Figure 3.5). There was no evidence of a practice effect in either the line-bar (p=0.36, $\eta_{partial}$=0.00) or the bar-line (p=0.23, $\eta_{partial}$=0.00) displays, but there was an effect of the initial location of the response probe with a small effect size for both line-bar (p<0.001, $\eta^2_{partial}$=0.03) and bar-line conditions (p<0.001, $\eta^2_{partial}$=0.02).

3.3.3. Perceptual Pull

As in Experiment 2, we still found an underestimation in average line positions and an overestimation in average bar positions for all compound displays. Similarly, we observed an effect of perceptual pull between the two data series (one line and one set of bars) in the display.

We used a mixed-effect model with display type (single or compound), graphed data series type (line or bar), the interaction between display type and graphed data series type, trial, and initial probe position as fixed effects, and with participants as a random effect. In the compound line-bar condition, we found a significant main effect of graphed data series type (Est=-11.14, $CI_{95\%}$=[-12.16, -10.12], p<0.001, $\eta_{partial}$=0.188), a significant interaction of display type and graphed data series type (Est=4.72, $CI_{95\%}$=[3.06, 6.39], p<0.001, $\eta_{partial}$=0.014), a small effect of initial probe location (Est=3.41, $CI_{95\%}$=[2.60, 4.21], p<0.001, $\eta_{partial}$=0.188), and negligible effects from other predictors (all $\eta_{partial}$<X, all ps>X). The underestimation of the top line was *exaggerated* when compared to the single-line displays ($MD_{topLine,single-compound}$=-2.96,

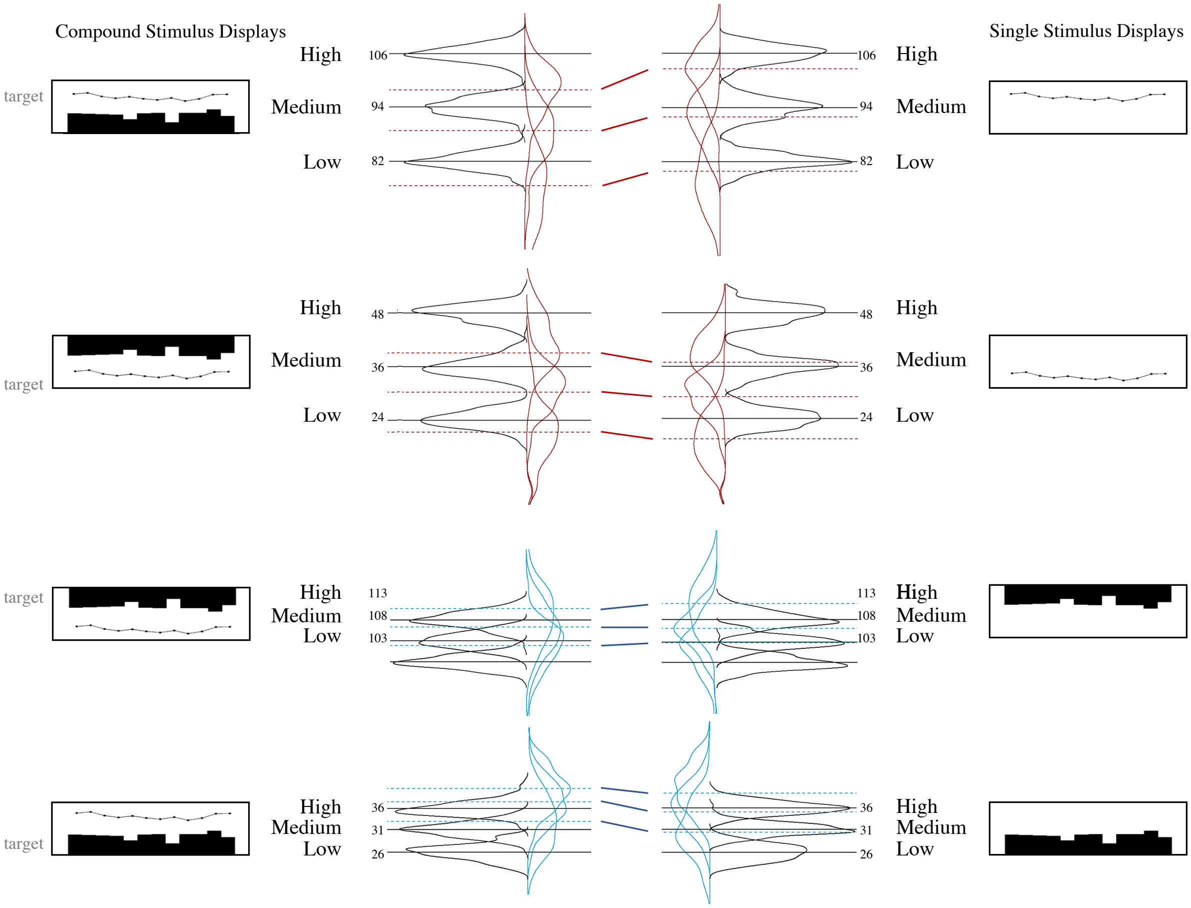


Figure 3.6. Results from Experiment 3 for compound stimulus displays (complex graphs; i.e., two lines or two sets of bars) and single stimulus displays (simple graphs; i.e., one line or one set of bars).

SE=0.90, *p*<0.001), as the bottom bars ‘pulled’ the average positional percept of the top line down. Overestimation of the bottom set of bars was also *exaggerated*, such that participants overestimated the bar position more when there was an above line ‘pulling’ the average positional percept of the set of bars up ($MD_{bottomBar,single-compound}$=-1.37, *SE*=0.53, *p*<0.001).

Similarly, in the compound bar-line condition, a mixed-effect model showed a significant main effect of graphed data series type (*Est*=7.90, $CI_{95\%}$=[6.94, 8.85], *p*<0.001, $\eta_{partial}$=0.209), a significant interaction effect between display type and graphed data series type (*Est*=3.45,

$CI_{95\%}$=[1.88, 5.01], p<0.001, $\eta_{partial}$=0.008), a small effect of initial probe location (Est=2.86, $CI_{95\%}$=[2.10, 3.62], p<0.001, $\eta_{partial}$=0.023), and negligible effects from other predictors (all $\eta_{partial}$<0.01, all ps>0.10). Overestimation of the top set of bars was significantly *reduced* when a line appeared below its position on the display ($MD_{topBar,single-compound}$=1.49, SE=0.67, p<0.001), suggesting that the bottom line 'pulled' the average positional percept of the set of bars down. Underestimation for the bottom line was also *reduced*, such that the average position for the bottom line was underestimated less ($MD_{bottomLine,single-compound}$ = 1.50, SE = 0.65, p < 0.001), as the top set of bars 'pulled' the average positional percept of the bottom line up.

In the compound displays overall, the effect of perceptual pull *exaggerated* the underestimation of average line positions and the overestimation of average bar positions when a line was located in the top half and a bar was located in the bottom half of the display. But perceptual pull *reduced* the same underestimation and overestimation bias when a line was located in the bottom half and a bar was located in the top half of the display.

3.3.4. Strength of Influence

The effect of perceptual pull occurs across graphed data series type, but is the *extent* of this perceptual pull dependent on the type of data series present? In other words, would a data series pull the same or different type of series more strongly, or would it pull all data series equally?

With data from Experiment 2 and 3, we conducted a between-subject ANOVA comparison examining the variation in average line and bar position estimations depending on the non-target data series (a line or a set of bars) (see Figure 3.7, which compares the average estimated line/bar positions when the non-target data series is a line (*red*) or a set of bars (*blue*)). Neither top nor bottom target line position estimations in a compound line-line display were significantly different from that in a compound line-bar display (top: *MD*=0.46, *SE*=0.70, *p*=0.57, with Tukey correction; bottom: *MD*=1.31, *SE*=0.56, *p*=0.42). Similarly, neither top nor bottom target bar position estimations in a compound bar-bar display were significantly different from that in compound bar-line displays (top: *MD*=3.11, *SE*=0.52, *p*=0.21; bottom: *MD*=1.98, *SE*=0.43, *p*=0.29). This suggests that the extent of perceptual pull does *not* depend on data series type.

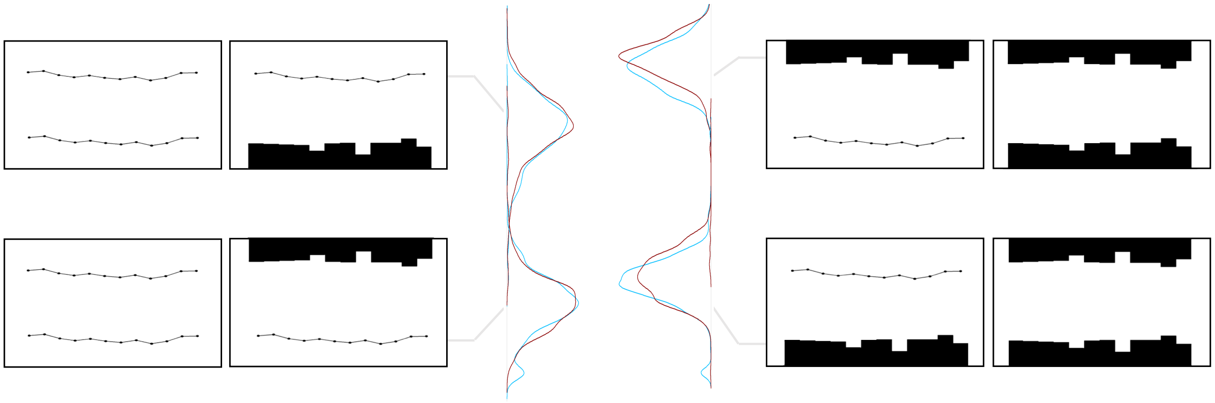


Figure 3.7. Estimation error for target data series is not dependent on the irrelevant graphed data series.

3.3.5. Discussion

Experiment 3 showed that perceptual pull is not dependent on graphed data series type, but can generalize across data series (e.g., lines and/or bars). A single, irrelevant line has a similar pulling force for a target set of bars as for a target line, and vice versa for a single set of irrelevant bars on a target set of bars or a target line. To further examine the extent to which average position estimates for a target line or set of bars are perceptually pulled by non-target, graphed data series, we introduce a perceptual mixture model.

3.4. Mixture Modeling

We propose a simple perceptual mixture model to computationally quantify the strength of the perceptual pull exerted by other graphed data series onto the target data series. The basic idea of the model is that a compound stimulus generates independent percepts of both graphed data series (the target and the ‘non-target’), and the reported target estimate is thought to be a weighted combination of these two independent percepts. An ideal observer would have a weight associated with the target data series of 1 and a weight for the non-target data series of 0, indicating no influence of the irrelevant non-target on perception of the target. However, the perceptual pull phenomenon indicates that the weight associated with the target data series will be less than 1, due to influence from the non-target data series.

The aim of this section is simply to assess the viability of this idea. As such, we develop the model based on the ‘top line, bottom bar’ compound stimulus from Experiment 3. For the sake of illustration, suppose the top line is the target data series. To model the compound stimulus judgments, we proceed as follows:

(1) Draw a random sample from the empirical distribution of single line estimation *errors*, X_{target}.

(2) Draw a random sample from the empirical distribution of single bar estimation *errors*, $X_{non-target}$.

(3) Draw a random *pair* of true line and bar positions in the compound stimulus. Add the sampled single line and bar estimation errors to these true positions, giving the single data series percepts in a compound stimulus: x^*_{target} and $x^*_{non-target}$.

(4) For a given weight associated with the target, compute the target line estimate in the compound stimulus: $y^*_{target} = w_{target}x^*_{target} + (1 - w_{target})x^*_{non-target}$.

(5) Repeat steps 1-4, M times ($M = 10,000$ below). The resulting distribution may be regarded as a synthetic distribution of compound target estimates, Y^*.

(6) Fit the resulting distribution of compound target estimates with a kernel density. This is a kernel density approximation of the likelihood Hartig et al. (2011).

(7) For each observed target estimate, $y_{i,obs}$ for $i = 1,\ldots,N$, in the compound condition, compute the approximate likelihood by plugging it into the kernel density (using linear interpolation).

(8) Across all observations, compute the overall log-likelihood as $\mathscr{L} = \sum_{i=1}^{N} \ln p(y_{i,obs}|\theta)$, where θ is the vector of model parameters (i.e. the weight associated with the target data series).

(9) Find the best-fitting weight parameter(s) through maximum likelihood estimation.[2]

In modeling the data, we made a number of other necessary assumptions. First, we must assume that the only graphical elements influencing position estimations are the lines and bars on the display, and do not include other potentially biasing elements, such as the display frame. Second, all participants are assumed to behave in the same way. We have limited data for individual participants, so estimating model parameters at the level of the individual would involve sampling from a limited set of single data series estimates, which is likely to bias the approximate likelihood. Third, position estimation errors are assumed to be independent of the true positions of the graphed data series. That is, we sample from the single data series distributions pooled over the three different target means. As a result, any effects of the true position of the data series in a compound stimulus are not captured by the model. However, our data indicate that any such effects were very small ($\eta_{partial}$=.006; see section 6.2). Fourth, we

[2] Note that the likelihood is stochastic, because it is based on weighted combinations of many random samples. We will deal with this issue below.

assume that the weight associated with the target data series can depend on whether the target is the line (and the non-target is the bar) or whether the target is the bar (and the non-target is the line) (i.e., $w_{target=line} \neq w_{target=bar}$). Since we previously observed asymmetrical behaviors in average line and bar position estimations (underestimating lines and overestimating bars), this final assumption allows for asymmetric effects of non-target bars on target lines and vice versa.

These assumptions mean that a given compound stimulus configuration (e.g., top line and bottom bars) is modeled with just two free parameters: the relevant target weight, which is allowed to depend on which data series is the target[3]. We make no assumptions about the nature of the single data series error distributions. The only assumption we make about the compound error distribution is that it is a weighted combination of the distributions of single data series errors. Overall underestimation or overestimation of the graphed data series is accounted for by the single data series error distributions, and the model captures how these overall shifts are 'pulled' by competing, irrelevant data series.

For the test case here (top line and bottom bars), there were 1552 target estimates in a compound stimulus (773 line estimates; 779 bar estimates). Best fitting weight parameters were found using the standard 'optim' function in R (R Core Team, 2014). Comparison with an optimal observer was performed by comparing the model fit with a 0-parameter model in which the weight associated with the target data series was set to 1 and the weight associated with non-target data to 0. For this optimal model, the distribution of compound target estimates is simply equivalent to the corresponding distribution of single target estimates. Model comparison was performed using the Akaike Information Criterion. To accommodate stochasticity in the likelihood function, we performed the optimization 50 times with random starting points

[3] It could be argued that the bandwidth of the kernel density is also a model parameter, but we simply adopted the default rule of thumb for a Gaussian kernel density (Silverman, 2018).

for the weight parameter over the interval $[0.9, 1]$. We report the average AIC difference with the optimal model (mixture model AIC - optimal model; negative values indicate a better fit of the mixture model). Any variation in the log-likelihood for the mixture model stems from two sources: the best-fitting parameters may depend slightly on the randomly chosen starting values. However, even with the same (best-fitting) parameters, the model would return a slightly different likelihood due to the random sampling from the single data series error distributions. For the baseline model, only this second source of variation was considered.

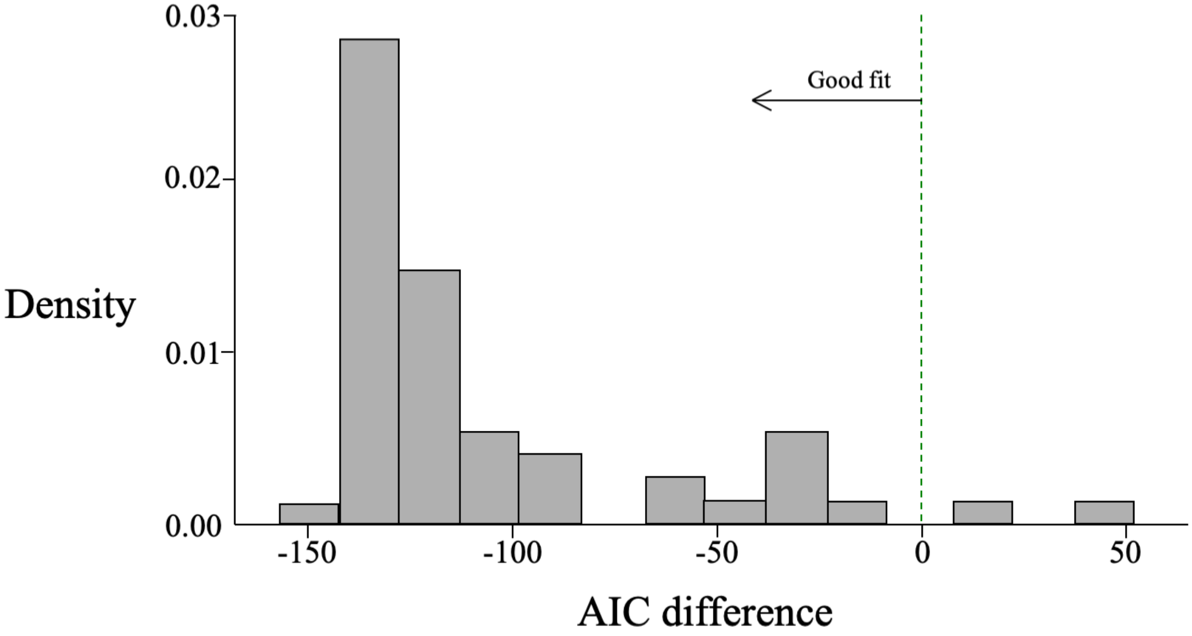


Figure 3.8. The mixture model of perceptual pull is suggested to be a good fit. In 50 iterations of optimization, only 4% had positive AICs.

For the 50 model fits conducted in this way, the average AIC difference was -104 (95% highest density interval=$[-145, -22]$). Only 2 out of 50 model fits had positive AICs, as shown in Figure 3.8. The mean weight for the line target was $0.945, CI = [0.901, 0.990]$, and the mean weight for the bar target was $0.971, CI = [0.933, 0.989]$. These weights are clearly close to 1, but

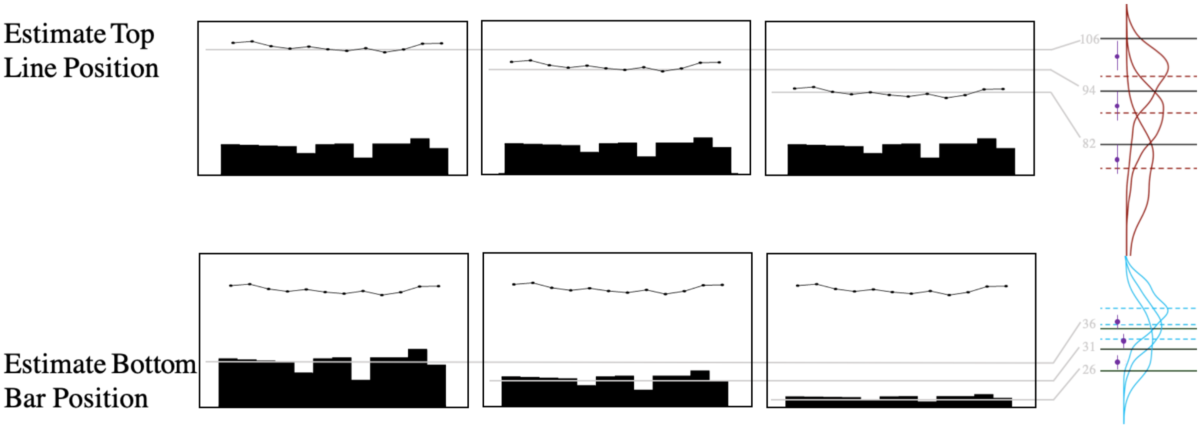


Figure 3.9. Comparison of mean model fit and 95% confidence intervals of model predictions (represented by *purple lines*).

the 95% highest density intervals do not include 1. We would expect the weights to be close to 1, because the perceptual pull effects are relatively subtle but reliable. This suggests that when participants viewed compound line-bar displays, they made average line position estimates by attributing 94.5% weight to the actual line position and the remaining 5.5% to the bars below the lines, on average. Participants made average bar position estimates by attributing 97.1% weight to the actual bar positions and the remaining 2.9% to the lines above, on average. Figure 3.9 illustrates the model fit compared to the observed data.

This model isolates the unique contribution of perceptual pull from the irrelevant depiction of the second data series. When there were two data series present on the graph, participants were not able to completely filter out the irrelevant data series. As a result, their target position estimate reflected a weighted combination of the target and non-target positions, reflecting an effect of perceptual pull.

3.5. Discussion

In three experiments, we empirically test estimation accuracy of average line and bar positions in simple or complex visualizations. Experiment 1 investigates whether systematic biases exist in position estimates for simple graphs, where a single line or a single set of bars is the only present, graphed data series. Experiment 2 and 3 investigate potential positional biases found in complex graphs. Specifically, Experiment 2 explores how average position perception of a graphed data series can be distorted in the presence of an identical type of data series (e.g., single line presented with another line on the display, or a single set of bars presented with another set of bars). Experiment 3 combines graphed data series (e.g., lines and bars on the same display) to examine how the average perception of one type of data series can be distorted by a different type of data series. Figure 3.2 illustrates this design space.

Much of the previous literature investigating perceptual biases has focused on biases within visualizations in a higher-level, cognitive context — how do elements (e.g., priming, titles, axes, etc) influence perception of relevant data? In the current study, we are interested in potential biases found in lower-level perception of data — can people accurately perceive purely graphical information? Specifically, we focus on possible biases within average position estimations to better understand perception of distributed visual information in both simple and complex graphs. Position averaging is not only one of the most common and crucial tasks when interpreting visualized data, but also an area few have studied.

These findings show that people systematically perceive graphed data series in a biased manner[4], underestimating the average positions of lines and overestimating the average positions of bars in a graph. We also found that under- and over-estimation of the target line or set of bars, respectively, can be exaggerated or diminished by the presence of other such graphed data series. The average position estimates for lines or bars tend to gravitate toward the positions of the other lines and/or bars present on the same graph. We call this perceptual bias of irrelevant, graphed data series on relevant, targeted series "perceptual pull".

[4] Note that any mention of "underestimation" or "overestimation" is in relation to the 140-pixel display frame, where values of $[0, 140]$ map onto the bottom and top of the frame, respectively. For example, for a set of top (downward-pointing) bars with an average of 100 pixels, overestimation (100-140 pixels) in this display-based frame will actually reflect shorter bar lengths. We chose this naming convention to be more closely aligned with real-world scenarios where downward-pointing bars are observed, such as bars that depict negative values. See Figure 6.1 for clarification.

CHAPTER 4

The Curse of Expertise

Imagine a scientist showing experimental results at a conference or colloquium, or a data analyst updating the company leadership on recent customer feedback with a dashboard. These people are experts in their respective fields, yet they overwhelm their audiences with overly complex visualizations, delivered too quickly, oblivious to the fact that others do not see what they see. We replicated this phenomenon in the lab, providing empirical evidence for a 'curse of knowledge' in data visualization â once an expert recognizes a given pattern in data as visually salient, the expert assumes that it is also visually salient to naïve observers.

This 'curse of knowledge' is a well-studied psychological phenomenon that appears in many domains. Well-informed decision makers fail to predict the judgments of less-informed decision makers, implicitly allowing their own knowledge to guide those predictions (Camerer et al., 1989). People given disambiguating information about ambiguous sentences, like "the daughter of the man and the woman arrived," assume that the sentence would no longer be ambiguous to other naïve listeners (Keysar and Henly, 2002). When people have access to additional information, e.g. that a message is sarcastic, they tend to perceive ambiguous messages such as "that restaurant was marvelous, just marvelous" as sarcastic – but they also predict that other people would read the same tone (Gilovich et al., 1998).

In one particularly powerful demonstration, people were asked to tap the rhythm of a set of well-known songs, such as "Happy Birthday," on a desk, and listeners guessed the songs based on the recorded rhythm of the tappers (Newton, 1990). Tappers estimated that listeners would identify around 50% of the songs, but in reality, listeners could only identify around 3%, revealing a vast overconfidence in how much information they communicated. The tappers ‘filled in’ missing information in their own heads, such as the pitches of the ‘notes’, and it appears impossible to turn off this filling-in process to simulate the experience of others. Taking a naïve perspective can be literally inconceivable (Roese and Vohs, 2012).

This curse of knowledge has powerful consequences for communication, because people generally do not convey information to others if they assume that it is already shared (Grice et al., 1975). Presenters must therefore have an accurate idea of what their audiences know and do not know, so that they can include only the information the audiences still need (Hart, 1996). Unfortunately, this knowledge is too often not present or not leveraged. Even teachers misjudge their students’ abilities and understanding, hindering effective instruction (Allbritton et al., 1996; Keysar and Henly, 2002; Ward et al., 1997).

Existing work in cognitive psychology shows that the curse of knowledge bias can impact interpersonal communication (Grice et al., 1975). The curse of knowledge can have particularly strong effects in children, who have more trouble inhibiting their own knowledge. In the ‘Sally-Ann’ task, children hear a story about Sally, who put her candy in a box before leaving the room. While she was gone, Ann removed the candy from the box and put it in a basket. Where will Sally look for the candy when she returns? Unable to inhibit their own knowledge of the illicit swap, most 4-year old children will assume that Sally will look in the basket (Bernstein et al., 2004; Pohl and Haracic, 2005). A modified ‘Sally-Ann Task’ targeting adults introducing several “boxes” and “baskets,” demonstrated that adults also make this error with a more complex scenario and a subtler measure (Birch and Bloom, 2007).

The curse of knowledge can also occur within a single person (Leary, 2007), in the form of ‘hindsight bias’. This bias, studied in business decision making, political strategizing and marketing, is the irrational belief that an outcome was more predictable after it becomes known (Roese and Vohs, 2012). People seem unable to recreate the novel and uncertain feelings from their own mind prior to the revelation of the outcome (Zwick et al., 1995; Cassar and Craig, 2009; Blank et al., 2003).

While the curse of knowledge is well-studied in the psychology of language, decision making and reasoning, there is less direct research on potential consequences for communication with data visualizations. Compared to numerical and textual formats, data visualizations are effective in highlighting the relationships and patterns in data to facilitate understanding (Card, 1999). But at the same time, understanding complex visualizations can be similar in time and effort to reading a paragraph (Hegarty, 2005; Khan and Khan, 2011). Critically, just as one can read many possible sentences from the paragraph, providing multiple perspectives on a topic, a graph or figure can be seen and interpreted in multiple ways depending on the how they select and interpret visual information over time (Shah and Freedman, 2011; Michal and Franconeri, 2017). The present experiment demonstrates that different experience with a dataset can cause people to adopt a particular perspective, which can substantially change their predictions about what naïve viewers will find salient in a visualization.

Given the primary role visualizations play in the communication of analytic data across science, education and industry (McKenzie et al., 2016; Knaflic, 2015), focusing on different patterns in the same dataset harbors the potential for miscommunications between the presenters and their audiences (Gilovich et al., 1998; Shah and Freedman, 2011; Yarbus, 1967). We suspect that the inability to separate one’s own knowledge and expertise from that of their audience can make visual data communication more difficult and less clear than presenters realize. This means that among the many features and patterns within a visualization, graph viewers could selectively focus on some while ignoring others, and in turn predict that naïve viewers would focus on the same feature and patterns.

Across four experiments, we demonstrate that the 'curse of knowledge' indeed extends to data visualizations. Knowledge, specifically, makes an expert recognize a given pattern in data as more visually salient, and the expert assumes that it is also visually salient to observers that they know to be naïve. To my knowledge, this would be provides the first examination of the curse of expertise in data visualization: whether a viewer's background knowledge will affect their predictions about what naïve others will see in a visualization.

4.1. Methods

Participants completed a Qualtrics (Qualtrics, 2013) survey in which they read a story that conveyed background knowledge about a graph depicting political polling data. They were told that the experimenters will show the same graph they saw to 100 people, along with only the following short description – "in the months before the elections of 2014 in a small European country, a polling organization asked citizens about their voting intentions on a daily basis." They were then asked to predict what uninformed viewers (with no knowledge of the story) would find to be the most visually salient features or patterns in the graph. The participants then predicted a second most salient feature, up to a fifth most salient feature. We intentionally did not specify what types of "features or patterns" the participants should predict, and did not provide them with examples. We defined "saliency" as "the most noticeable and important feature" for our participants. After writing down each feature they predicted, the participants also circled regions on the graph corresponding to each feature on a paper copy of the graph. They then reported how salient (1 = not at all salient, 5 = very salient) they thought their five predicted features were to themselves. Finally, they matched their five predictions as best as possible with five pre-determined features, as shown in Figure 4.1.

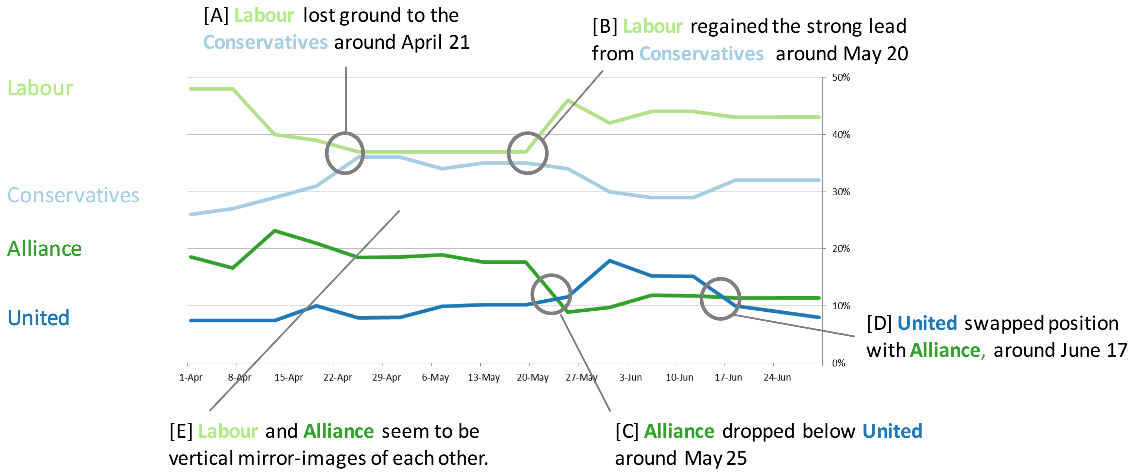
This within-subject experiment compares individual participant's saliency ratings of primed features (a subset of five critical graph features that were highlighted with a particular story) vs. unprimed features. We introduced three stories to counterbalance the possible primed or unprimed features, and randomly assigned participants to read one of those stories. The critical comparison in this experiment is between the salience ratings that participants assign for primed features vs. unprimed features. The independent variable is therefore whether a feature was primed or not, and the dependent variable the salience ratings for those features. We also measured a second dependent variable of how visually salient each participant rated their predicted features to themselves, on a continuous scale from one (very salient) to five (not at all salient).

4.2. Experiment 1a

Eighteen Northwestern University students (10 women) participated in this experiment in exchange for course credits in an introductory psychology class. All participants were asked to bring corrective eye wear if needed.

The participants read a story highlighting a competition between two out of four political parties, illustrating how citizen voting intentions fluctuated with current events. Figure 4.2 shows a sample display of the story highlighting the Labour and Alliance party.

Please rank the following statements (A, B, C, D, and E) to match your written ranking predictions, as best as you can. If you didn't write something down, select N/A.



[A] Labour lost ground to the Conservatives around April 21

[B] Labour regained the strong lead from Conservatives around May 20

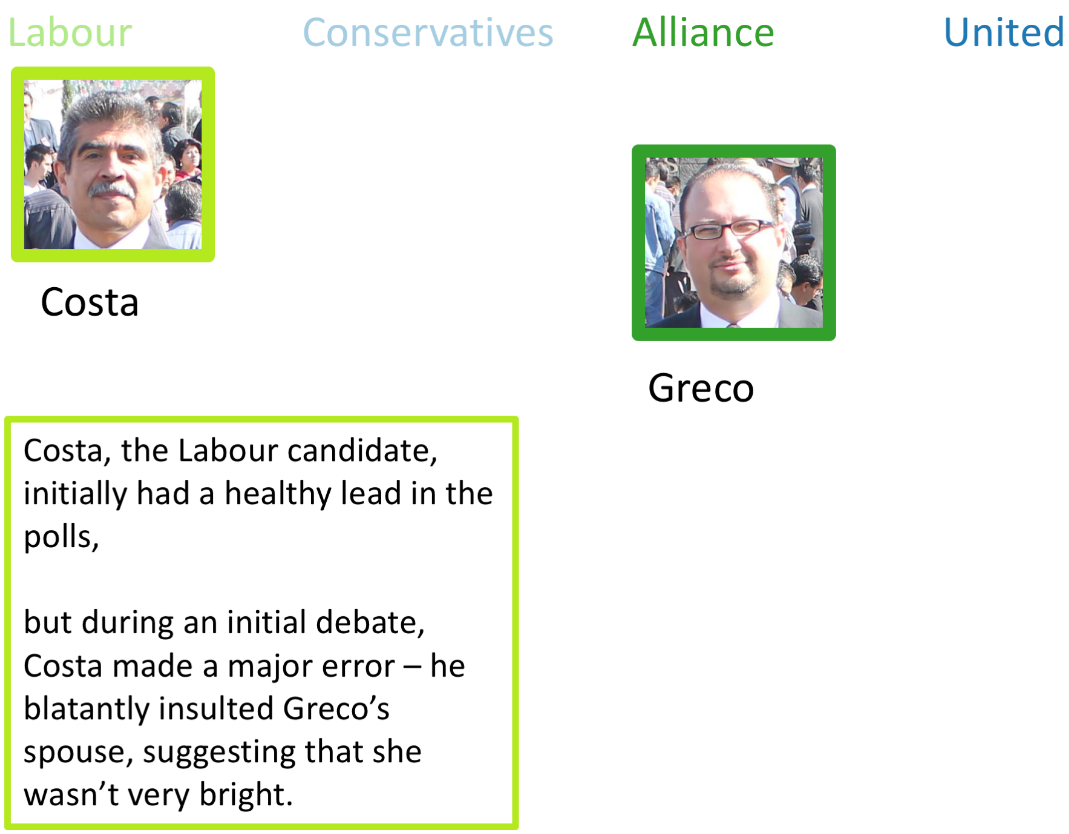
[C] Alliance dropped below United around May 25

[D] United swapped position with Alliance, around June 17

[E] Labour and Alliance seem to be vertical mirror-images of each other.

Figure 4.1. Matching five pre-determined features in Experiment 1.

According to the story, initially, between the two highlighted parties, one had a healthy lead in the polls. During an initial debate, the leading party lost voters to the less popular party and eventually lost the lead. In a later debate, the originally leading party was able to take back the votes the candidate lost and take the lead back again after a bad debate performance by his opponent. The three versions of the story all describe this same competition over time, but ascribing it to the top two parties (Top-Prime Story), the top and third party (Middle-Prime Story) or the bottom two parties (Bottom-Prime Story), highlighting the corresponding fluctuations. As shown in Figure 4.3, participants were randomly assigned to read a version of the story and were shown polling data after reading the story. In each pair of lines, the party with the higher line cedes votes to the party with the lower line (initial debate), and then the higher line gains back that ground (later debate).



Costa, the Labour candidate, initially had a healthy lead in the polls,

but during an initial debate, Costa made a major error – he blatantly insulted Greco’s spouse, suggesting that she wasn’t very bright.

Figure 4.2. Snapshot of the story participants read.

When participants predicted what an uninformed graph viewer would see as the most visually salient feature on the graph, they were shown an unannotated version of the line graph, depicted in Figure 4.4. They were told that this unannotated graph (with no story), was all that the uninformed graph viewers would see. Paper copies of this non-highlighted graph were provided to the participants to mark down their five predictions separately. We attempted to construct this graph in a way that balanced the relative salience of several critical features. The bottom two lines were made darker in color to balance the top two lines, which we expected to be more salient as a baseline (McKenzie et al., 2016). We further added two intersections to the bottom two lines to counter the top two lines’ natural visual saliency for just being on the top.

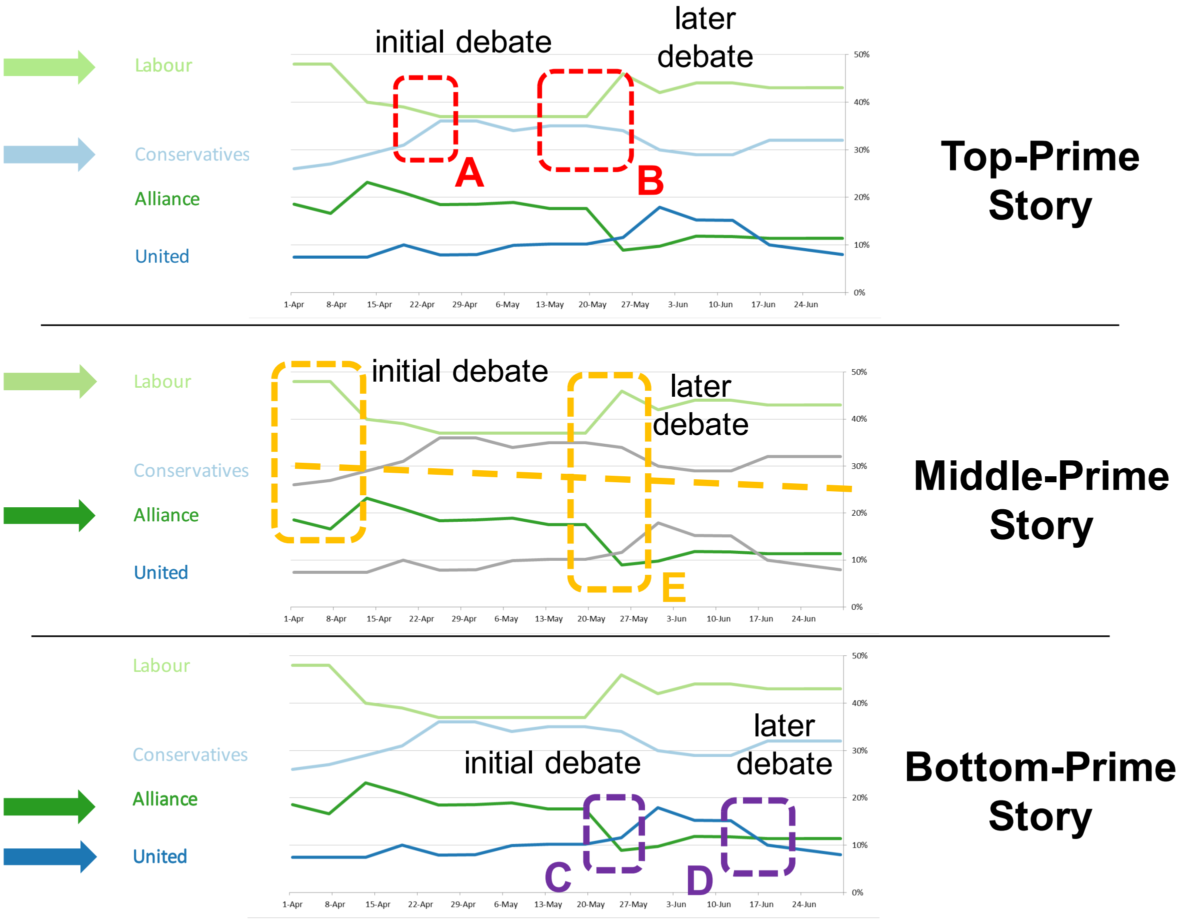


Figure 4.3. Three stories highlighting different features in Experiment 1.

We worried that the green ‘mirror image’ lines would form a less salient pattern, so we aligned their major change points to maximize the salience of that pattern. We also conducted several pilot versions of this experiment where we tracked the most salience features regardless of what was primed, and adjusted its appearance to equate those salience values (e.g., by making a peak less sharp, or a color difference stronger).

The participants then matched their own predictions to the five pre-determined features, referring to their markings on the paper copies of the unannotated graph, shown in Figure 4.4. A

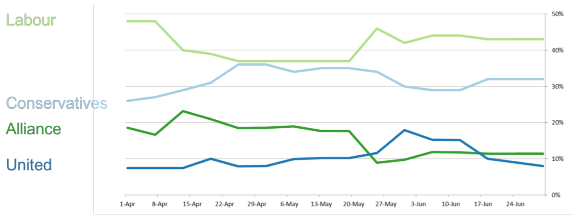


Figure 4.4. The unannotated graph of the line graph experiment.

subset of the five pre-determined features are highlighted in each of the three stories, as shown in Figure 4.3. The top-prime version of the story highlighted features A and B on top (describing the two almost-intersections of the top two lines). The middle-prime version of the story highlighted features C and D in the bottom right corner (describing the two intersections of the bottom two lines). The middle-prime version of the story highlighted feature E pointing towards the center section of the graph (describing the mirroring trend of the two green lines). Participants' referred to their freely identified salient feature drawings and matched them with the five features mentioned above. If the feature they drew did not match any of the five, they indicated it as "N/A." The subsequent quantitative data analysis of the saliency predictions and rankings were done on the rankings of the five pre-determined features. Among the five pre-determined features, 48% matched with the participants' freely identified salient feature drawings, and 56% matched if we only look at the participant's top three predictions. We discuss potential limitations of this approach at the end of this paper. We include the actual freely identified salient feature drawings of the predicted top three salient features in the qualitative results section to provide a fuller picture of the participants' responses in addition to our quantitative analysis.

4.2.1. Qualitative Results

Examining what the participants marked down on their physical copies of the unannotated graph, we find qualitatively observable differences among the three story versions. Figure 4.5 shows what the 18 participants who read different versions of the story (6 for each top, middle and bottom-prime story) marked on paper as their predictions of the most, 2nd most and 3rd most salient features to an uninformed viewer.

The top and bottom rows of Figure 4.5 directly compares the story versions and the respectively highlighted features to the overall predictions participants made. We see that depending on what version of the story participants read, free predictions reflected that they thought other uninformed viewers would see the features highlighted in their particular story as visually salient, even though participants were explicitly told to ignore the story when making their predictions. For example, looking at the bottom row of Figure 4.5, participants who read the top story identified features highlighting the top two lines to be salient more often than participants who read the bottom prime story and middle prime story. The participants who read the middle-prime story identified global and mirroring features to be salient to other viewers (notice how participants often circled pairs of features spanning a larger area), as opposed to local features identified by participants who read the top and bottom prime story.

4.2.2. Quantitative Results

Using the data from the feature matching section of the experiment, rankings were assigned to the five pre-determined features (ABCDE). The results are shown in Figure 4.6. For example, if a participant matched their most visually salient feature to uninformed viewer prediction to

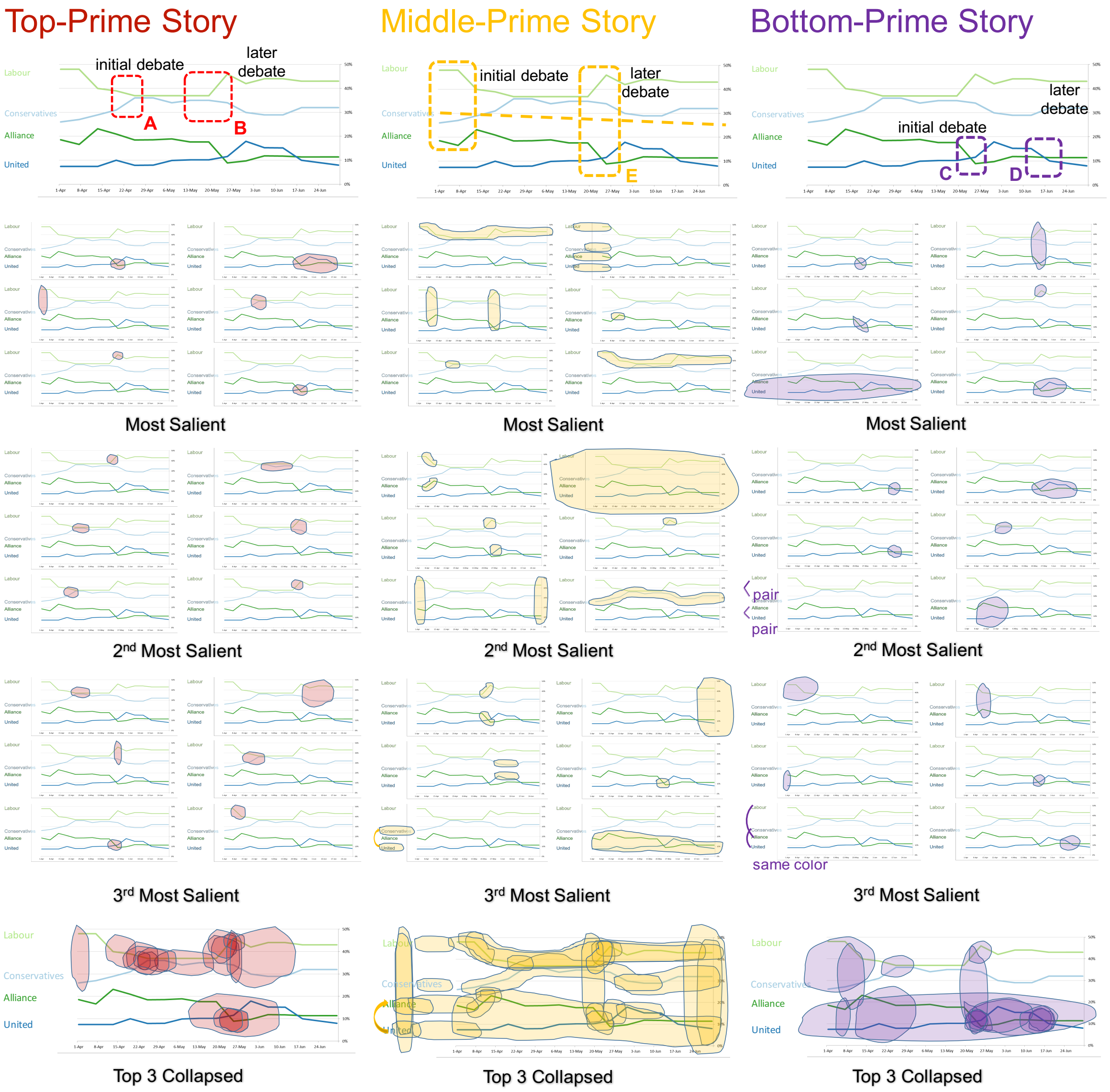


Figure 4.5. Summary of Qualitative Results from Experiment 1a. Each column represents one story version, read by 6 participants who marked their most, 2nd most and 3rd most salient feature predictions.

feature C (which is a bottom feature), feature C would receive a rank of ‘1’ for this participant. The rank ‘1’ would be entered in R for statistical calculations.

We reverse coded the rank in Figure 4.6, renaming it "saliency prediction," to be more intuitive (e.g. a feature ranked '1' will have a saliency prediction of '5'). For example, if a participant matched their predicted fourth-most feature to feature B (which is a top feature), feature B would receive a rank of '4' and reversely coded as '2' on the 'saliency prediction' axis in Figure 4.6.

If a participant matched pre-determined features to multiple predictions, then the feature would receive the ranking of the highest rank. For example, if a participant matched their predicted second and third salient features to feature A, then feature A would receive a ranking of two.

If a participant did not think any of the five pre-determined features matched to one of their predictions, that specific prediction would be matched to "N/A." The ranking spot of this prediction would be counted as taken. For example, if a participant matched the predicted second most visually salient feature to feature E, the fourth most visually salient feature to feature D, and every other prediction they made did not match to any of the five pre-determined features, feature E would receive a rank of '2' and feature D would receive a rank of '4.' Remaining unranked features (ABC) would take on a rank of '6,' which translate to "saliency predictions" of '0.'

If participants matched two features to a predicted feature, the two features would receive the same rank (e.g., if a participant wrote down a feature to be the second most visually salient feature to an uninformed viewer and matched both feature A and B to it, then both feature A and B would receive a rank of '2.')

4.2.2.1. Wilcoxon Signed-Ranked Test. We conducted a non-parametric Wilcoxon Signed-Rank Test comparing the participants' saliency rankings of primed and not primed features

(Kerby, 2014). Primed feature rankings are rankings of features highlighted in the story the participant read. For example, the middle feature (E) rankings ranked by participants who read the middle-prime version of the story are primed feature rankings. Non-primed feature rankings are rankings of features not highlighted in the story the participant read. For example, top (AB) and bottom (CD) feature rankings ranked by participants who read the middle-prime version of the story are non-primed feature rankings.

The Wilcoxon Signed-Rank test indicates that the overall primed feature rankings, Wilcoxon mean score = 59.77, rank mean = 2.63, were significantly higher compared to the overall not primed feature ranks, Wilcoxon mean score = 38.37, rank mean = 0.87, $Z = 4.03$, $p < 0.01$. Primed features were given higher saliency rankings and thus were predicted to be more visually salient to other uninformed viewers than not primed features.

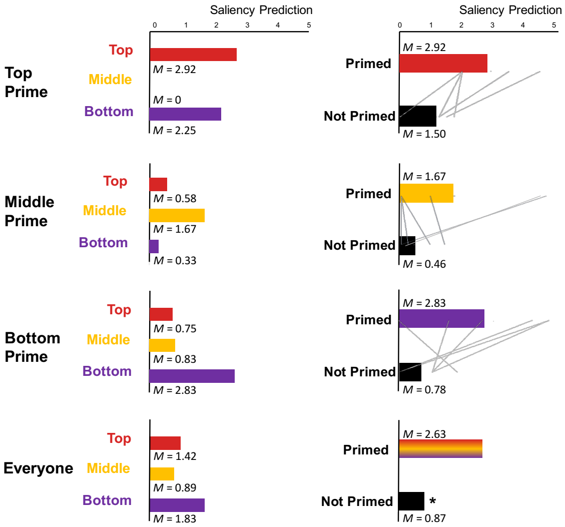


Figure 4.6. Ranking details for each story version. The grey oriented lines represent individual participant ratings. The right column shows saliency ratings of primed and not primed features (e.g. in Top Prime, top is primed; middle and bottom are not primed), across the three stories.

4.2.2.2. Descriptive Statistics. In order to more clearly illustrate the differences in saliency rankings, we visualized their descriptive statistics. Since there are two pre-determined features highlighted in the top-prime and bottom-prime stories, and only one pre-determined feature is highlighted in the middle-prime story, the rankings of the top features (A and B) were averaged to generate a top feature average ranking. Similarly, the rankings of the bottom features (C and D) were averaged. The left column of Figure 4.6 shows the participant prediction rankings of the top features (AB), middle feature (E) and bottom feature (CD) for the three story versions (no standard deviation is shown because ranking data is nonparametric). The right column of Figure 4.6 shows saliency ratings of primed and not primed features (e.g. in Top Prime, top is primed; middle and bottom are not primed), across the three stories.

Overall, most participants rated features that were highlighted in the story (primed), as more visually salient than other features that were not highlighted in the story (not primed). This supports the results of our Wilcoxon Signed-Rank Test. Inspecting the grey lines in the right column of Figure 4.6, we also see that some participants did not rate the primed features as more visually salient. This might mean that these participants were relatively immune to the curse of knowledge, though the present design cannot distinguish robust individual differences from measurement (or other sources of) noise.

4.2.2.3. Salience Prediction Ranking. After participants marked down a feature that they predicted other uninformed graph viewers would find visually salient, participants also rated how visually salient that predicted feature was to themselves. We see from Figure 4.6 that not everyone predicted the story-primed features to be visually salient to others. In the present analysis, we take a different approach here by looking at whether the participants would find features they predicted to be salient to other people also salient to themselves, regardless of whether they were primed features or not.

In Figure 4.7, 'Saliency to Self' is how salient each participant's predictions were to themselves on a continuous scale, where one means not at all visually salient, and five means very visually salient. Feature Rank is the order of the predictions. For example, 1 corresponds to the feature the participant predicted to be the most salient and 5 corresponds to the feature the participant predicted to be the 5th most salient, to a naïve viewer. Each dot represents one rating from one participant and the three lines are regression lines based on the scattered points.

There was a negative correlation between the Feature Rank and Saliency to Self, showing that regardless of whether the features were primed or not, participants rated the features predicted to be the most/least visually salient to a naïve viewer also to be the most/least visually salient to themselves, suggesting a curse of knowledge where they could not separate their own perspectives from that of another person. Using Spearman's Correlation, we found a moderately strong association ($r_s = 0.55$, $p < 0.001$) between the self-rated salience of a feature, and the predicted salience rating for other naïve observers.

4.2.3. Discussion

The knowledge the participants obtained by reading the story biased their predictions such that, in general, they saw the features depicted in the story as more visually salient than features not depicted in the story. More importantly, after acquiring this background knowledge, participants were biased to predict that other uninformed graph viewers would rate those features as more visually salient as well.

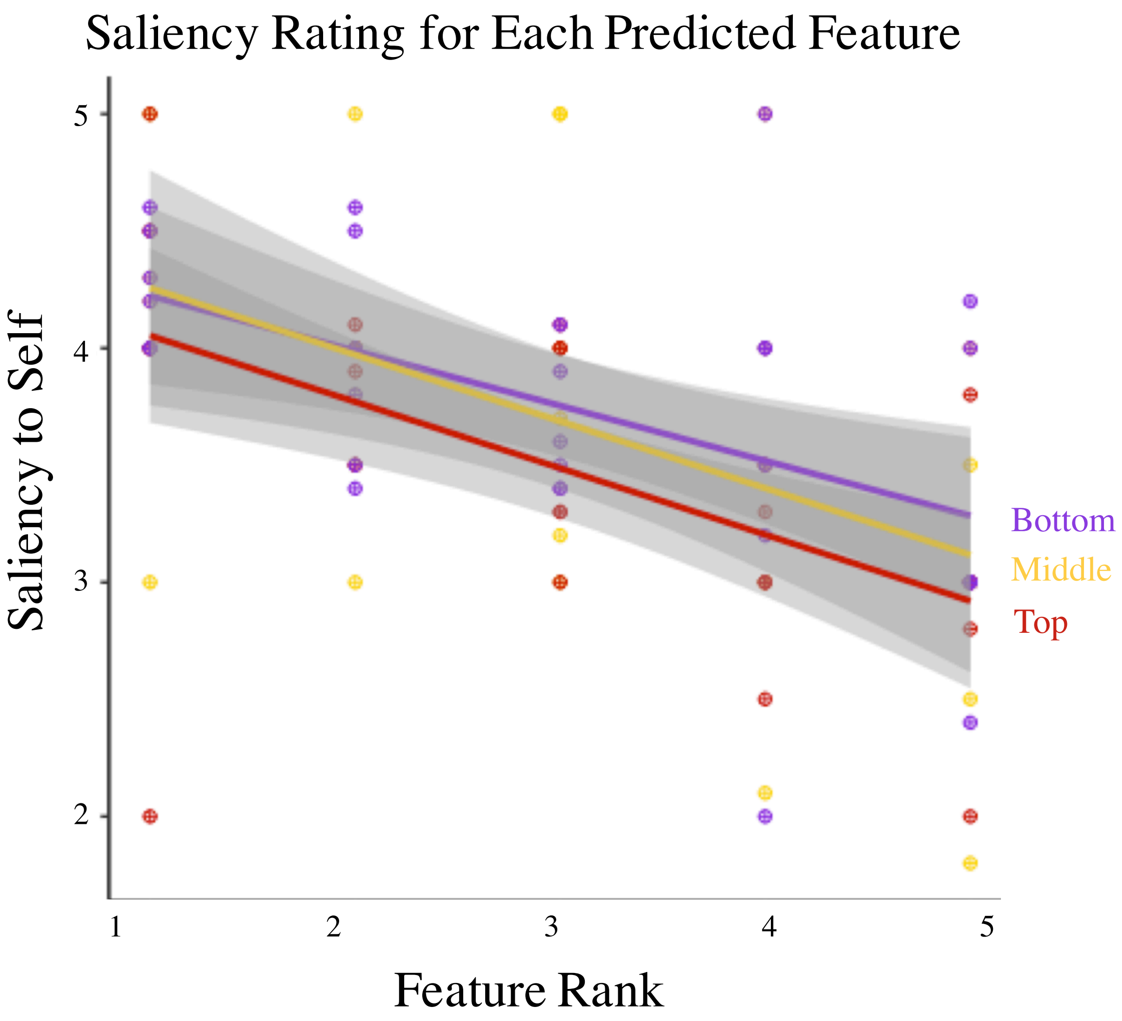


Figure 4.7. Regression of predicted saliency and saliency to self in Experiment 1a.

Both qualitative and quantitative statistical analyses for this experiment were done post-hoc. To ensure the validity of our findings, we conducted two follow up experiments with slight modifications with a new set of participants, and analyzed the data following similar procedures and an identical data analysis.

4.3. Experiment 1b

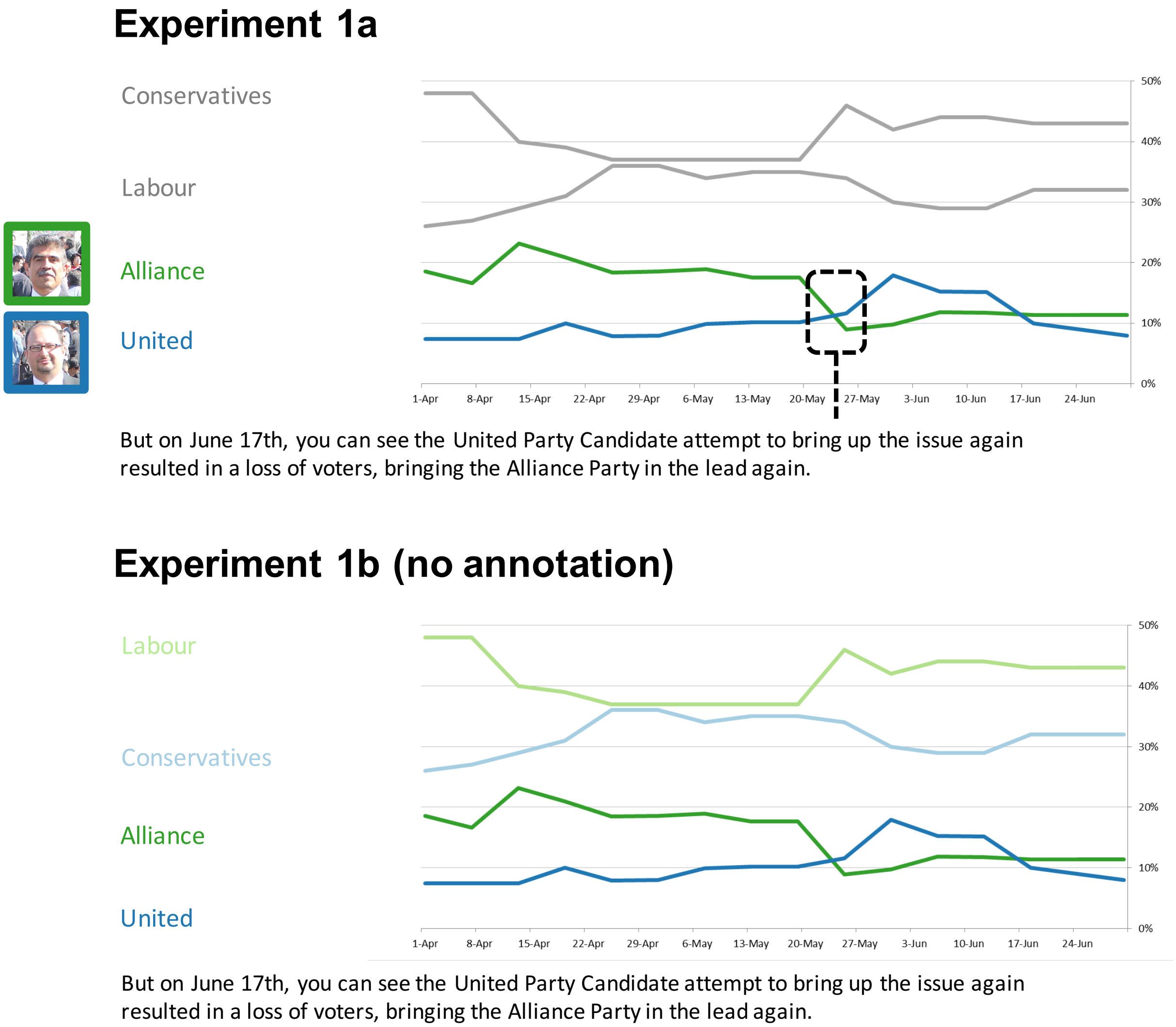
In Experiment 1a, participants were told the story and then shown a graph visually highlighting the story content before they made their predictions. Experiment 1b hoped to tease apart the priming effect of the visual annotations and that of the story by only including the story and removing the graph visual highlighting the story. The procedures and data analyses of Experiment 1b were identical to that of Experiment 1a, except we removed the feature cue after viewing the story (see Figure 4.8). The participants read the story and were presented the same unannotated line graph to draw and predict what other uninformed viewers would see. I hypothesize that even without the visual cue the participants would be just as biased in predicting what other uninformed viewers would see, thinking they would see the same features as visually salient.

Twenty-nine Northwestern University students (23 women) participated in this experiment in exchange for course credits in an introductory psychology class or monetary payment. All participants were asked to bring corrective eyewear if needed.

Participants again referred to their freely identified salient feature drawings and matched them with the five features mentioned above. Among the five pre-determined features, 66% matched with participants' freely identified salient feature drawings, and 78% matched if we only look at the top three predictions.

4.3.1. Quantitative Results

The Wilcoxon Signed-Rank Test Kerby (2014) indicates that the overall primed feature ranks, Wilcoxon mean score = 83.15, rank mean = 2.48, were significantly higher compared to the overall not primed feature ranks, Wilcoxon mean score = 67.98, rank mean = 1.79, $Z = 2.13$,



Experiment 1a

But on June 17th, you can see the United Party Candidate attempt to bring up the issue again resulted in a loss of voters, bringing the Alliance Party in the lead again.

Experiment 1b (no annotation)

But on June 17th, you can see the United Party Candidate attempt to bring up the issue again resulted in a loss of voters, bringing the Alliance Party in the lead again.

Figure 4.8. Comparison between Experiment 1a and 1b annotations.

$p = 0.035$. Primed features were given higher saliency rankings and were predicted to be more visually salient to other viewers than not primed features, even without visual annotations.

Inspecting the grey lines in the right column of Figure 4.9, we again see that some participants did not rate the primed features as more visually salient. This might mean that these participants were relatively immune to the curse of knowledge. Compared to Experiment 1a, we see that by taking away the visual annotations, the curse of knowledge effect weakened and the number of people might be immune to the curse of knowledge increased.

We also observed an interesting change in the middle prime saliency prediction from Experiment 1a to 1b, such that the participants in 1a who were primed with the middle feature rated it slightly more visually salient than participants in 1b. The middle feature â- the mirroring pattern of the two green lines, are more spatially separated than the top and bottom features. Since the participants in 1b only received a story prime without the visual annotation, the more spatially separated middle feature may have become harder for them to see compared to the participants in 1a who were shown clear visual annotations of this spatially separated middle feature. We speculate that while background story and visual annotation both contribute to the curse of knowledge, as shown in Experiment 1a and 1b, for spatially separated features, the visual annotation may play a more influential role in creating a curse of knowledge effect.

4.3.1.1. Salience Prediction Ranking. We found a significant relation ($r_s = 0.31$, $p < 0.01$) using Spearman's Correlation between the predicted salience ranking of features for other naïve observers and the self-rated salience of these features, see Figure 4.10. This indicates that even without the visual annotation cue, the more visually salient a feature participants rated to themselves, the more visually salient they think the features were to a naïve viewer.

4.3.2. Discussion

We observed a statistically weaker curse of knowledge effect without the visual annotations in the present experiment. However, most participants nonetheless reported features primed by the story to be more visually salient than features not primed by the story, even without visual annotations. This suggests that only having the background knowledge, without any visual

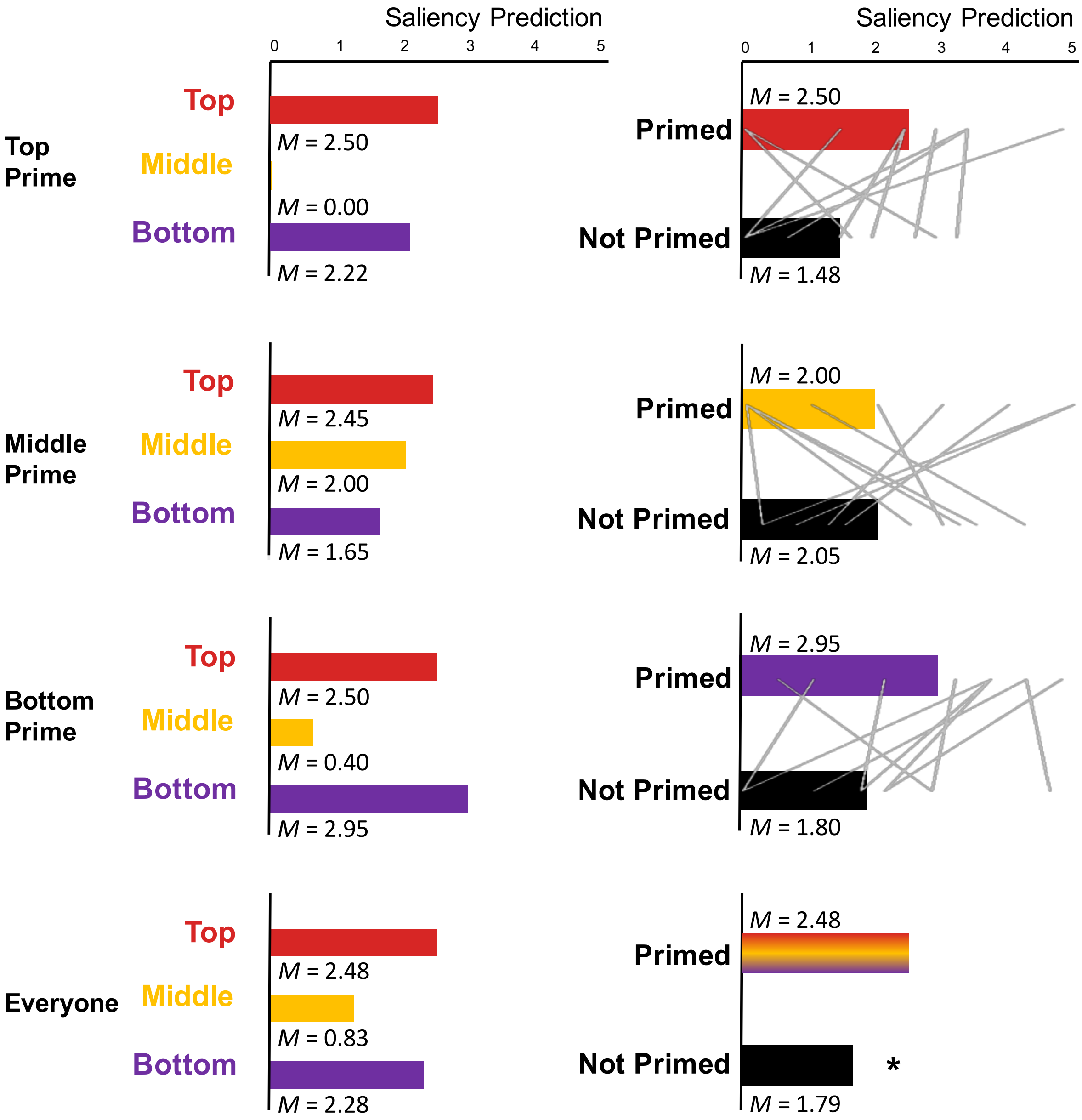


Figure 4.9. Saliency prediction ranking for Experiment 1b.

annotation cues, is still enough to bias people to predict that other naïve graph viewers would see features primed by the story as more visually salient.

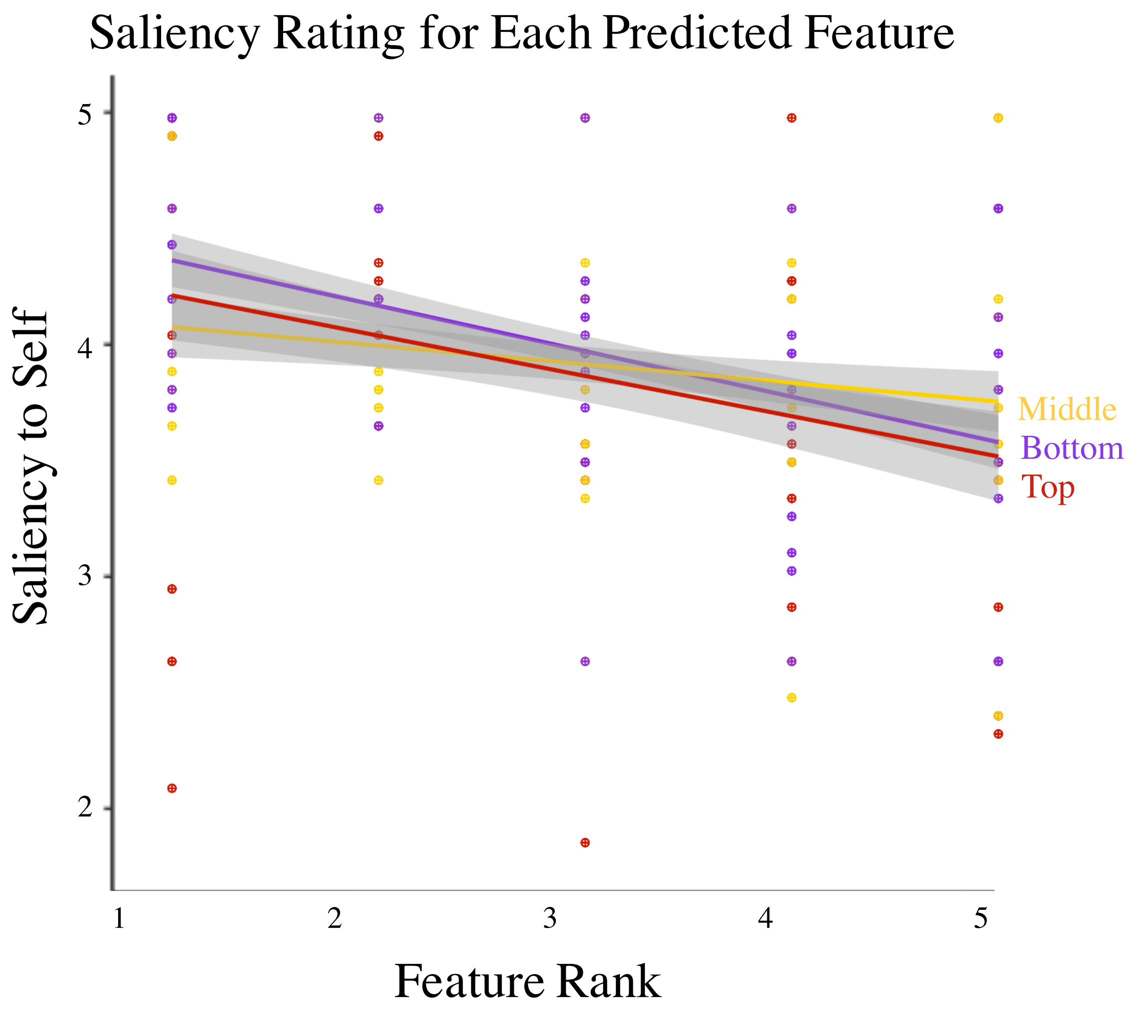


Figure 4.10. Regression of predicted saliency and saliency to self 1b.

4.4. Experiment 1c

We conducted a third follow up experiment on a new set of participants and analyzed the data following the same procedures and data analysis method. Since Experiment 1a and 1b did not specify in the instructions what types of features the participants should be predicting or drawing, we designed Experiment 1c with more specific instructions to maximize the amount of matching between freely identified salient features and the five pre-determined features. This experiment 1c also serves as a conceptual replication of Experiment 1a and 1b.

Previously, participants predicted features with no specific restrictions or requirements, leading some to pick out features irrelevant to the study (e.g., one participant circled the entire graph as being visually salient, another circled the y-axis, see Figure 4.5). To decrease such uninterpretable responses in the feature free-identification stage, participants were instructed to only describe features that involved two or more parties.

Twenty-one Northwestern University students (10 women) participated in this experiment in exchange for course credits in an introductory psychology class. All participants were asked to bring corrective eyewear if needed.

Among the five pre-determined features, 64% matched with participants' freely identified salient feature drawings, which is a 16% increase from Experiment 1a. When we look at the top three predictions, 83% matched in Experiment 1c, which is a 27% increase compared to Experiment 1a and a 5% increase compared to Experiment 1b.

4.4.1. Quantitative Results

The Wilcoxon Signed-Rank test indicated that the overall primed feature ranks, Wilcoxon mean score = 65.93, rank mean = 3.26, were statistically significantly higher than the overall not primed feature ranks, Wilcoxon mean score = 46.54, rank mean = 1.80, $Z = 3.17$, $p < 0.01$. The descriptive statistics are shown in Figure 4.11. This result is consistent with the Experiment 1a and 1b such that the primed features were given higher saliency rankings and were predicted to be more visually salient to other naïve viewers than unprimed features.

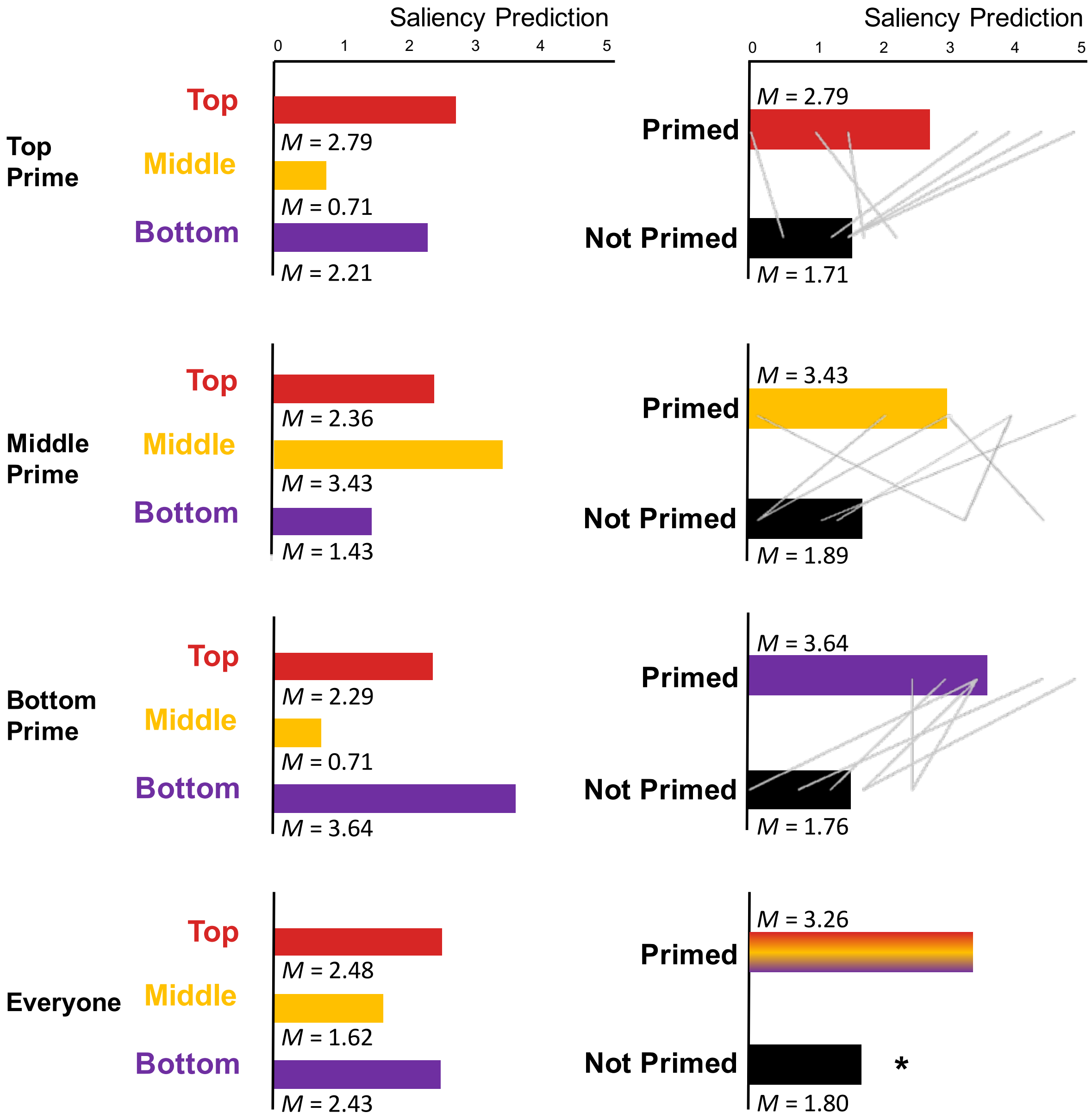


Figure 4.11. Saliency prediction ranking for Experiment 1c.

Spearman's Correlation again showed a moderately strong relationship ($r_s = 0.43$, $p < 0.001$) between the self-rated salience of a feature, and the predicted salience rating for other uninformed graph viewers, shown in Figure 4.12.

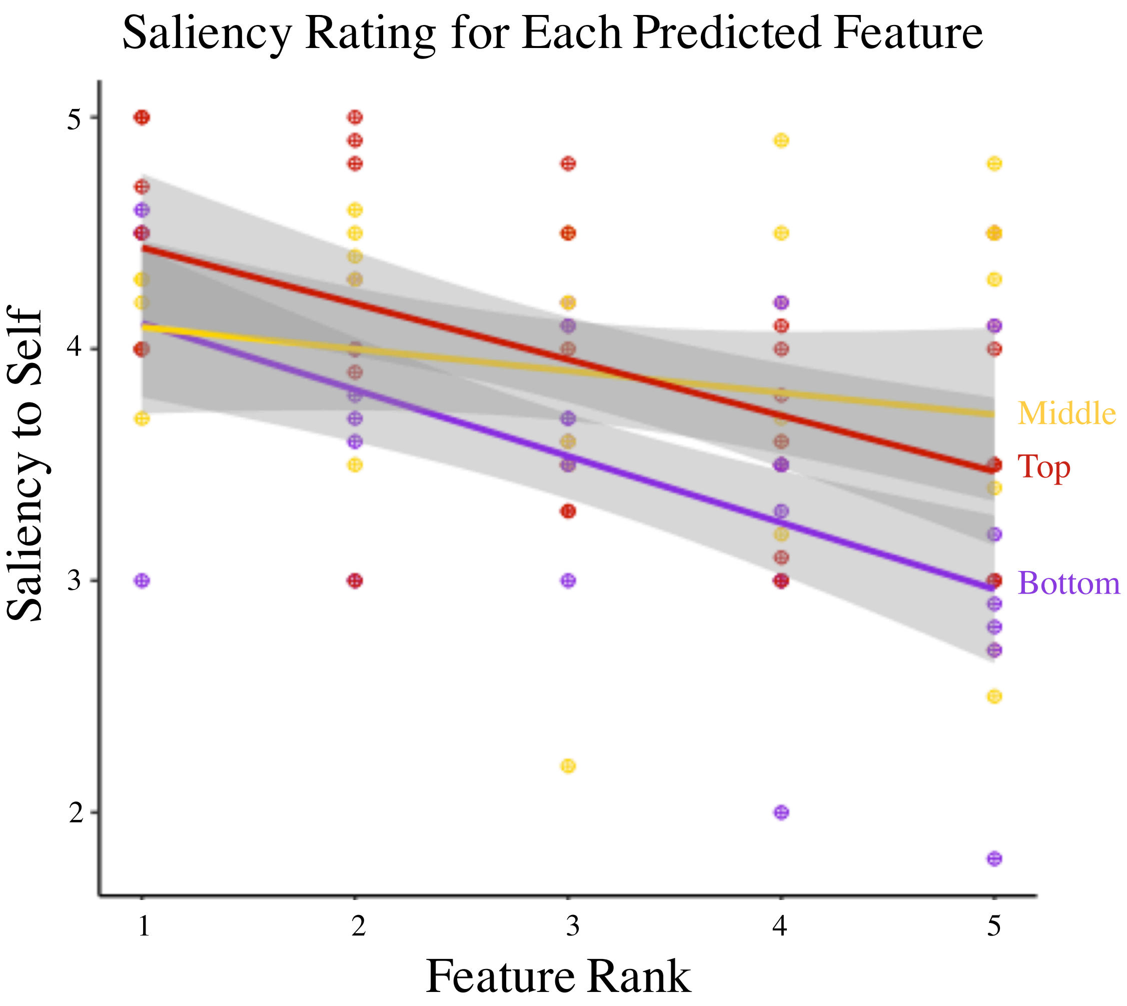


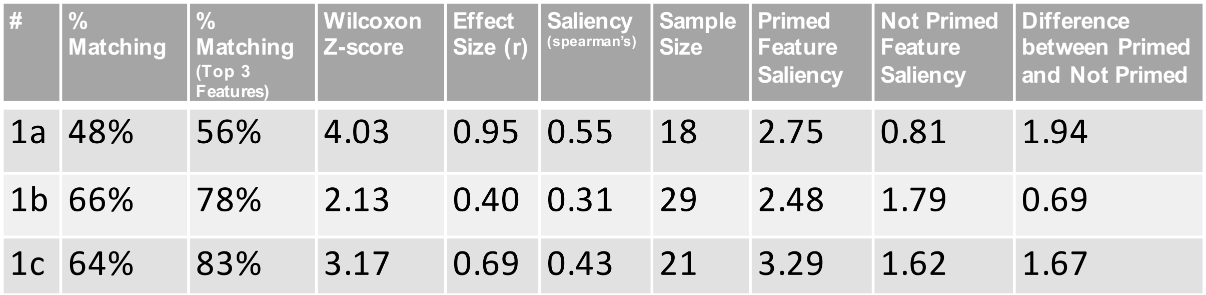
Figure 4.12. Regression of predicted saliency and saliency to self 1c.

4.4.2. Discussion

Both Experiment 1a, where we primed participants with both a story and visual annotations, and Experiment 1b, where we took away the visual annotations, show a curse of knowledge effect where people predict features they themselves see as visually salient to also be salient to naïve viewers. This effect decreased by half in Experiment 1b when we took away the visual annotations, suggesting that both background story and visual annotations contributed to this effect, as shown in Figure 4.13.

Comparing Experiment 1a and 1c (where we gave the participants more specific instructions on what types of features and patterns to identify), we observed a higher number of matches between the freely identified features and the pre-determined features. We also see that overall feature saliency for primed and not primed features increased from Experiment 1a to 1c. This instruction phrasing seems to have strengthened the curse of knowledge effect. There was also a decrease in effect size from Experiment 1a to 1c, though not statistically robust. But it is also possible that, by asking participants to predict features that include two or more parties in Experiment 1c, participants were able to match more of their own predictions to the pre-determined features (which involves two parties). This may have increased the likelihood of unprimed features to be included in the participants' predictions, which in turn increased the saliency rating of not primed features and decreased the differences between primed and not primed feature saliency ratings, resulting in a smaller effect size for Experiment 1c.

A comparison of the "Everyone" row across Figure 4.6, 4.9 and 4.11 shows that people gave similar saliency ratings to top and bottom features overall, but slightly lower ratings for the middle features. We speculate this to be due to the middle feature â the mirroring of the two green lines being more spatially separated than the top and bottom features, which makes the middle feature a more difficult feature to see without annotation. Participants still rated this less salient middle feature as the most visually salient to both themselves and other people when they read a story highlighting this feature, supporting the hypothesis that participants predict features they see as more visually salient also visually salient to an uninformed viewer, and that they rated the feature predicted to be the most/least visually salient to an uninformed viewer to also be the most/least visually salient to themselves.



#	% Matching	% Matching (Top 3 Features)	Wilcoxon Z-score	Effect Size (r)	Saliency (spearman's)	Sample Size	Primed Feature Saliency	Not Primed Feature Saliency	Difference between Primed and Not Primed
1a	48%	56%	4.03	0.95	0.55	18	2.75	0.81	1.94
1b	66%	78%	2.13	0.40	0.31	29	2.48	1.79	0.69
1c	64%	83%	3.17	0.69	0.43	21	3.29	1.62	1.67

Figure 4.13. Comparison across all three Line Graph Experiments.

4.5. Experiment 2

To evaluate the generalizability of this specific curse of knowledge effect, we replicated our findings using a novel type of graph, and a new story.

Seventeen Northwestern University students (9 women) participated in this experiment in exchange for course credits in an introductory psychology class. All participants were asked to bring corrective eyewear if needed.

This bar graph experiment followed the same within-subject design and experimental procedures as the line graph experiments. Participants were randomly assigned to read one of three different backstories describing events leading to a presidential election between the Liberal and the Conservative parties.

After reading the story, they were shown public polling data highlighting a key aspect of public opinion that eventually led to the victory of the winning candidate. They were asked to freely identify top five features they predict to be visually salient to a naïve graph viewer on an unannotated graph (Figure 4.14), rank the saliency of these predicted features to themselves, and match the freely identified predictions to five pre-determined features, as shown in Figure 4.15.

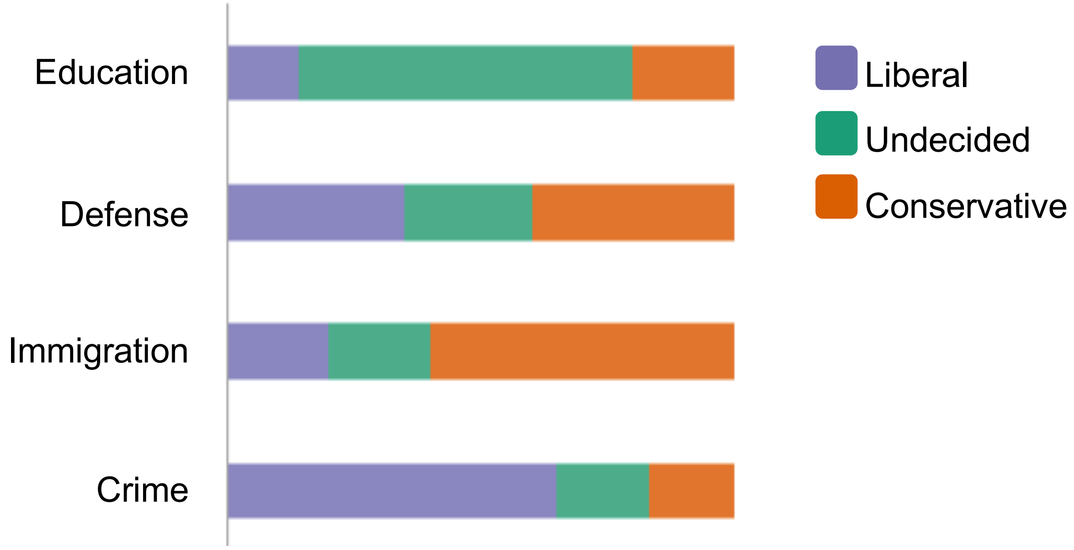


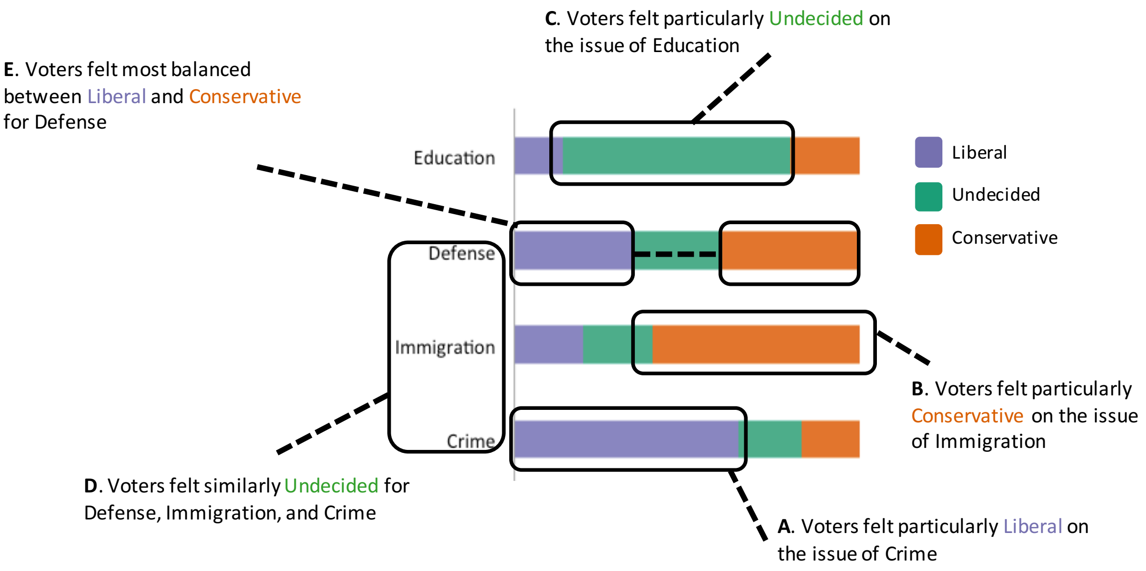
Figure 4.14. Unannotated bar graph in bar graph experiment.

Figure 4.14 shows an unannotated version of the bar graph the participants freely drew their predictions on. The stacked bar represents how people with different political stances (e.g., Liberal vs Conservative) view the topics listed, such as education. The length of the bars represents the number of voters.

We pre-determined five features on this graph, as shown in Figure 4.15. The graph and the features are balanced such that from the top to bottom, the four issues the public polls demonstrate correspond to education, defense, immigration, and crime issues. In the top two bars, the areas of purple and orange bars are the same. Between the bottom two bars, the area of the orange bar on the immigration issues equals the area of the purple bar on the crime issue. Similarly, the area of the purple bar on the immigration equals the area of the orange bar on the crime issue. Additionally, the area of the two undecided bars are equal. Overall, the total area of purple bars equals the total area of the orange bars.

Critically, they were told that this unannotated graph (Figure 4.14) was all that the uninformed graph viewers had access to, and that there was no background story provided for the uninformed graph viewers. Also, paper copies of this unannotated graph were provided to the participants to mark down their predictions, prior to matching their predicted features to the five pre-determined features, as shown in Figure 4.15.

Please rank the following statements (A, B, C, D, and E) to match your written ranking predictions, as best as you can. If you didn't write something down, select N/A.



C. Voters felt particularly Undecided on the issue of Education

E. Voters felt most balanced between Liberal and Conservative for Defense

B. Voters felt particularly Conservative on the issue of Immigration

D. Voters felt similarly Undecided for Defense, Immigration, and Crime

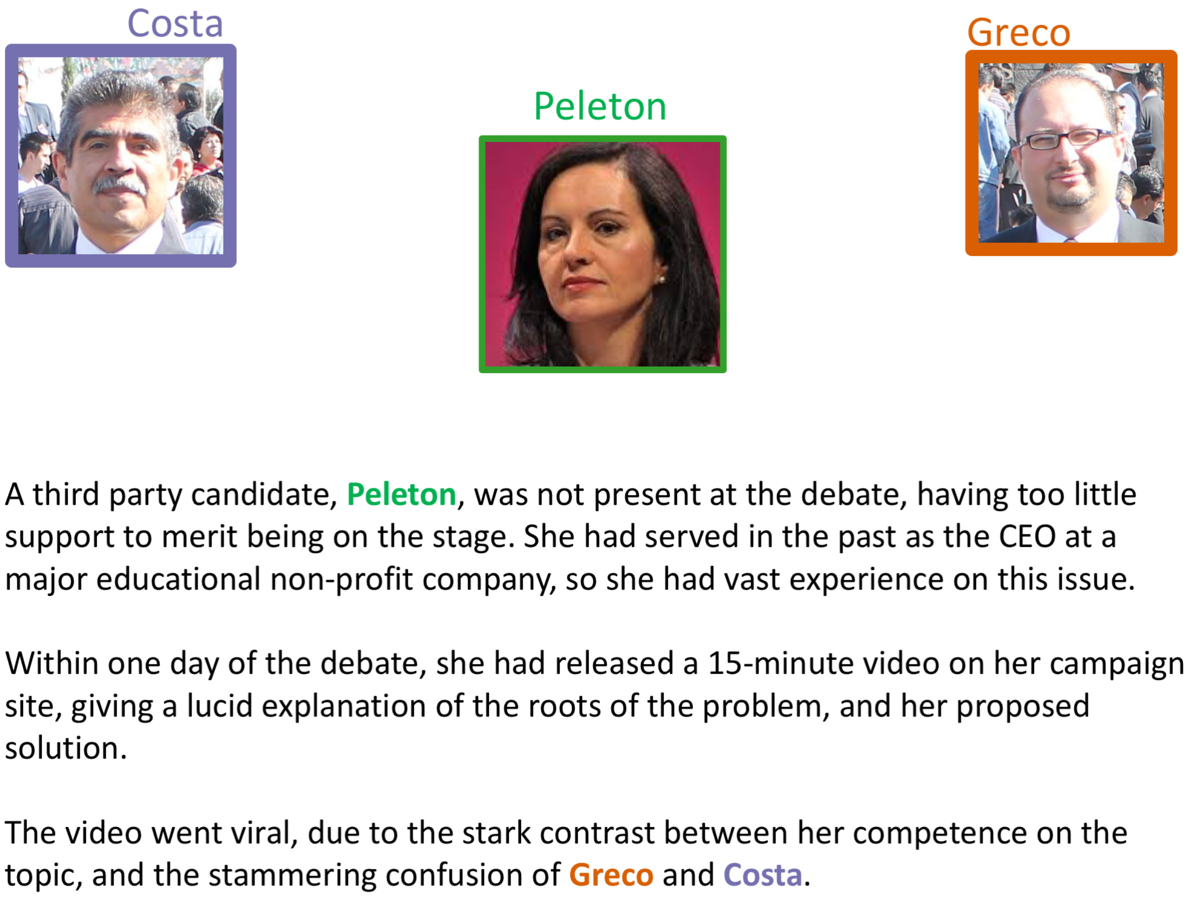
A. Voters felt particularly Liberal on the issue of Crime

Figure 4.15. Matching five pre-determined features.

4.5.0.1. Story. There are three versions of the story in this experiment: crime, immigration and education, Figure 4.16 shows a snapshot of the stories. The crime story was a story about police brutality toward specific minority groups. The Conservative Party leader supported the police, brazenly stating that people in the minority group deserved such punishment, which was an unpopular position to take. Meanwhile, the Liberal Party advocated for reform in police departments and better treatment of suspected criminals. Participants saw graphs that highlighted the majority's Liberal public opinion of crime, explaining it as the reason behind the Liberal Party's victory, as shown in left most column in Figure 4.17.

The immigration story described a terrorist attack on the country's bus system two weeks before the election. The Conservative candidate had predicted in the past that immigrants posed a threat to the country's citizens. There was no information whether terrorists were immigrants, but the public was too frightened to care. While the Liberal candidate had laughed at his opponent for being too overly paranoid, the frightened public supported the Conservative view on immigration, leading to the victory of the Conservative candidate at the election. The graph the participants saw corresponded to the story highlighting the majority's Conservative public opinion on immigration, explaining it as the reason behind the Conservative Party's victory, shown in the right-most column in Figure 4.17.

The education story described a debate between the Liberal and Conservative Parties on the country's education system. They were told that the country had not been performing well compared to other EU countries academically. Neither candidate could come up with a clear vision on how to solve this, and the public was shocked at their incompetence. This opened an opportunity for a third candidate, who was an expert on education (as well as being female, a salient characteristic), in the election. The graph corresponded to the story by highlighting the fact that most people in the country had been undecided (neither Liberal nor Conservatives) on the issue of education, opening the opportunity for the third candidate, shown in the middle column in Figure 4.17.


Costa
Peleton
Greco

A third party candidate, Peleton, was not present at the debate, having too little support to merit being on the stage. She had served in the past as the CEO at a major educational non-profit company, so she had vast experience on this issue.

Within one day of the debate, she had released a 15-minute video on her campaign site, giving a lucid explanation of the roots of the problem, and her proposed solution.

The video went viral, due to the stark contrast between her competence on the topic, and the stammering confusion of Greco and Costa.

Figure 4.16. Snap-shot of bar graph experiment story.

4.5.0.2. Matching Features. The participants referred to their paper copies of the unannotated graph and matched their own predictions to five pre-determined features, shown Figure 4.15.

Feature A corresponds to the feature reflected in the crime story, highlighting the purple section in the bottom bar representing public opinion on crime issues. Feature B corresponds to the feature reflected in the immigration story, highlighting the orange section in the second to bottom bar representing public opinion on immigration issues. Feature C corresponds to the feature reflected in the education story, highlighting the green section in the top bar on public opinion on education issues. These remaining features (DE) were not directly reflected in any stories, serving as “fillers”. Feature D highlighted how the public was equally undecided on the issue of defense, immigration, and crime. Feature E highlighted how the defense issue had equal Conservative and Liberal support.

Among the five pre-determined features, 82% matched with participants’ freely identified salient feature drawings, and 94% matched if we only look at the top two predictions.

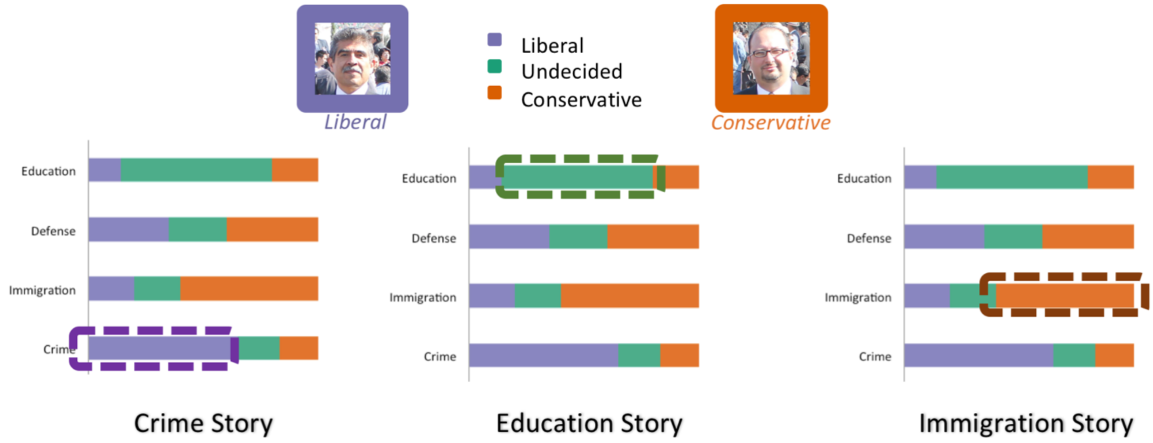


Figure 4.17. Highlighted feature for three story versions.

4.6. Qualitative Results

Examining what the participants marked down on their physical copies of the unannotated graph, we find observable differences in the order of feature predictions for the three story versions.

In Figure 4.18, each column represents the responses of participants who read that version of the story. The top row shows the highlighted feature in that story version. Underneath, the first and second rows show the most and second most visually salient predicted features. There are participants who indicated multiple features to be salient for each of the five predictions, therefore the numbers on the graph represent the number of times the highlighted feature was chosen to be visually salient to a naïve viewer. Because the predictions can be overlapped visually across all participants, the darker the shading of a highlighted feature, the more frequently it was chosen to be visually salient to a naïve viewer.

Overall, the participants generally indicated features primed by the story version they read as what others would see as visually salient. Figure 4.18 compares the story versions and their respectively highlighted features to the overall predictions participants made, supporting our hypothesis.

4.6.1. Quantitative Results

We analyzed our data using the same method and criteria as the line graph experiments 1a, 1b and 1c. Comparing the feature highlighted in the story (primed feature) with the average rankings of all the features not explicitly highlighted in the story (unprimed features) as shown in Figure 4.19, across all three stories, descriptive statistics show that participants predicted primed features to be more visually salient than not primed features to naïve viewers. For example, for participants who read the crime story, feature A (the crime feature) was ranked to be more visually salient to naïve viewers than not primed features BCDE.

The non-parametric Wilcoxon Signed-Rank Test indicated that the overall primed feature ranks, Wilcoxon score mean $= 34.74$, rank mean $= 4.29$, were statistically significantly higher than the overall not primed feature ranks, Wilcoxon score man $= 21.63$, rank mean $= 3.38$, $Z = 3.09$, $p < 0.01$. This result adds to the line experiments 1a, 1b and 1c, supporting that the features depicted in the story were given higher priority rankings and predicted to be more visually salient to naïve viewers than features not depicted by the story.

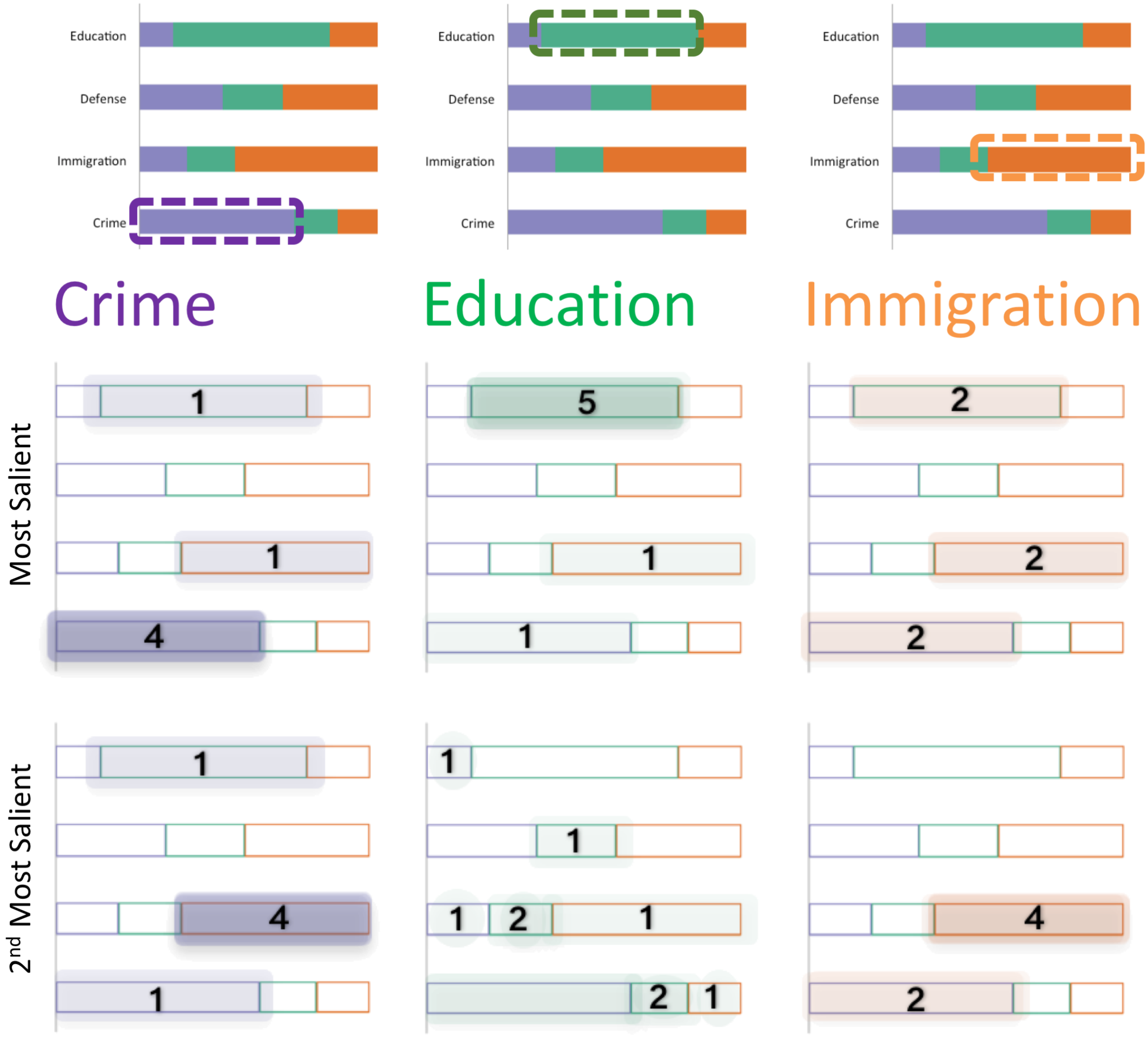


Figure 4.18. Qualitative result of bar graph experiment. Heat map shows participants indicated the primed features to be more visually salient to naïve viewers than other features.

We also found strong, significant correlation between predicted features' saliency ranking and self-rated saliency of these features using Spearman's Correlation, $r_s = 0.65$, $p < 0.001$, indicating that participants predicted features which were visually salient to themselves to also be salient to naïve viewers, consistent with previous experiments, see Figure 4.20.

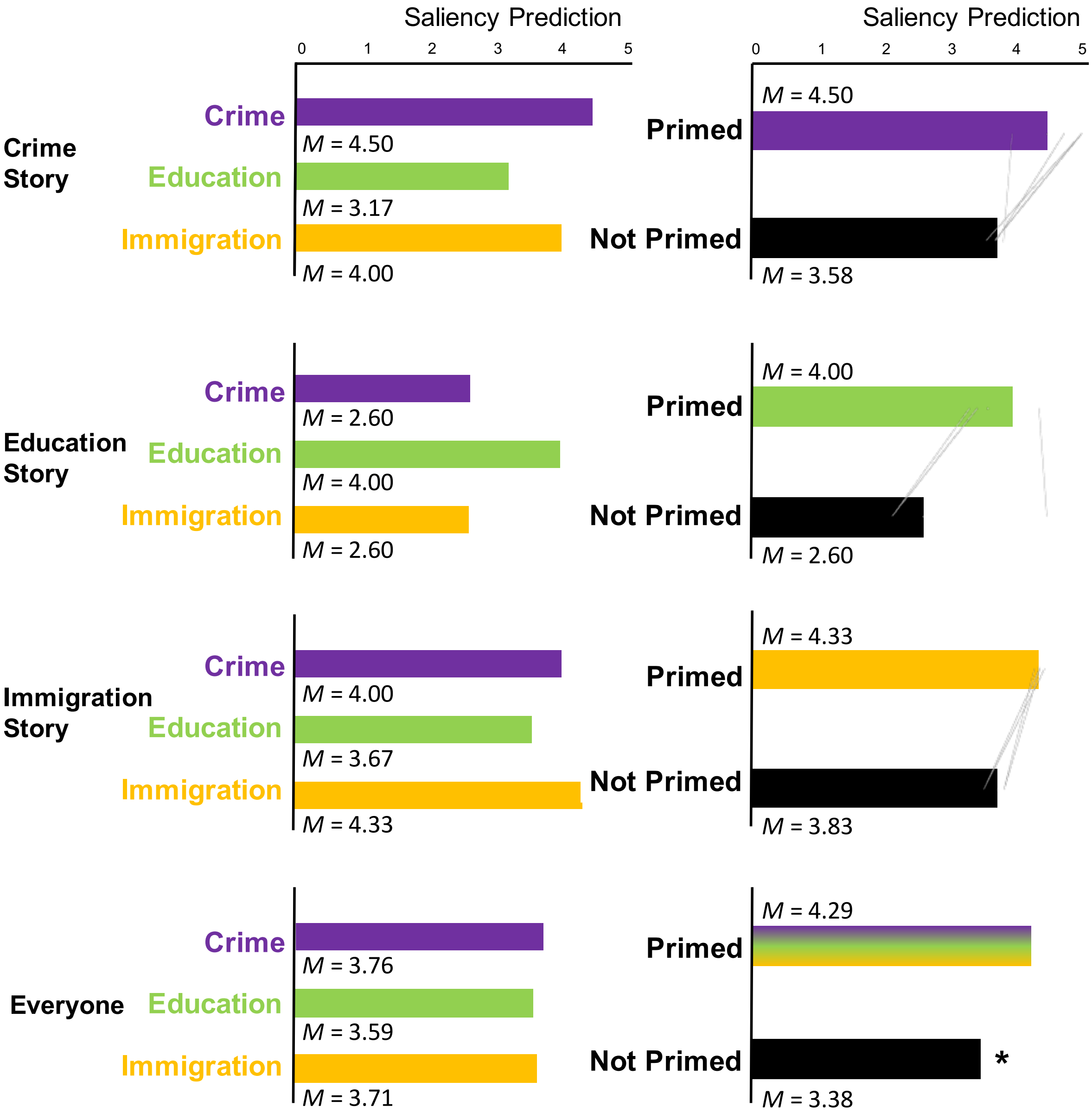


Figure 4.19. Prediction rankings break down by story version and primed/not primed in bar graph experiment. The grey lines represent individual participant responses.

4.6.2. Discussion

The significant differences between saliency rankings of the primed and not primed features reveals the curse of knowledge bias in viewing bar graphs. This result is consistent with the line graph experiments, showing that this curse of knowledge can be generalized to bar graphs with different data sets and visual features.

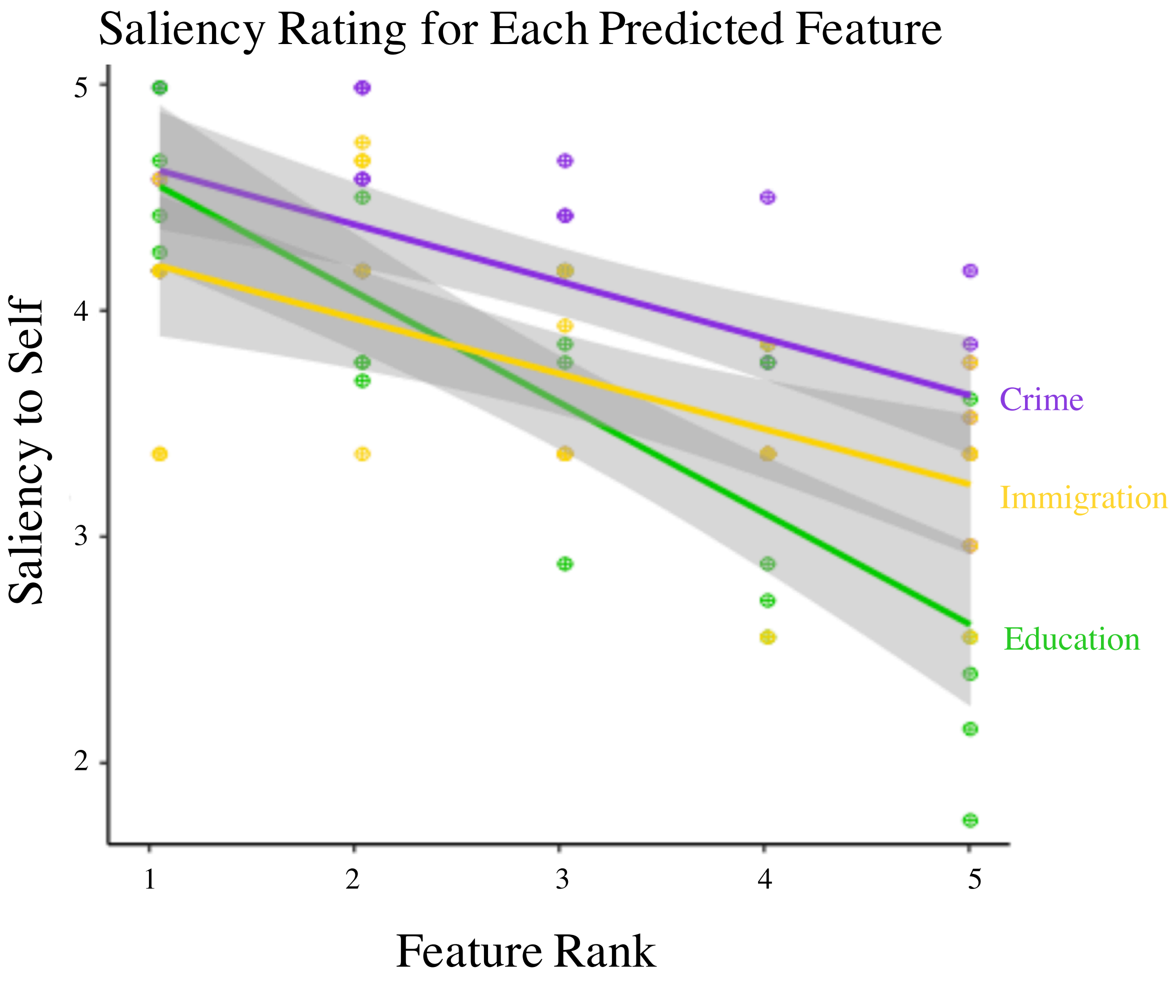


Figure 4.20. Regression of predicted saliency and saliency to self.

4.7. Conclusion

Across four experiments and two types of graphs, it is clear that participants are susceptible to a ‘curse of knowledge’ when asked to simulate what others would see in a visualization. When a participant was told one of three possible background stories, each of which made a particular pattern within a graph visually salient to them, that participant assumed that naïve viewers would also see the same pattern as visually salient. This effect occurred despite explicit instructions to ignore what they knew, and to take a naïve perspective. To our knowledge, this is the first empirical demonstration of the curse of knowledge in the realm of data visualization, and even in the broader realm of visual perception.

This result joins other recent explorations of the influence of perceptual and cognitive biases on interpretations of patterns in data visualizations, many of which cannot be easily mitigated (Bateman et al., 2010; Borgo et al., 2012; Dimara et al., 2017; Herman et al., 2000; Michal and Franconeri, 2017; Moere et al., 2012; Pandey et al., 2014; Segel and Heer, 2010). Some of this research has begun to explore visual designs and interactive decision-making environments that mitigate these biases (Dimara et al., 2019).

This experiment simulates the real-world context of focusing on a particular pattern out of multiple possibilities. But if the visualization contained a single dominantly salient pattern (e.g., a downward trend among upward trends), the dominance of that pattern could hide any effect of the curse of knowledge. We attempted to balance the salience of the alternative patterns and the data suggests that these patterns were roughly balanced, according to the ‘Everyone’ section of Figure 4.6, 4.9, and 4.11, which collapse over the instructional primes (though there is a trend in Experiment 1a for ‘bottom’ to be more salient). Similarly, we used a set of stories, party names, and pictures, that we hoped would maintain a balance across the experiments. For example, we picked a salient female candidate, and a top position in the visualization, to balance out the presumed lower salience of the green ‘Education’ bar, which was in the horizontal-center of the graph. Such balancing is critical for finding any experimental effect of a single factor among multiple other factors that potentially compete. However, we have not tested the baseline saliency of the graph in the absence of story primes. Future research could test the robustness of this bias with less balanced visualizations, or more complex visualizations, to more closely emulate real world situations and further explore how stronger baseline salience differences might prevent the curse of knowledge bias.

We recognize that there are many kinds of visual data communication across many types of conversation partners. Communication could be between the creator of the visualization and an audience listening to the creator’s story, or between people who did not create a visualization, but are sharing their interpretations with each other. This experiment focused on the later situation where the experts did not create the visualizations themselves. Future research could investigate if the curse of knowledge persists if the communication is between the visualization designer and a naïve audience, perhaps even in more realistic situations instead of lab simulations. We predict the curse to be stronger in these conditions as visualization creators would have richer expertise and deeper understanding of the data pattern and trends, making it even more difficult for them to separate their knowledge with that of their audience.

While most participants predicted primed features to be more visually salient to uninformed others, some participants did not. Why are some people immune to the curse of knowledge, at least for this case study? Are some people simply better at simulating the thoughts of others, or do they use different strategies? The curse of knowledge can manifest not just from differences in perceived salience, as tested here, but by memorability, context, or impact of the data. Future research could investigate other consequences of the curse and evaluate different methods to discover the manifestation of the curse of knowledge. While these question veers more closely toward the psychology literature (see (Epley and Waytz, 2010; Zhou et al., 2017) on discussions of strategy differences in inferring and simulating the perspectives of others), understanding the underlying difference could lead to prescriptions for mitigating the bias.

CHAPTER 5

The Illusion of Causality

Visualization designs affect decisions. Imagine coming across a piece of BBC news, as shown in Figure 5.1, showing that the number of crimes in London rises with temperature. It can be easy for viewers to conclude that warmer temperature causes violent crimes (News, 2018; Matute et al., 2015; Kahneman, 2011).

Concluding causality from the visualized data alone is misguided. We can only establish a correlation - the tendency of two variables changing together - between temperature and crime rate because it is possible that other factors not shown on the graph caused the difference in the number of violent crimes. For example, when the temperature gets warmer, more people go outside, more crimes may happen overall, and thus more violent crimes. If the amount of people outside is kept constant, decreasing temperature would not likely lower crime rates. While the variables illustrated are linked, they are not necessarily causally linked. Yet, people routinely see causal relationships in data.

Confusing correlation with causation is a ubiquitous decision-making error. Just because two factors are correlated (i.e., they tend to co-occur together), it does not mean that one is causing the other. A large portion of work in economics, education, epidemiology, psychology and public health involves analyzing correlations in observed data, which cannot definitively establish causation (Robins et al., 2000). Researchers and journalists can sometimes exaggerate causal implications from these results, making it even more difficult for people to decide what kind of conclusions are sound (Shiffrin, 2016; Sumner et al., 2014). This can pave way for misunderstanding of correlation and causation (Halpern, 1998; Shaklee and Elek, 1988; Koslowski, 1996), potentially having detrimental impact. When researchers or journalists misinterpret or misrepresent correlation for causation, for example, the general public may be misled into thinking correlated factors, such as time of getting vaccinated and time of autism diagnosis, or national debt and GDP growth, are also causally related (Dixon and Clarke, 2013; Reinhart and Rogoff, 2010).

It is difficult to distinguish causation from correlation (Rothman, 2012). Even for people who learned ‘correlation is not causation’ with classroom examples, it could still be challenging to apply their learning to new contexts (Shtulman and Valcarcel, 2012; Rhodes et al., 2014). Because establishing causal inference is complex, even trained scientists can sometimes struggle with correlation and causation (Halpern, 1998). We are interested in whether a simple change in the visualization design can reduce unwarranted conclusions of causality.

Although many have looked at the effect of visualization designs on *perceptual* analytic tasks such as determining anomalies or estimating data trends (Saket et al., 2018; Correll and Gleicher, 2014; Croxton and Stryker, 1927; Eells, 1926; Harrison et al., 2014; Spence and Lewandowsky, 1991; Cleveland and McGill, 1984b; Kay et al., 2016), researchers have only begun to explore the effect of visualization design on *cognitive reasoning* tasks, such as understanding uncertainty (Hullman et al., 2018; Kale et al., 2019), persuading attitude or belief change (Kim et al., 2019; Pandey et al., 2014) or eliciting empathy (Boy et al., 2017). Previous work has demonstrated visualization designs could influence data interpretation. For example, many people conclude “on average, Dutch are taller than Americans” from a bar graph visualizing the height of Americans and Dutch, but when the same information is visualized

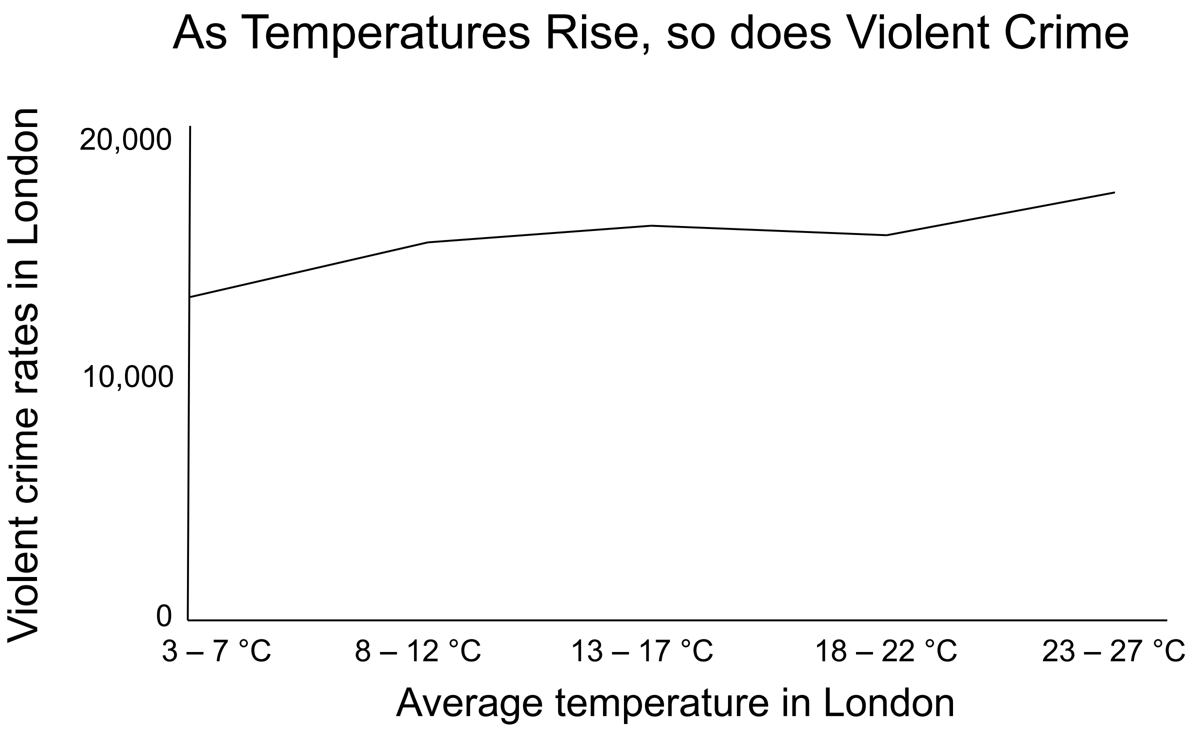


Figure 5.1. Recreation of BBC news article figure, "Heatwave: Is there more crime in hot weather?"(News, 2018)

with a line graph, people are more likely to conclude “people get taller as they become more Dutch.”(Zacks and Tversky, 1999). We suspect visualization designs can also afford different cognitive reasoning routines in data, triggering perceived causal links more or less strongly in data.

What types of visual formats are commonly used to present correlated data? Bar graphs, line graphs and scatter plots are common ways to depict correlated data in media (News, 2018; Guo, 2016), alongside text, as shown in Figure 5.1 and Figure 5.2. We investigate how bar graphs, line graphs, scatter plots and text influence causal reasoning of data.

Research on perceptions of causality indicates that they can be context-dependent, in addition to being visualization design-dependent. When the evidence presented aligns with people’s

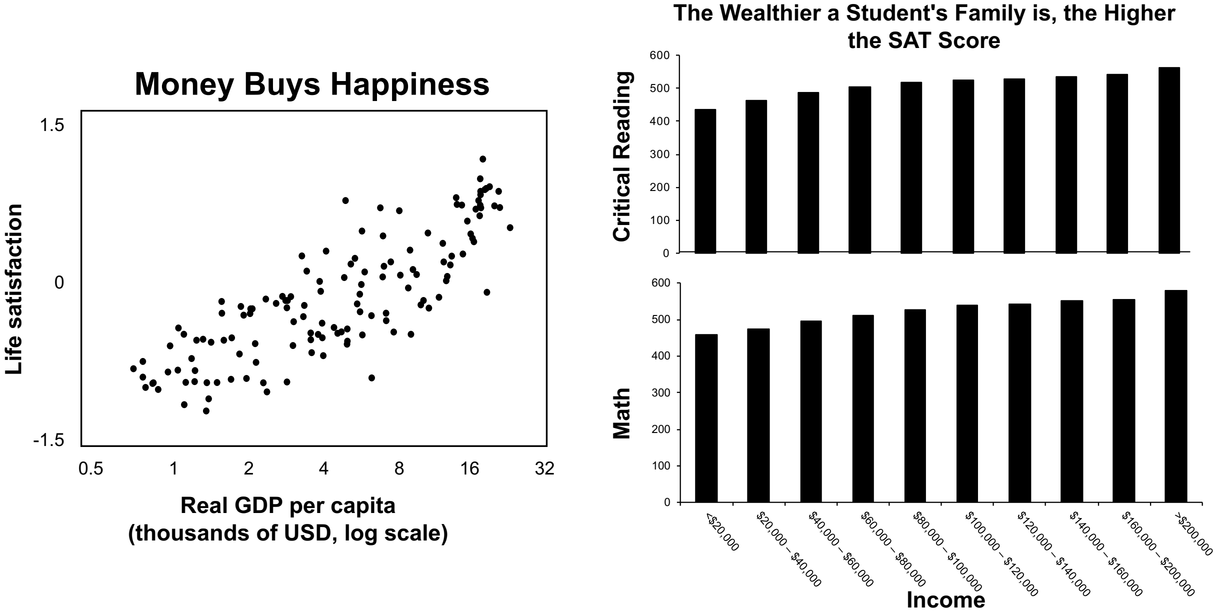


Figure 5.2. Recreation of NPR article "Money Buys Happiness," showing a correlation between GDP and life satisfaction (Radio, 2011) and of the Washington Post news article “Researchers have debunked one of our most basic assumptions about how the world works” showing a correlation between income and SAT scores(Guo, 2016).

prior experience, emotional response or beliefs, they become more likely to judge the evidence as sound Shah et al. (2017). People often perceive high causality when they judge the evidence as sound and stop thinking through other possible explanations Kahneman (2011). Prior work suggests that persuasiveness of visualized data depends on both context (does the topic align with the viewers’ prior beliefs?) and visualization designs (tabular design or bar graphs) Kim et al. (2017, 2019); Pandey et al. (2014). Thus we also examine the effect of context by testing a set of paired variables that vary in the plausibility of their causal link, which we establish in a pilot experiment.

The task people perform when viewing the visualizations may also influence the conclusions they draw. Our experiments consider two common tasks people perform when interacting with data. The first is a *judgment* task in which they decide whether they agree or disagree with the presented information. For example, media often present people with visualizations alongside text describing a correlational or a causal relation between depicted variables Bromme and Goldman (2014). In this scenario, information consumers have to decide how much they agree with the description based on the visualized data. Judgment tasks can be evaluated by comparing participant ratings of how much they agree with statement describing a correlation or a causation. The second is a *generative* task where people have to independently interpret a visualization to draw their own conclusions. One example is when a data analyst working to make sense of their data hoping to deliver a research report on the newest scientific findings. In this scenario, the data analyst has to actively interpret some visualizations and generate a conclusion. Generative tasks may shed more insights on how participants interpreted data and arrived at possible correlational/causal conclusions, but because they are open-ended, they tend to be more difficult to formally evaluate. In our pilot experiment, we asked participants to generate interpretations of data, then used their interpretations to develop a taxonomy to facilitate analysis of generative tasks in Experiment 1.

5.1. Pilot Experiment

Taking inspiration from the anecdotes of a set of local instructors of research methods and data analytics, we generated 19 potential variable pairs, from those with plausible causal relations to those with implausible causal relations. We conducted a pilot experiment to test the perceived correlation and causation of these variable pairs, and identified pairs within a range from low to high perceived correlation and causation for use in Experiment 1.

Specifically, we surveyed 21 participants for their perceived plausibility of correlational and causal relations of the 19 variable pairs through Qualtrics on Amazon's Mechanical Turk (MTurk) (Qualtrics, 2013). Participants viewed the 19 correlation and causation statement sets in random orders. For each pair, they first interpreted its message and justified their reasoning in a text box. This is the *generative* task. Then, on a separate page, they read a correlation statement and a causation statement, as shown in Table 5.1. The correlation statement accurately describes the relation between the depicted data variables, while the causation statement attributes causal relations to the depicted data variables. They gave a plausibility rating for each (0 = extremely implausible, 100 = extremely plausible). This task reflects the *judgment* tasks people would perform in real life.

The participants rated their perceived plausibility of both the correlation and causation statements. Table 5.1 shows the four contexts we picked with varying plausibility. These four context differed significantly in their perceived correlation and causation ratings, based on an analysis of variances, as shown in Figure 5.3. We visualized information using these four contexts in Experiment 1 to investigate the effect of visualization design on perceived causality.

5.1.1. Qualitative Coding: Interpretation Taxonomy

To provide a structured way of interpreting participants' statements in our experiments, we analyzed the freeform written response from the generative task in the pilot, in which participants drew conclusions from the information and justified their correlation and causation ratings, to create a taxonomy to characterize these conclusions in the experiment. We identified six dimensions that could help us characterize and evaluate the conclusions participants generated – whether the participant concluded correlation, concluded causation, mentioned third variables,

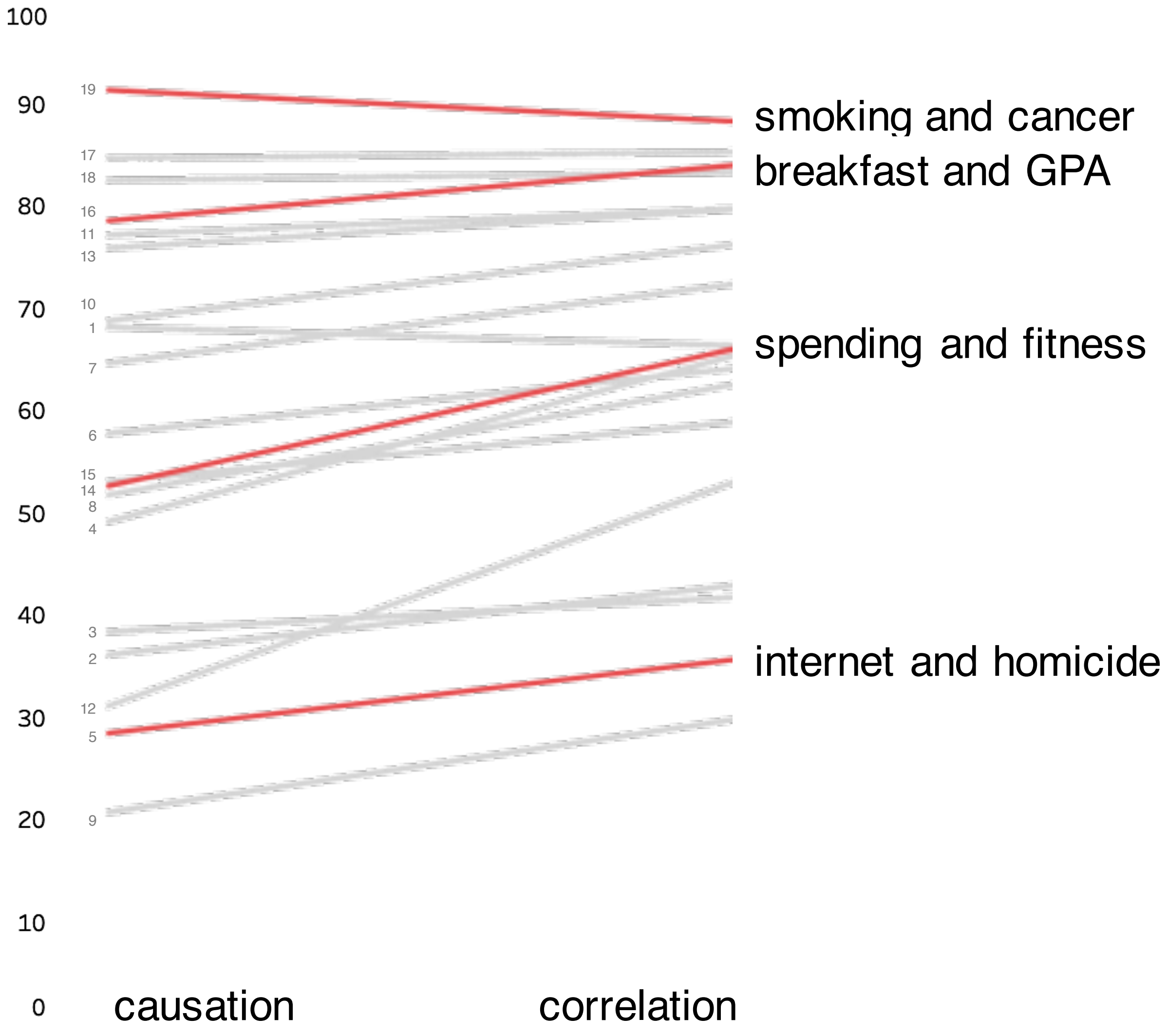


Figure 5.3. Pilot results. Grey numbers indicate the index of the 19 statements, details see supplementary. The line positions represent mean correlation and causation plausibility ratings. Red lines are the correlation and causation plausibility ratings for the selected contexts, intended to cover a range of plausibility.

grouped variables together, made direct observations or explicitly stated the data to be inconclusive. Each response is coded independently on these six dimensions, which means the same response could fit into multiple categories.

Distinguishing Correlation from Causation: Referencing past work outlining a taxonomy of causal arguments Oestermeier and Hesse (2000), we looked for causal inference patterns in the verbal responses in the generative task, to distinguish a causal interpretation from a correlational one. Specifically, words such as "causes", "leads to" and "results in" depending on the context, suggests causal interpretations, while phrases such as "as X increases, Y tend to increase" were classified as correlational interpretations.

Mentioning Third Variables: If participants discussed variables not depicted in the visualization as influencing the relations between the two depicted variables, we additionally labelled the response as "considered third variables."

Grouping Variables: Participants could also group the levels of a variable together when justifying their reasoning. For example, one could say "when X is high, Y is high, but when X is low, Y is low," which arbitrarily divides the x–variable into two dimensions. Grouping of variables may be associated with misattributed causal relations. Thus we examine variable-grouping as part of our taxonomy.

Direct Observations: We also anticipated that not all participants would provide high-level reasoning. Some could make direct observations, stating the values depicted in a visualization verbatim. "When X is 2, Y is 3" and "there is a vertical line starting at 15000" are both instances of direct observations.

Inconclusive Responses: Participants could also deem the amount of data present inconclusive without drawing any correlational or causal conclusions.

Table 5.1. Correlation and causation plausibility ratings for the four selected statement sets from the pilot experiment.

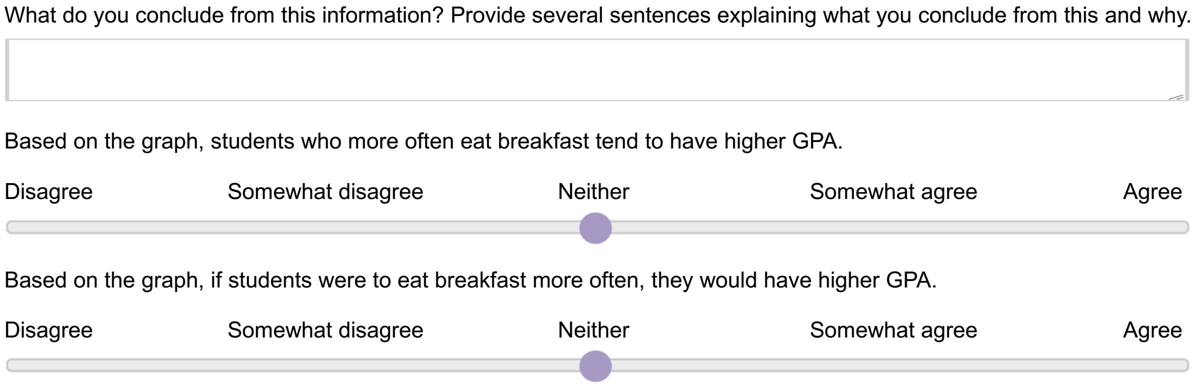
Variables	Statement	Type	Plausibility Rating
spending and fitness	People who spend more on admission to sporting events tend to be more physically fit.	correlation	65.91
	If people were to spend more on admission to sporting events, they would be more fit.	causation	52.52
smoking and cancer	People who smoke more have a higher risk of getting lung cancer.	correlation	88.14
	If people smoke more, they would have higher risk of getting lung cancer.	causation	91.19
breakfast and GPA	Students who more often eat breakfast tend to have higher GPA.	correlation	83.86
	If students were to eat breakfast more often, they would have higher GPA.	causation	78.43
internet and homicide	When there are more people using Internet Explorer, the homicide rates in the United States tend to be higher.	correlation	35.57
	If more people used Internet Explorer, there would be more homicide in the United States.	causation	28.38

5.2. Experiment 1 Causality in Context

Experiment 1 investigates whether visualization design influences how people interpret correlation and causation in data, using the four variable pairs selected from the pilot experiment. We asked participants to complete both judgment and generative tasks, in which they rate how much they agree with a correlation or causation statement, and verbally interpret the information and justify their judgment task reasoning, as shown in Figure 5.4.

5.2.1. Method

Participants were recruited through the Human Intelligence Task (HIT) postings on MTurk. We excluded workers who are not based in the United States, have an approval rate below 95%, failed the attention checks, entered nonsensical answers for the free response questions or failed the graph reading comprehension checks (details of these checks are included in the supplementary materials). An omnibus power analysis based on pilot effect sizes suggested



What do you conclude from this information? Provide several sentences explaining what you conclude from this and why.

Based on the graph, students who more often eat breakfast tend to have higher GPA.

Disagree | Somewhat disagree | Neither | Somewhat agree | Agree

Based on the graph, if students were to eat breakfast more often, they would have higher GPA.

Disagree | Somewhat disagree | Neither | Somewhat agree | Agree

Figure 5.4. Example of generative task (top) and judgment task (middle and bottom) in Experiment 1. The three questions were shown on *separate* pages in Qualtrics in the order from top to bottom.

a target sample of 136 participants would give us 95% power to detect an overall difference between visualization designs at alpha level of 0.05. We iteratively surveyed and excluded participants until we reached this sample size.

This experiment had a 4×4 Graeco Latin Square design. As shown in Figure 5.5, each participant saw four sets of data in the four variable pairing chosen from the pilot experiment, presented using four visualization designs. We will refer to the variable pairing as '*context.*' We replicated each condition 34 times with different participants to increase the reliability in our measures. We chose three simple visualization designs commonly seen in media and education News (2018); Guo (2016); Tufte (2001); Knaflic (2015) – bar graphs, line graphs and scatter plots as well as a plain text, as shown in Figure **??**. The plain text was written to parallel the bar graph, including identical information in which one variable (X) was arbitrarily divided into two groups and the corresponding average value for the other variable (Y) at those two groups were specified.

Our independent variables are the visualization design and context plausibility. Visualization design is a categorical variable indicating the design we presented the information to the

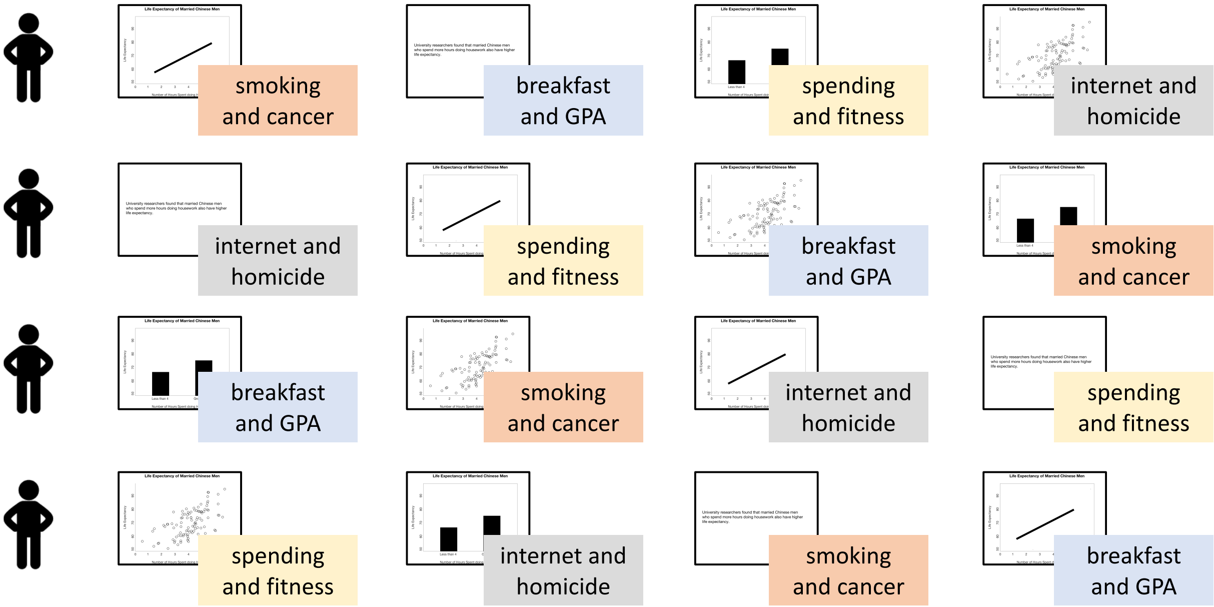


Figure 5.5. Graeco-Latin Square design showing the four conditions for Experiment 1. Each row represents a condition. Each column represents the order in which the participants saw the stimuli, with the left-most seen first and the right-most seen last.

participants, which could be bar graphs, line graphs, scatter plots or plain text. Context plausibility is the correlation and causation statement plausibility collected from the pilot experiment, which is a continuous variable from 0, extremely implausible, to 100, extremely plausible. We recorded the order in which the participants viewed the visualizations. We also collected demographic information such as participant age, gender, political orientation and level of education.

There were two dependent variables. Four researchers blind to both the study design and the condition manipulations coded the response in the *generative* task based on the interpretive taxonomy, and the participant count in each category (e.g., direct observation) was one dependent variable. The other dependent variable was participants' ratings on how much they agreed with the correlation and causation statements listed in Table 5.1 in the *judgment* task.

We used MATLAB to randomly generate 100 pairs of data points from a normal distribution with a correlation of 0.6 to avoid ceiling and floor effect of rating the underlying correlation as too high or too low. We visualized this dataset into a bar graph, line graph and scatter plot, as shown in Figure **??**. To ensure all participants viewed the same visualized data across all conditions, we relabeled the axis to fit the context without changing the underlying dataset. For example, Figure 5.6 shows the bar graph depicted in the four contexts.

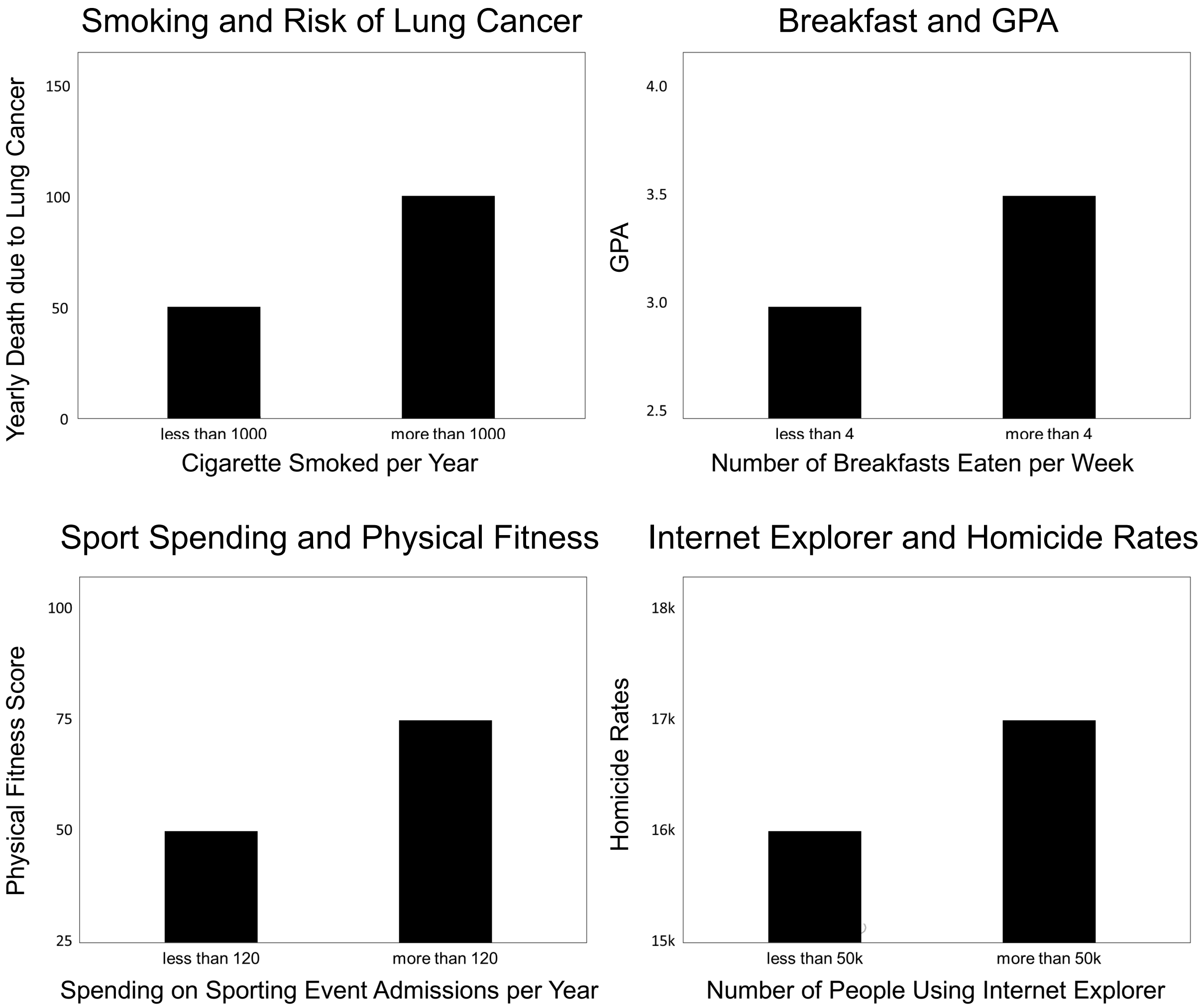


Figure 5.6. The bar graph stimulus in the four contexts.

Upon accepting the HIT, participants clicked on a Qualtrics link to access the experiment. Participants completed the four task trials and finished with demographic questions. On each trial, participants viewed a visualization (bar, line, scatter or text) and answered two graph reading comprehension check questions. They then completed the generative task in which they wrote several sentences explaining what they concluded from the visualization and why. This was followed by the judgment task in which participants read a correlation and a causation statement (presented separately on two pages), and rated how much they agree with each on a scale from 0 (disagree) to 100 (agree), as shown in Figure 5.4.

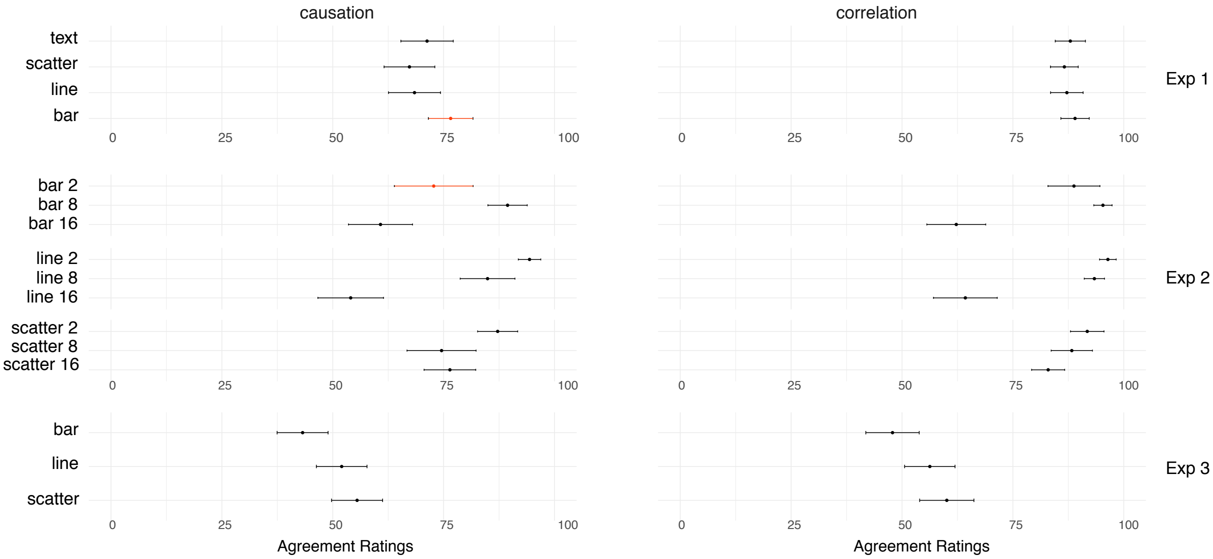


Figure 5.7. Quantitative results from all three experiments showing participants' correlation and causation agreement ratings.

5.2.2. Causation Judgment Results

We used a mixed-effect linear model to fit the causation ratings (Bates, 2005), which was how much each participant agreed with the causation statements, under the four visualization designs (bar, line, scatter and text). For fixed effects, we used visualization design, causation statement plausibility, trial order and demographic information (age, gender, education and political orientation) as predictors. Because it seemed plausible that certain combinations of contexts (pairs) and visualization designs could interact to increase or lessen perceived causality (i.e., based on conventions for showing data in certain domains), we also considered an interaction between visualization design and causation statement plausibility. We used a random intercept term accounting for individual differences as random effects.

The regression model indicated a relatively large effect of causation statement plausibility (context), χ^2=162.70, $\eta^2_{partial}$=0.274,p<0.001, a relatively small effect of visualization design (χ^2=11.65,$\eta^2_{partial}$=0.026,p<0.01), and negligible interaction effect between causation statement plausibility (context) and visualization design (χ^2=0.97,$\eta^2_{partial}$=0.002,p=0.81). Referencing Figure 5.7, participants rated bar graphs to be the most causal (M=76.59, $CI_{95\%}$=[71.51, 81.76]) and text the second most causal (M=71.26, $CI_{95\%}$=[65.30, 77.23]). This largely agreed with the results from the generative tasks where participants also made causal interpretations and the most group-wise comparisons in bar graphs and text. Given the similarity between bar graphs and text, which was written to contain identical information as the bar graph (grouping the data into two groups), we suspected that perceived causality differed between visualization designs because information was organized and presented differently among them.

Line graphs and scatter plots, unlike bar graphs and text, did not group variables together. Participants rated line graphs (M=68.43, $CI_{95\%}$=[62.52, 74.35]) and scatter plots (M=67.29, $CI_{95\%}$=[61.52, 73.07]) the least causal, which were the two designs with the most correlation interpretation in the generative task. This suggests that the effect of visualization design on perceived causality could be driven by data aggregation and visual encoding marks.

There is negligible effect of the order the visualizations were presented (χ^2 = 0.11, $\eta^2_{partial}$ = 0.002, p = 0.74), which means perceived causation does not depend on what was presented to them previously nor was there a learning effect. Results also indicated a comparatively small effect of gender (χ^2 = 4.23,$\eta^2_{partial}$ = 0.007,p = 0.040), such that male participants gave higher causation ratings, and education (χ^2 = 0.4.53, $\eta^2_{partial}$ = 0.011,p = 0.033), such that participants with higher levels of educating gave lower causation ratings.

5.2.3. Correlation Judgment Results

We used a similar mixed-effect linear model to predict how much each participant agreed with the correlation statements. We kept all predictors the same with the exception of swapping the causation statement plausibility with the correlation statement plausibility. Only correlation statement plausibility had a sizable effect predicting perceived correlations (χ^2 = 71.02, $\eta^2_{partial}$ = 0.141, p < 0.001), there was negligible effect of visualization design (χ^2 = 1.98, $\eta^2_{partial}$ = 0.005, p = 0.58), a small interaction between the two (χ^2 = 6.15, $\eta^2_{partial}$ = 0.012, p = 0.10), a tiny effect of education (χ^2 = 2.99, $\eta^2_{partial}$ = 0.007, p = 0.08), such that participants with higher levels of education gave lower correlation ratings. There were negligible effects of order, age and gender (details included in the supplementary materials). We can see this from the similar correlation confidence intervals in Figure 5.7. This suggests visualization design does not significantly influence people's judgment of correlation from data, at least when participants were given a concrete context.

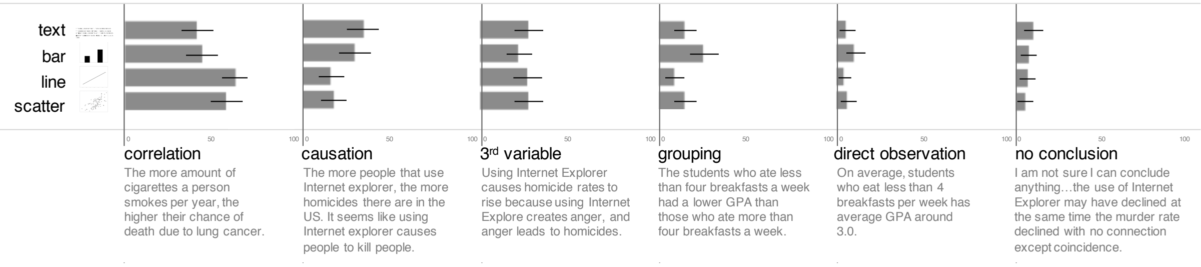


Figure 5.8. Qualitative coding results of Experiment 1. Each bar represents the percentage of participants that mentioned the indicated dimension (e.g., third variable) for a certain visualization design.

5.2.4. Qualitative Results from Generative Task

Each of generative task responses was coded as "yes" or "no" on each of the six categories, as shown in the top row of Figure 5.8.

Correlation Conclusions: Many participants appropriately inferred correlation between depicted variables, using words and phrases such as "tend to" and "the more X the more Y." A chi-square test of independence with Bonferroni adjustment suggests that varying proportion of participants drew correlation conclusions from different visualization designs (χ^2 = 27.84, $p <$ 0.001). On average, in 75.7% of the trials participants drew correlation conclusion from line graphs ($CI_{95\%}$=[68.7, 82.9]), 69.1% from scatterplots ($CI_{95\%}$ = [61.4, 76.9]), 52.9% from bar graphs ($CI_{95\%}$ = [44.6, 61.3]), and 50.0% from text ($CI_{95\%}$ = [41.6, 58.4]). Figure 5.8 shows one example of a correlation interpretation.

Causal Conclusions Among the participants who generated causal conclusions from the data, some used causation suggestive words such as “leads to” or “causes”, while others seemed to have assumed causation without using causation suggestive words. Some of these participants dismissed the visualized information as illogical because the causal relation they interpreted went against their belief or intuition. As a result, some did not reach a conclusion from the visualization, not because they were aware that correlation is not causation, but because they thought the visualization was depicting a causal relation that did not make sense to them.

For example, in response to the "spending and fitness" visualization, one participant suggested that the visualization did not make sense because "there is no correlation between the two," mistaking correlation for causation. In this case, the participant seemed to understand the notion that correlation is not causation, but assumed that the visual results implied more than just correlation nonetheless. We coded the response as both "causation" and "no conclusion."

There were also two participants who mentioned "experiments" in their responses with bar graphs, even though we specifically noted that the visualizations are generated from survey data. It is possible that some people associate bar graphs with controlled experiments, from which causal conclusions can be validly drawn.

We found several common characteristics among participants who did not assume causal relations. They questioned the directionality and predispositions, or mentioned third variables at play. For example, in the "breakfast and GPA" context, participants who did not assume causation questioned whether it is people who ate breakfast more were more likely to get good grades, or that people who were more likely to get good grades were more organized, and thus more likely to get up early and eat breakfast.

A chi-square test of independence revealed an overall effect of visualization design on whether people drew causal conclusions as defined by their generated responses ($\chi^2 = 21.77$, $p < 0.0001$). As shown in the causation column in Figure 5.8, in 39.0% of the trials participants drew causal conclusion from text ($CI_{95\%} = [30.8, 47.2]$), in 33.8% from bar graphs ($CI_{95\%} = [25.9, 41.8]$), in 20.6% from scatter plots ($CI_{95\%} = [13.8, 27.4]$), and 18.4% from line graphs ($CI_{95\%} = [11.9, 24.9]$).

Third Variables Visualization designs might influence whether people think of third variables when drawing conclusions from visualizations. We observed participants justifying both correlation and causation by connecting a third variable to the two visualized. For example, in the "internet and homicide" context, one participant speculated that "*using Internet Explorer causes homicide rates to rise because using Internet Explore[r] creates anger, and anger leads to homicides.*" Anger is not visualized on the graph, therefore it is a third variable.

A chi-square test of independence suggested that there was *no* relation between visualization design and mentioning of third variables (χ^2=2.03, p=0.57), suggesting no particular visualization design makes people more or less likely to think of third variables, as shown in the 3^{rd} variable column in Figure 5.8. On average, in 30.9% of the trials participants mentioned third variables in scatter plots ($CI_{95\%}$=[23.1, 38.7]), 30.9% in text ($CI_{95\%}$=[23.1, 38.7]), 30.2% in line graphs ($CI_{95\%}$=[22.4, 37.9]), and 24.3% in bar graphs ($CI_{95\%}$=[17.1, 31.5]).

Grouping in Response We observed an overall effect of visualization design on the number of group-wise comparisons made (χ^2=15.57, p<0.001). Researchers coded responses as group-wise comparisons when the participant described the visualized data in two groups by one dimension and compared the two grouped values in the other dimension. For example,

"The students who ate less than four breakfasts a week had a lower GPA than those who ate more than four breakfasts a week."

In 27.9% of the trials participants made group-wise comparisons of variables in bar graphs ($CI_{95\%}$=[20.4, 35.5]), 16.2% in text ($CI_{95\%}$=[9.99, 22.4]), 16.2% in scatter plots ($CI_{95\%}$=[9.99, 22.4]), and 9.6% in line graphs ($CI_{95\%}$=[4.6, 14.5]).

Direct Observations While no visualization elicited more direct observations than others χ^2=5.09, p=0.17), we observed several direct, number-specific comparisons instead of global pattern or trend observations across all designs. For example, when viewing a bar visualization on "breakfast and GPA," one participant concluded –

"On average, students who eat less than 4 breakfasts per week has average GPA around 3.0."

As shown in Figure 5.8, in 11.0% of the trials participants made direct observations in bar graphs ($CI_{95\%}$=[5.8, 16.3]), 6.6% in scatter plots ($CI_{95\%}$=[2.4, 10.8]), 5.9% in text ($CI_{95\%}$=[1.9, 9.8]), and 4.4% line graphs ($CI_{95\%}$=[0.96, 7.9]).

No Conclusions All visualizations elicited the same proportion of non conclusions (χ^2=2.57, p=0.46). In 11.0% of the trials participants drew no conclusion in text ($CI_{95\%}$=[5.8, 16.3]), 8.1% in bar graphs ($CI_{95\%}$=[3.5, 12.7]), 7.4% in line graphs ($CI_{95\%}$=[3.0, 11.7]), and 5.9% in scatter plots ($CI_{95\%}$=[1.9, 9.8]).

We observed two types of no conclusion responses, one in which participants inferred causality from the visualization but decided to draw no conclusion because it went against their intuition, and the other in which participants made a conscious decision not to. This could be a result of them choosing to be skeptical about the completeness of the information or being aware of "correlation is not causation." For example, in response to the "internet and homicide" context, one participant wrote

"I am not sure I can conclude anything —the use of Internet Explorer may have declined at the same time the murder rate declined with no connection except coincidence."

In general, many people drew from their personal experience or knowledge to make sense of the visualized information. Congruent with prior research, most participants' first intuition is to justify a potential relation between the variables visualized, despite the plausibility of the causal link (Ibrahim et al., 2016; Kahneman, 2011). Few stopped and thought of "counter examples," questioned the validity of the data, or showed clear signs of understanding that correlation is not causation.

Some participants used "template" words or phrases, such as "correlation is not causation" or "Y tend to increase with varying levels of X" to frame their conclusions. For example, one participant made the following conclusion in the "internet and homicide" scenario.

"The graph shows that in cities with more people using Internet Explorer, there tend to be many more homicides. While the results are pretty clear, I think "correlation is not causation" should be applied here. I'm not a scientist, but I don't think the two variables are really related in any meaningful way."

It is also apparent when a participant only memorized the phrase "correlation is not causation" without truly understanding the concept. They read correlation from the data, and assumed the data to be telling a causal story as they confuse correlation for causation. But, because they were superficially aware that "correlation is not causation," they dismissed the *correlation* in data despite the observable correlation in data. For example, this participant was clearly aware of the phrase "correlation is not causation," but instead of critically thinking through third variables or other possibilities, quickly dismissed the data and the apparent correlation.

"With only this information I can't conclude anything since I do not see any correlation. In my opinion these two variables are uncorrelated..."

Furthermore, all participants interpreted the visualization assuming the X -> Y directionality, such as "as X increases Y increases." For people who made causal conclusions, all of them described the x-axis variable as the cause and the y-axis variable as the effect. This suggests that there may exist a conventional interpretation of causality in data for the x-axis variable to be seen as the cause and the y-axis variable to be seen as the cause.

5.2.5. Discussion

In general, the quantitative and qualitative results told similar stories of how, when given specific pairs of common variables, people perceived causality as more likely in bar graphs and less likely in scatter graphs. Context also had a relatively large effect on perceived causality, but the effect of visualization design on perceived causality was not context dependent. We took away the specific pairs of common variables in subsequent experiments to further examine *how* visualization designs influence perceived causality.

5.3. Experiment 2 Aggregation levels

Experiment 1 found that people perceived high causality from bar graphs and low causality from scatter plots. But is this driven by properties of the visual encoding marks (e.g., rectangular bars versus circular points versus lines), or by how aggregated data is? For example, the bar graph we showed aggregated the data into 2 groups while the scatter plot did not aggregate any data, showing each data point individually. Experiment 2 tested the effect of the amount of aggregation in data on perceived causality, and whether the visual encoding marks interact with this effect by comparing bar graphs, line graphs and scatter plots.

5.3.1. Method

Because visualization context (i.e., what specific pair of variables was shown) did not influence the effect of visualization design on perceived causality, we omitted context from the visualizations in Experiment 2. Instead of presenting the data in four scenarios with varying plausibility, we stripped the variable names (e.g., "GPA") and replaced with abstract variable labels (e.g., "X","Y"). We operationalized the amount of aggregation as the number of bins the data is sorted

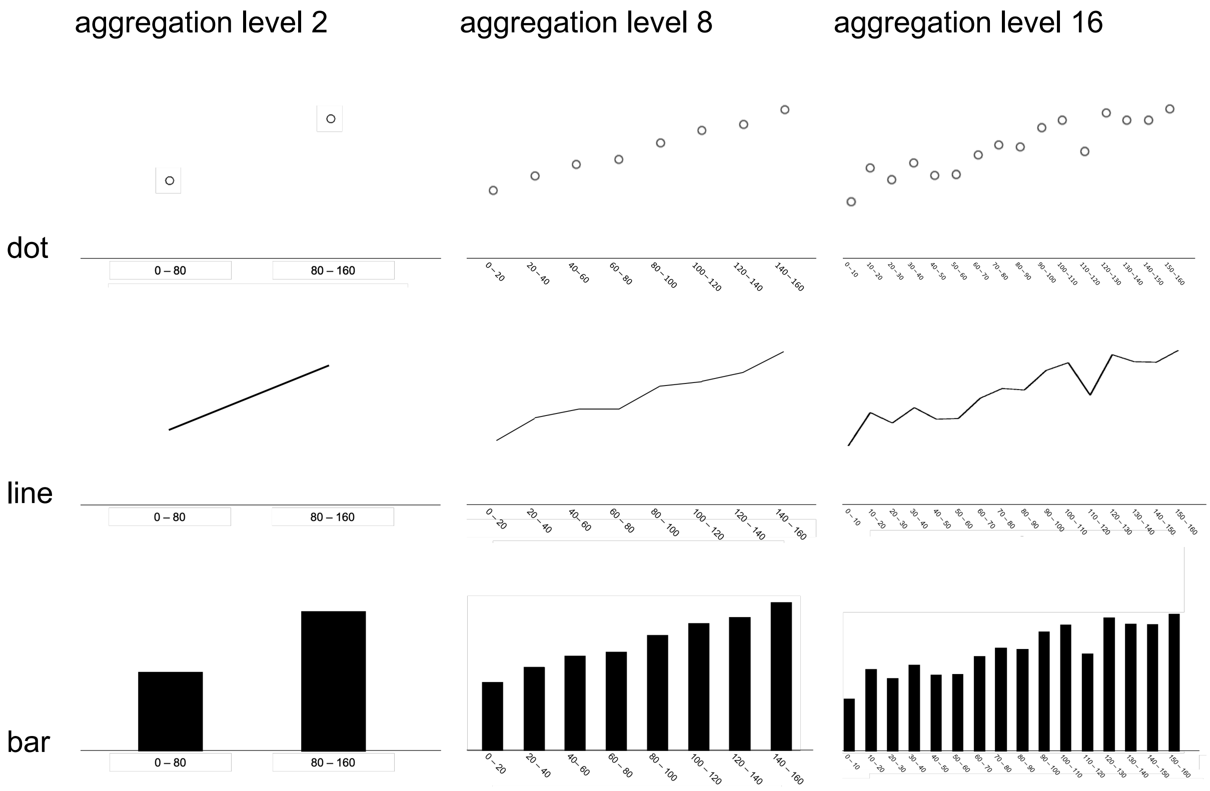
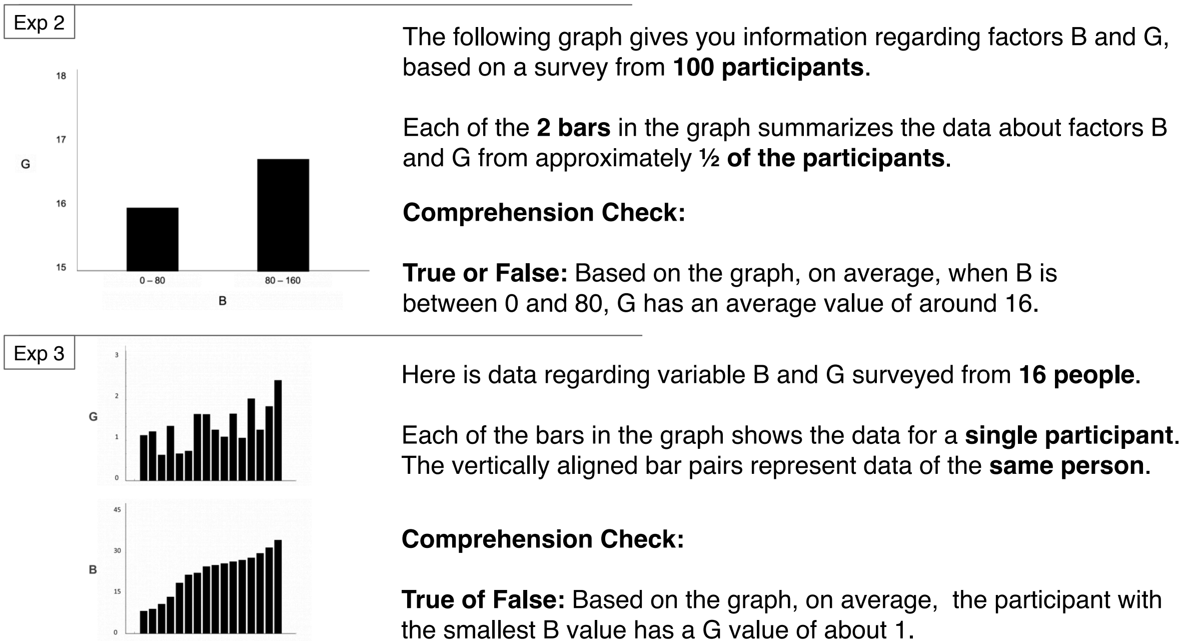


Figure 5.9. Three aggregation levels tested in Experiment 2 for bar, line and dot type encoding marks.

in. The bar graph used in Experiment 1 aggregated the data into two bins. For Experiment 2, we additionally created bar graphs that aggregated the data into eight bins and 16 bins. We created dot plots and line graphs using the same binned data in the bar graphs, but replacing the rectangular bars with circles and lines, as shown in Figure 5.9. Here, bar graphs depict comparisons of data between two, eight or 16 groups, which fit regular conventions of graphic communication using bar graphs(Zacks and Tversky, 1999). Line charts are also sometimes aggregated, such as when showing daily, weekly, or monthly estimates. However, conventional scatter plots typically illustrate each dot as an individual data value (Sarikaya and Gleicher, 2018), making our scatterplot stimuli less realistic but useful for the sake of a controlled comparison.

We explicitly told the participants that the visualized data were generated by summarizing and binning data as they viewed the visualizations, as shown in the left figure in Figure 5.10. To ensure the participants understood the plotted data, we created instructions with examples for participants to read through (see supplementary for the example). We asked each participant six graph comprehension questions on the specific visualizations we examined for the experiment, to confirm that participants understood the visualizations, as shown in Figure 5.10. Similar to Experiment 1, participants who failed the comprehension checks were excluded from analysis as they did not appear to have understood the data (the full experiment and data are available as supplementary materials). Participants completed the judgment task by rating how much they agreed with correlation and causation statements, similar to Experiment 1, but we excluded the generative task as the variables were abstract.



The following graph gives you information regarding factors B and G, based on a survey from **100 participants**.

Each of the **2 bars** in the graph summarizes the data about factors B and G from approximately ½ **of the participants**.

Comprehension Check:

True or False: Based on the graph, on average, when B is between 0 and 80, G has an average value of around 16.

Here is data regarding variable B and G surveyed from **16 people**.

Each of the bars in the graph shows the data for a **single participant**. The vertically aligned bar pairs represent data of the **same person**.

Comprehension Check:

True of False: Based on the graph, on average, the participant with the smallest B value has a G value of about 1.

Figure 5.10. Snapshots from Experiment 2 (left) and Experiment 3 (right).

The independent variables in this experiment are visual encoding marks, which can be rectangular bars, lines or dots, and aggregation level, which can be two, eight or 16. The dependent variables are correlation ratings and causation ratings, similar to Experiment 1. We used a 3×3 Graeco Latin Square design crossing visualization design and aggregation groups, similar to that in Experiment 1, which crossed visualization design and context. Each participant saw three visualizations — bar graph, line graph and dot plot, one of which aggregated into two groups, one into eight groups and other into 16 groups. We recruited 129 participants for Experiment 2 using the same method and exclusion criteria.

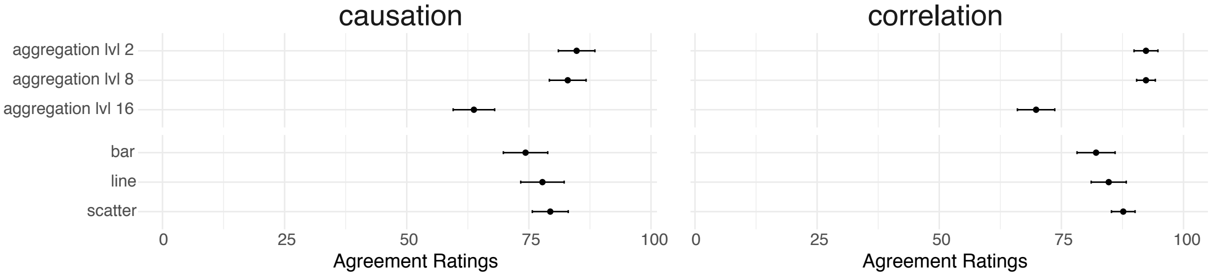


Figure 5.11. Main effect of aggregation levels (top) and visual encoding types (bottom) on correlation and causation ratings in Experiment 2.

5.3.2. Causation Judgment Results

We used a similar mixed-effect linear model from Experiment 1 to fit the causation ratings with fixed effects of visual encoding marks, aggregation level, an interaction between encoding marks and aggregation level, trial order and demographic information (age, gender, education and political orientation), and a random intercept term accounting for individual differences as random effects.

The regression model indicated a relatively small main effect of visual encoding marks (χ^2 = 5.97, $\eta^2_{partial}$ = 0.020,p = 0.050), such that aggregated dot plots had the highest causality ratings (*M* = 79.38, $CI_{95\%}$ = [75.67, 83.09]), followed by line encodings (*M* = 77.78, $CI_{95\%}$ = [73.29, 82.26]), and rectangular bar encodings had the lowest causality ratings (*M* = 74.32, $CI_{95\%}$ = [69.73, 78.90]), as shown in Figure 5.11 (top).

There is relatively large main effect of aggregation level, such that visualizations with the more data aggregation were perceived as more causal ($\chi^2 = 117.05, \eta^2_{partial} = 0.29, p < 0.001$). Visualizations with aggregation level two, the most aggregation which binned data into two groups, had the highest average causality ratings (*M* = 84.76, $CI_{95\%}$ = [81.00, 88.55]), followed by visualizations with aggregation level eight (*M* = 82.95, $CI_{95\%}$ = [79.16, 86.75], and visualization with the least aggregation, which binned data into sixteen groups, had the lowest average causality ratings (*M* = 63.74, $CI_{95\%}$ = [59.46, 68.03]), as shown in Figure 5.11 (bottom).

There is an interaction effect between visual encoding marks and aggregation level ($\chi^2 = 28.10, \eta^2_{partial} = 0.089, p < 0.01$) on perceived causality, as shown in Figure 5.7. For dot encodings, perceived causality did not differ significantly between aggregation level two (*M* = 87.19, $CI_{95\%}$ = [82.54, 91.84]), aggregation level eight (*M* = 74.53, $CI_{95\%}$ = [66.51,82.56]) and aggregation level 16 (*M* = 76.42, $CI_{95\%}$ = [70.42,82.41]). For line encodings, perceived causality significantly decreased as the number bins increased, such that aggregation level two (*M* = 94.37, $CI_{95\%}$ = [91.76,96.98]) was perceived the most causal, followed by aggregation level eight (*M* = 84.91, $CI_{95\%}$ = [78.55,91.26]), and aggregation level 16 was perceived the least causal (*M* = 54.05, $CI_{95\%}$ = [46.43,61.67]). For bar encodings, aggregation level eight was perceived as being the most causal (*M* = 89.42, $CI_{95\%}$ = [84.85,93.98]), followed by aggregation level two (*M* = 72.77, $CI_{95\%}$ = [63.62, 81.92]), and aggregation level 16 the least causal (*M* = 60.77, $CI_{95\%}$ = [53.33, 68.20]). There is a negligible effect of the order the visualizations were presented (χ^2 = 0.14,$\eta^2_{partial}$ = 0.002, p = 0.71) as well as participant age, political orientation, gender and education.

5.3.3. Comparing Experiment 1 and Experiment 2 Bars

Experiment 1 seemed to indicate that bar graphs conveyed a greater impression of causation than other representations, Experiment 2 suggests that this impression is due to an interaction between the visual encoding marks and aggregation level. Comparing the causation ratings of bar graphs in Experiment 2 with that in Experiment 1, as shown marked in red in Figure 5.7, we see that although participants gave lower causation ratings for bar encodings overall, if we only compare the aggregation level two bar condition from Experiment 2 with the bar condition in Experiment 1 (which is an aggregation level two bar graph with context), the two results match ($p = 0.47$), suggesting that bar graphs with two bars may be an interesting case study, see section **??**. Examining participant quotes for the Experiment 1 in Section **??** (Causal Conclusions), one explanation may be that many participants associate aggregation level 2 bar graphs with controlled experiments, which can be a valid way to establish causal relationships.

5.3.4. Correlation Judgment Results

We used the same mixed-effect linear model to fit the correlation ratings. The model indicated a relatively small main effect of visual encoding marks (χ^2 = 9.93,$\eta^2_{partial}$ = 0.03,p < 0.01), such that aggregated dot plots had the highest correlation ratings (M = 87.67, $CI_{95\%}$ = [85.23, 90.11]), followed by line encodings (M = 84.69, $CI_{95\%}$ = [81.06, 88.32]), and rectangular bar encodings had the lowest ratings (M = 82.10, $CI_{95\%}$ = [78.17, 86.03]), as shown in 5.11.

There is a relatively large main effect of aggregation level, such that visualizations with more data aggregation were perceived as more correlational (χ^2 = 212.31, $\eta^2_{partial}$ = 0.40,p < 0.001). Visualizations with aggregation level two, the most aggregation which binned data into two groups, had the highest average correlation ratings (M = 92.32, $CI_{95\%}$ = [89.85, 94.79]), followed by visualizations with aggregation level eight (M = 92.31, $CI_{95\%}$ = [90.39, 94.25], and visualization with the least aggregation, which binned data into 16 groups, had the lowest average ratings (M = 69.82, $CI_{95\%}$ = [65.96, 73.68]), as shown in 5.11.

There is a medium interaction effect between visual encoding marks and aggregation level (χ^2 = 30.32,$\eta^2_{partial}$ = 0.088,p < 0.001) on perceived correlation, as shown in Figure 5.7. For dot encodings, perceived correlation did not differ significantly between aggregation level two (M = 91.77, $CI_{95\%}$ = [87.88, 95.66]), aggregation level eight (M = 88.28, $CI_{95\%}$ = [83.49,93.06]) and aggregation level 16 (M = 82.95, $CI_{95\%}$ = [79.12,86.79]). For line encodings, perceived correlations significantly decreased as the number bins increased, such that aggregation level two (M = 96.42, $CI_{95\%}$ = [94.49,98.35]) was perceived to be the most correlational, followed by aggregation level eight (M = 93.37, $CI_{95\%}$ = [91.03,95.72]), and aggregation level 16 was perceived to be the least correlational (M = 64.28, $CI_{95\%}$ = [56.88,71.68]). For bar encodings, aggregation level eight was perceived to be the most correlational (M = 95.30, $CI_{95\%}$ = [93.18,97.43]), followed by aggregation level two (M = 88.77, $CI_{95\%}$ = [82.74, 94.80]), and aggregation level 16 the least correlational (M = 62.23, $CI_{95\%}$ = [55.39, 69.07]).

There is a relatively small effect of the order the visualizations were presented (χ^2 = 10.65, $\eta^2_{partial}$ = 0.022, p = 0.001), indicating a learning effect, which is reasonable given the novelty of the visualization designs. There was negligible effect of age and gender, but a relatively small effect of political orientation (χ^2 = 1.85,$\eta^2_{partial}$ = 0.013, p = 0.17), such that more liberal participants gave higher correlation ratings overall, and education (χ^2 = 3.5, $\eta^2_{partial}$ = 0.019, p = 0.84), such that participants with higher levels of education gave higher correlation ratings.

5.3.5. Discussion

Bar visual encoding marks received the lowest causal ratings, followed by line, and dot encodings received the highest causal ratings. These ratings could be further increased or decreased by the amount of data aggregation, such that decreased aggregation (increasing the number of bins) decreased perceived causality, and increased aggregation increased perceived causality in data. However, the visualizations in this experiment all aggregated data, even at the smallest aggregation level (with 16 bins). In order to isolate the effect of visualization encoding, we test how visual encoding marks influence perceived causality when *no* data is aggregated in Experiment 3.

5.4. Experiment 3

The bar graphs and line graphs examined in our first two experiments aggregated data. Experiment 1 showed aggregated bars binned into two groups and a continuous line, which essentially aggregated across all levels. Experiment 2 used aggregated plots which are not commonly seen, because scatter plots and to some extent line charts don't typically depict binned data, as least as often as bar charts do. Scatter plots, for example, usually show non-aggregated raw data. One familiar instance where data is naturally dis-aggregated is a nominal list, which usually shows ranking data, such as (Gratzl et al., 2013).

5.4.1. Method

We created modified bar graphs, line graphs and scatter plots to present non-aggregated data, as shown in Figure 5.12. This modification aims to parallel the non-aggregated way that scatter plots present data in bar and line charts. For each graph, the x-axis shows the index of each data point. This is a nominal dimension in which order is typically not meaningful, such as an index assigned to each unique name of a person or university. Each of the two graphs shows the value of one variable associated with the index, and the vertically aligned bar pairs represent the variable values associated with the same index. One of the variables was sorted in increasing value to mimic the x-axis and the other is left unsorted mimicking the y-axis in a scatter plot. We made the same modification to line graphs and scatter plots, as shown in Figure 5.12.

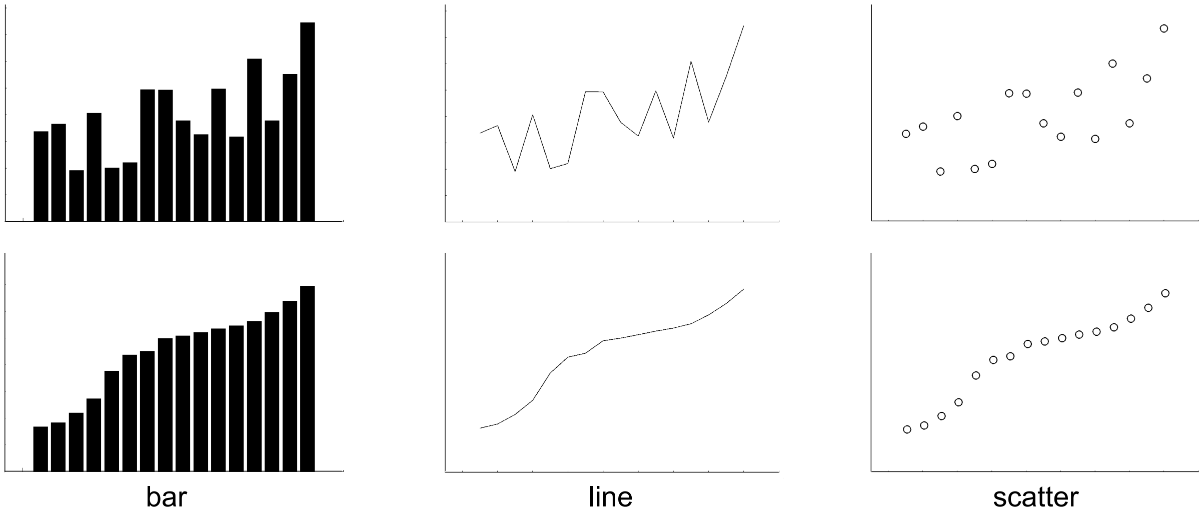


Figure 5.12. Non-aggregated data visualized with bars, lines and dots.

Similar to Experiment 2, the visualizations created for this experiment are not conventional and therefore may seem unintuitive to some viewers (although we do sometimes see them in the real world, as shown in the left column of Figure 5.2). To ensure the participants in this experiment understood the plotted data, we created instructions with examples for participants to read through (see supplementary for example details). We applied the same exclusion criteria as those in Experiment 2.

In this within-subject design, every participant viewed all three visualization designs in different order, counterbalanced with different axis values labels. An omnibus power analysis, based on pilot effect sizes, suggested a target sample of 62 would yield enough power to detect an overall difference between visualization designs. We collected data following the same data collection and exclusion method as the previous experiments.

5.4.2. Visual Mark Encoding Types

As shown in Figure 5.7, a mixed-model linear regression model predicting perceived causality using visual encoding type, trial order and demographic information as fixed effects and individual participants as random effects showed an effect of visual encoding types (χ^2 = 15.44,$\eta^2_{partial}$ = 0.10, p < 0.01), such that dot encodings were perceived to be the most causal (M = 55.49, $CI_{95\%}$ = [49.62, 61.36]), closely followed by line encodings (M = 52.02, $CI_{95\%}$ = [46.19, 57.84]) and bar encodings the least causal (M = 43.21, $CI_{95\%}$ = [37.35, 49.07]). There is a relatively small effect of order (χ^2 = 2.58,$\eta^2_{partial}$ = 0.019) suggesting that participants showed comparatively small learning effects towards the potentially unfamiliar non-aggregated visualizations, age (χ^2 = 3.43,$\eta^2_{partial}$ = 0.014), such that older participants rated causation less on average, and education (χ^2 = 4.84,$\eta^2_{partial}$ = 0.035), such that participants with higher levels of education gave higher causation ratings.

A mixed-model linear regression model predicting perceived correlation using the same fixed effects and random effects showed an effect of visual encoding types (χ^2 = 15.17,$\eta^2_{partial}$

= 0.10,p < 0.01), such that dot encodings were perceived to be the most correlational (M = 60.10, $CI_{95\%}$ = [53.86, 66.33]), closely followed by line encodings (M = 56.27, $CI_{95\%}$ = [50.48, 62.06]) and bar encodings the least correlational (M = 47.86, $CI_{95\%}$ = [41.71, 54.00]). There is a relatively small effect of order (χ^2 = 7.68,$\eta^2_{partial}$ = 0.055) suggesting a relatively small learning effect, and negligible effects of age, gender, political orientation and education.

5.4.3. Aggregated and Non-Aggregated Data

We did a post-hoc between-subject comparison using a mixed-effect linear model comparing the non-aggregated visualization causality ratings in Experiment 3 to the ratings of the visualization with aggregation level 16 in Experiment 2, since both conditions showed 16 data values (16 pairs of values in Experiment 3), differing only in data manipulation – whether the data was explicitly stated to be aggregated or not. We found a relatively large effect of data manipulation (χ^2 = 93.38,$\eta^2_{partial}$ = 0.17,p < 0.001) such that visualizations that aggregated data (Experiment 3, M = 50.24, $CI_{95\%}$ = [46.84, 53.64]) were perceived to be more causal than visualizations that did not (Experiment 2, M = 77.16, $CI_{95\%}$ = [74.70, 79.62]).

5.5. General Discussion

Overall, the choices authors make between visual encoding marks and the amount of data aggregation likely contribute to perceived causality in data. Although our results from Experiment 1 suggest that bar charts were perceived as most likely to be causal, controlling for the amount of data aggregation in Experiment 2 and Experiment 3 suggested that the level of aggregation was the driving factor of higher perceived causality in bar graphs. We also found an effect of visual encoding marks such that bars were perceived to be less causal than line and dot encodings. However, as discussed in section 5.3.3, two-bar bar graphs seemed to be a special case where participants consistently perceived the relationship it depicted to be highly causal.

Our qualitative characterization of verbal responses could be improved. We encountered several instances of ambiguous language, such as "*there is some sort of relationship between A and B*," which made it difficult for researchers to decide whether the participants meant a correlation or a causal relation. Some participants used template phrases such as "*correlation is not causation*" and "*A is correlated with B*" to describe relations in data, but we lacked ways of evaluating whether they *actually* read a causal relation from the data or not.

This work took an initial step toward showing that visualizations can be designed to mitigate misinterpretation of correlation and causation. Future experiments could investigate how other techniques, such as verbal annotation on the visualization, could reinforce better interpretation of correlation and causation in addition to visualization designs, potentially contributing to data journalism and education.

CHAPTER 6

Data Arrangement Matters

One key task viewer tend to do with visualizations is making comparisons. Comparison is a foundational perceptual operation in data visualizations (Gleicher et al., 2011). For example, among the four data values depicted by the bar charts in Figure 1, you might compare the average of the two A values to the average of the two B values, the average of West vs. East, or any two particular points. An experienced graph reader might compare the interaction – the delta among deltas – between the two factors. Within the scatterplots, you might compare average values along X or Y, for the categories indicated by color or shape.

Even for an expert, reading a chart is less like instantly recognizing a picture, and more like slowly reading a paragraph (Shah and Freedman, 2011), with the sentences being individual comparisons that unfold as a sequence over time (Nothelfer and Franconeri, 2019). Imagine a data analyst who has carefully walked through this type of sequence for a dataset and has identified the specific pattern that they would like to communicate to an audience. We argue that the analyst can vastly improve the efficiency of their communication by not only choosing the right visualization, but also knowing how to *arrange and group* the values within the visualization to guide a viewer toward seeing that key pattern early in their sequence of comparisons. We demonstrate that visual comparison is guided not only by conscious decisions about which comparisons are of interest for a given problem or dataset, but also by the way data values are arranged by visual grouping cues like the proximity, size, color, and shape of the visualization's marks.

The importance of this factor should be visible in Figure 6.1: the two bar charts reflect an identical underlying dataset, yet each layout affords different comparisons. The same difference should be clear across the two scatterplots. We explore and model how grouping cues can control the salience of a given comparison, with the goal of producing guidelines for data communication design.

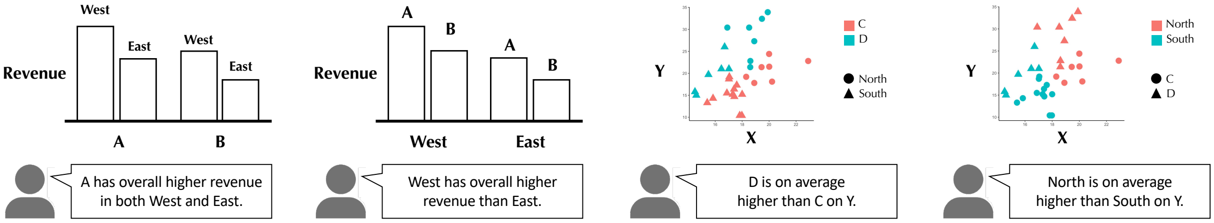


Figure 6.1. The two bar charts above depict the same data values. But different arrangements can guide the viewer to notice different patterns, as illustrated by the single most likely participant conclusion shown below each chart. The two scatterplots also depict the same data values, but viewers are most likely to compare across the category coded by color (instead of shape).

6.1. Pilot Experiment

We first tested whether crowdworkers on Amazon's Mechanical Turk (MTurk) could produce typed responses that we could interpret to confidently assess the comparison that they found most salient. We showed 58 participants bar charts through Qualtrics on MTurk (Snow and Mann, 2013). Participants viewed a bar chart depicting the revenues of two companies in two regions, as shown in the left-most chart in Figure 6.1. The underlying data comprises a 2x2 factorial design - two levels of company (A and B) and two levels of region (East and West). It illustrated a main effect of company (revenue of A is greater than that of B), a main effect of region (revenue in the West is greater than that in the East), and an interaction between company and region. In the experiment, the bars could either be spatially grouped by company (see first chart in Figure 6.1) or by region (see second chart in Figure 6.1).

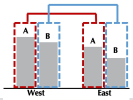
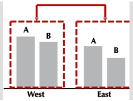
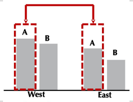
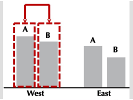
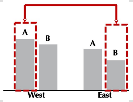
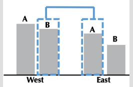
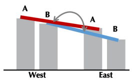
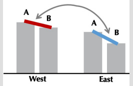
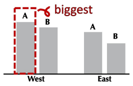
Participants were prompted to type the first conclusion they would draw from the chart. After typing the first conclusion, a new prompt would appear on the same page and ask them to enter the second conclusion. This continued until the participant entered five conclusions. Then, on a separate page, they entered demographic information such as their age and gender. The order in which they wrote the sentence conclusions were recorded as the sentence rankings. The effect sizes obtained from these results were also used to determine the sample size needed for Experiment 1.

6.1.1. Qualitative Coding: Interpretation Taxonomy

To provide a structured way of interpreting participants' sentence conclusions in our experiments, we analyzed the conclusions participants drew and coded them into the following categories, shown on Table 6.1.

Main Effect: Drawing inspiration from research on interpreting results from 2x2 factorial design experiments (Cozby, 2007), we categorized conclusions comparing how the two levels of each factor affect revenue as a main effect comparison. Comparing overall revenues of companies A and B and comparing overall revenues of regions East and West are both main effects. A comparison that is incongruent to the spatial grouping would be a *distal-group comparison*. That is, the two levels of the factor compared are spatially separated on the x-axis. A comparison that is congruent to the spatial grouping would be a *proximal-group comparison*. That is,

Table 6.1. Examples of conclusions, coded expressions, and categories

Example	Sentence Conclusions	Expression	Category	Comparison
	A has more sales than B.	$A > B$	Main	Distal
	East is worse than West.	$East < West$	Main	Proximal
	A does much better in the West as opposed to the East.	$West_A > East_A$	Pair-wise	Distal
	A's West region generates more revenue than that of B.	$West_A > West_B$	Pair-wise	Proximal
	A's West region has higher revenue than B's East region.	$West_A > East_B$	Pair-wise	Edge
	B's West region has higher revenue than A's East region.	$West_B > East_A$	Pair-wise	Middle
	A shows more difference in revenue between the two regions than B.	$\|West_A - East_A\| > \|West_B - East_B\|$	Interaction	Distal
	West's profitability fluctuates more than East.	$\|West_A - West_B\| > \|East_A - East_B\|$	Interaction	Proximal
	Company A West is the best.	$max(4) = West_A$	Superlative	

the two levels of the factor compared are nested on the x-axis. As shown in Table 6.1, comparing revenue between A and B would be a distal-group comparison and comparing revenue between East and West would be a proximal-group comparison.

Pair-Wise Comparisons: Comparing one level of a factor across the two levels of the other factor is a pair-wise comparison, such as comparing A's revenue from the West to that from the East. Similar to main effect comparisons, pair-wise comparisons can be categorized either as *distal-group*, which compares the two spatially separated levels of one factor, or as *proximal-group*, which compares the nested two levels of one factor. As shown in Table 6.1, comparing revenue between West A and East A would be a distal-group comparison and comparing revenue between West A and West B would be proximal-group. Although no participant has concluded these in the pilot, we identified a third type of pair-wise comparison: an *edge comparison* between the left most and the right most bars, as well as a fourth type of pair-wise comparison: a *middle comparison* between the middle two bars. Examples of these comparisons can also be found in Table 6.1.

Interaction: When the effect of one factor on the dependent variable depends on the particular level of the other factor, we have an interaction. We think of interaction comparisons as comparing the slopes of lines connecting two factor levels. There are two possible interaction comparisons. One is *distal-group interaction* comparison, which looks at the effect of the spatially grouped variable on the dependent variable (revenue), moderated by the nested variable (either location or company). This is essentially a comparison of the slope of the line connecting the first and third bars to the slope of the line connecting the second and fourth bars. The other is *proximal-group interaction* comparison, which looks at the effect of the nested variable on the dependent variable (revenue), mediated by the spatially separated variable (either location or company). This is a comparison of the slopes of lines connecting the spatially grouped pairs. Again, as shown in Table 6.1, a distal-group interaction would compare the difference between West A and East A to the difference between West B and East B. A proximal-group interaction would compare the difference between West A and West B to the difference between East A and East B.

Superlative: These sentences include conclusions about single values reflecting maximum and minimum revenues. Unlike comparative relations (e.g., X is bigger/taller/longer than Y), which take more effort to extract because they require an eye movement from one data value to the other (Michal and Franconeri, 2017; Wolfe, 1998), finding the maximal or minimal element is automatic and effortless (Picon and Odic, 2017), suggesting that a different cognitive/perceptual process is at work.

Other: Not all participants made comparisons or superlative conclusions. This category included conclusions about the number of companies, colors, title or axis-labels and are coded as *other*. Further, referencing methods from Shah and Freedman (2011), comments about implications, novelty or "obviousness" of data would be included in this category.

All responses were transformed into expressions by two human coders and categorized following the taxonomy above. These expressions categorize the human sentences into common comparisons in data. We discuss the potential of using this coding system for creating recommendation systems in Section **??**. Table 6.1 summaries the taxonomy, giving examples of responses, expressions and comparison categories.

6.1.2. Insights from Pilot

We noticed that several participants mentioned multiple comparisons in the same conclusion. For example, when prompted to report the first conclusion, the participant identified both a distal-group pair-wise comparison and a proximal-group pair-wise comparison. In these cases, we broke up the response in two parts and included each in its corresponding categories, preserving the rank order for both. We also noticed that many participants stopped giving interpretable answers after the third conclusion, typing answers such as "I can't draw any additional conclusions." We therefore limited the total number of comparison prompts to three.

The underlying data chosen for the pilot includes both main effects and an interaction, but we observed very few conclusions mentioning interactions. This may reflect interactions being cognitively more difficult to extract, but it also could be due to the particular data values used in the pilot. For the main experiment we therefore decided to sample other possible datasets for the 2x2 factorial visualizations. We also noticed that people tended to be guided by size-spatial grouping, such as whether a tall bar is spatially grouped with another tall bar, versus a short bar. Because size-spatial grouping depends on the values in the dataset, this factor is intertwined with the previous challenge of choosing new sets of data values.

Figure 6.2 depicts eight possible categories of datasets, depending on whether there exist main effects and/or interactions. Note that size grouping can be congruent or incongruent with spatial grouping, as shown in (b) and (c), respectively.

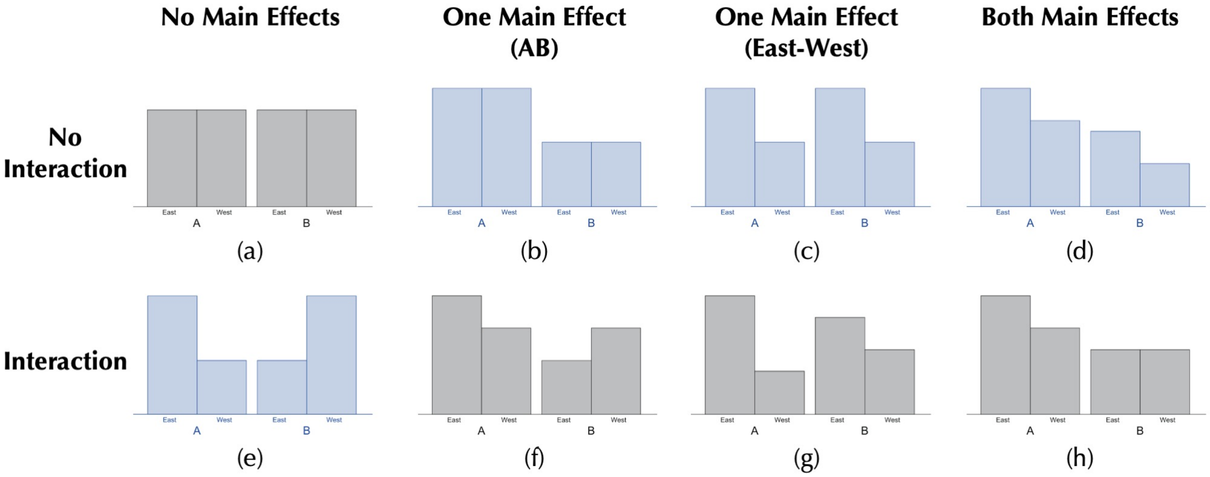


Figure 6.2. Eight possible relations of two two-level factors. Note that the outcomes are idealized examples; perfect outcomes rarely occur in actual data. Blue colored ones are examined in depth in this chapter.

6.2. Experiment 1: Bar Visualizations

To investigate the effect of spatial, color, and size groupings on comparisons, we chose a set of data values and arrangements that minimized potential confounds created by data relations and size grouping. To cover blue subset of the relationships from the eight possibilities shown in Figure 6.2, we selected three combinations of main effects and interactions (the three top-level columns of Figure 6.3). These three datasets have three different underlying relations: one dataset with no main effect but an interaction as in (e), one with one main effect but no interaction as in (b) or (c), and one with two main effects but no interaction as in (d). The datasets used to produce each chart were perturbed with noise (+/- 10% vertical length) so that no two values were identical. This created trivial interactions even in the no interaction charts; see Figure 6.3. Note that (f), (g), and (h) were not selected for examination in the experiment because they are visually similar to (d), (c), and (b), especially after perturbing each with noise.

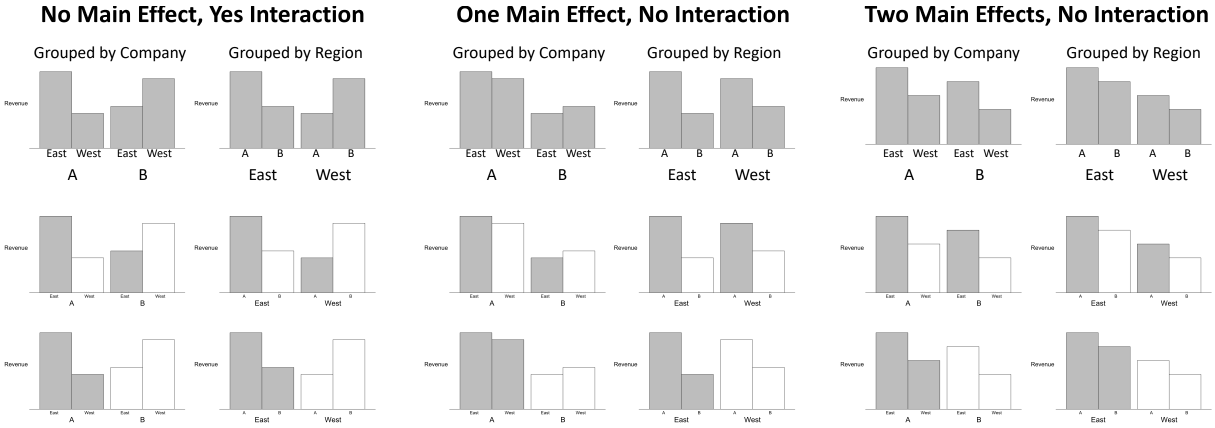
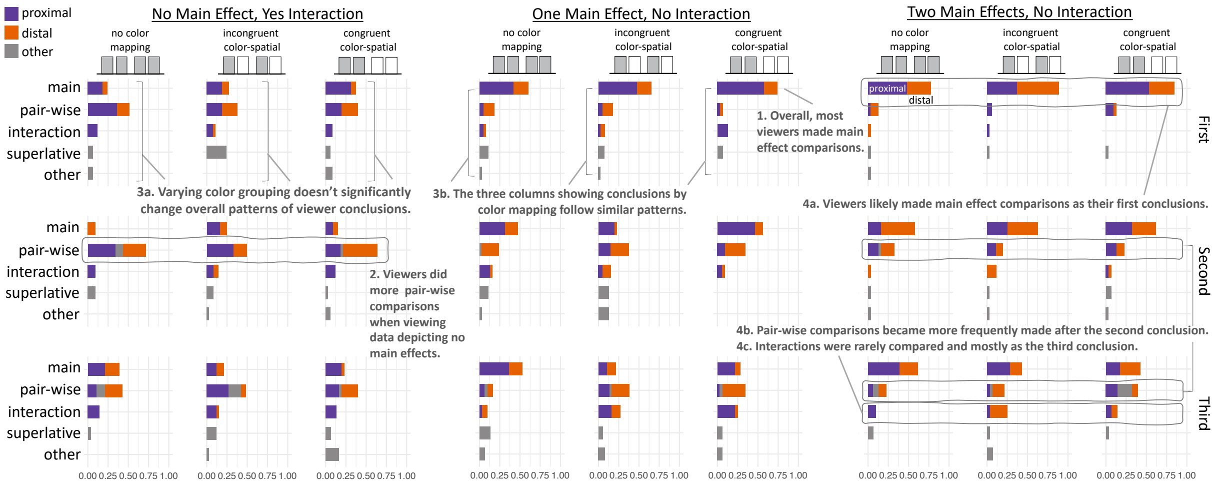


Figure 6.3. Design space for bar visualization experiment. These are the three data relations used for this experiment, chosen from the set of relations depicted in Figure 6.2. The first row shows no color mapping, the second row shows incongruent color-spatial mapping, and the third row shows congruent color-spatial mapping.

We used the same two factors (company and region) from the pilot. Each dataset was presented as a bar chart either grouped spatially by company or by region, subsequently making the size grouping either congruent or incongruent to the spatial grouping. We also incorporated color grouping for each chart. Color was either mapped onto region or company, making it either congruent or incongruent with spatial grouping. Congruent color-spatial mapping means the spatially proximate two bars share the same color (e.g., grey, grey, white, white), while incongruent color-spatial mapping means the spatially proximate two bars have different colors (e.g., grey, white, grey, white). Figure 6.3 maps this design space.

This experiment followed the same procedure as the pilot. Participants viewed these bar charts and wrote down their top three conclusions. A power analysis based on the pilot effect sizes derived from logistic regression models suggested a target sample of 45 conclusions per chart would give 80% power to detect patterns in conclusion frequencies at alpha level of 0.05 (Hsieh et al., 1998). Because we have no control over the conclusions a given participant would draw from a visualization, we iteratively surveyed and excluded participants on Amazon's Mechanical Turk who gave nonsensical answers or failed attention tests until we had at least 45 conclusions for each bar chart.



1. Overall, most viewers made main effect comparisons.

3a. Varying color grouping doesn't significantly change overall patterns of viewer conclusions.

3b. The three columns showing conclusions by color mapping follow similar patterns.

4a. Viewers likely made main effect comparisons as their first conclusions.

2. Viewers did more pair-wise comparisons when viewing data depicting no main effects.

4b. Pair-wise comparisons became more frequently made after the second conclusion.
4c. Interactions were rarely compared and mostly as the third conclusion.

Figure 6.4. Conclusions by underlying data relations. Each row section represents the order in which the sentence conclusion was made. Orange color shows distal-group comparisons, purple proximal, and grey others. Each grid box of five bars represents the percentage of participant responses for that fixed number of trials.

6.2.1. Sentence Conclusion Overview

We collected sentence conclusions from 312 unique participants ($M_{age} = 40.62$, $SD_{age} = 9.95$, 125 females), each writing their top three conclusions (in the order of which came to mind first). After excluding the nonsensical conclusions (4.2%), we ended up with 897 conclusions total. As shown in Figure 6.4 annotation 1, most were conclusions on main effects (45.4%), followed by pair-wise comparisons (31.9%), with few interactions (11.7%), superlatives (7.1%), and others (3.9%). More specifically, the prominence of pair-wise comparisons seemed to be largely driven by people more likely making pair-wise comparisons when the underlying data depicted no main effect but an interaction, see Figure 6.4 annotation 2 (more details discussed in Section 6.2.2).

As shown in Figure 6.4 annotation 3a+b, participants who viewed the bar charts with the additional color mapping seemed to follow the same pattern as those that viewed the bar charts without the color mapping. There was no significant effect of color mapping on the overall conclusion type or ranking order ($\chi^2 = 6.28$, $p = 0.18$) (Ripley et al., 2016), which suggested that color grouping did not have a significant effect on what conclusion categories people drew. Table 6.2 shows the odds ratio of participants writing sentence conclusions in each category (e.g., main effects) for each color mapping, using distal-group main effect conclusions as reference. Each number shows the percentage likelihood a viewer would make a specific category of sentence conclusion compared to making a distal-group main effect comparison for a specific color mapping. For example, for the no color mapping condition (first column), participants were 1.45 times more likely to make a proximal-group main effect conclusion compared to a distal-group main effect (one asterisk represents $p < 0.05$, two asterisks represents $p < 0.01$, three asterisks represents $p < 0.001$, and a dot represents $0.05 < p < 0.09$).

As shown in Table 6.3 and Figure 6.4 annotation 4a+b, multinomial logistic regression analysis (Ripley et al., 2016) predicting conclusion rankings with conclusion categories suggests that people were most likely to make a main effect comparison first. Those who did not make a main effect comparison were more likely to have made a pair-wise comparison than anything else ($p_{interaction} = 0.002$, $p_{superlative} = 0.001$, $p_{other} < 0.001$). By the second conclusion, people were more likely to make a main effect comparison or a pair-wise comparison than other conclusions, but the likelihood of them making main effect comparison and pair-wise comparison were not significantly different ($p = 0.74$). By the third conclusion, people became more likely to make main effect comparisons again, and similar to what they did for the first conclusion, those who did not make a main effect comparison more likely made pair-wise comparisons ($p_{interaction} = 0.02$, $p_{superlative} = 0.004$, $p_{other} < 0.001$). Participants were not likely to make interaction comparisons, and when they did, it was mostly as their third conclusion, see Figure 6.4 annotation 4c. More statistical details can be found in the supplementary materials.

6.2.2. Effect of Underlying Data Relation

A χ^2 test of independence with Bonferroni adjustment suggests that people more likely made main effect comparisons when they viewed data depicting one main effect and no interactions, or two main effects and no interactions, compared to data depicting an interaction but no main

Table 6.2. Bar Odds Ratio of Comparison Category by Color Mapping

Category	**No Mapping**	**Incongruent**	**Congruent**
main distal	1.00	1.00	1.00
main proximal	**1.45***	**1.39**•	**2.24*****
pair-wise distal	0.80	0.78	1.24
pair-wise proximal	0.64*	0.85	0.82
interaction distal	0.10***	0.42***	0.12***
interaction proximal	0.38***	0.37***	0.65·

Table 6.3. Bar Odds Ratio of Comparison Category by Order Compared

Category	First Rank	Second Rank	Third Rank
main effect	**1.00**	**1.00**	**1.00**
pair-wise	0.38***	**1.00**	0.93
interaction	0.12***	0.28***	0.48***
superlative	0.14***	0.17***	0.19***
others	0.04***	0.10***	0.15***

effects ($\chi^2 = 101.3$, $p < 0.001$), as shown in Figure 6.5 annotation 1. Post-hoc comparisons with Bonferroni corrections revealed that they were more likely to make pair-wise comparisons when viewing data depicting an interaction but no main effects ($z = 8.73$, $p < 0.001$), see Figure 6.5 annotation 2. Multinomial logistic regression analysis Ripley et al. (2016) predicting conclusion type with data relations further shows that people take into consideration the underlying data relation as they make comparisons of data. Table 6.4 shows the odds ratio of participants making a specific comparison for each of the three data relations, all using distal-group main effect comparisons as reference. We bolded the conclusion with the highest odds ratio for each color mapping per data relation.

Specifically, viewers were more likely to make pair-wise comparisons when they see charts depicting no main effect but an interaction. Referencing Table 6.4 top section row 3 and 4, when the chart has no color-spatial mapping or incongruent color-spatial mapping, participants more readily made proximal-group pair-wise comparisons. When the chart had congruent color-spatial mapping, participants more readily made distal-group pair-wise comparisons. When the data depicted one main effect, but no interaction, viewers were more likely to make proximal-group main effect comparisons, see Table 6.4 middle section row 2. When the data depicted two main effects and no interaction, although not significantly, viewers more readily made

Table 6.4. Bar Odds Ratio of Data Relations by Color Mapping

Data Relation	**Category**	**No Mapping**	**Incongruent**	**Congruent**
63.6emNo Main Effect, Yes Interaction	main distal	1.00	1.00	1.00
	main proximal	1.20	1.88	4.00**
	pair distal	2.00$^{\bullet}$	1.66	**5.40*** **
	pair proximal	**2.60***	**3.11****	3.60*
	int. distal	0.00	0.44	0.00
	int. proximal	1.10	1.11	2.20
63.6emOne Main Effect, No Interaction	main distal	1.00	1.00	1.00
	main proximal	**2.16****	**2.58****	**4.00****
	pair distal	0.83	**1.83$^{\bullet}$**	1.80
	pair proximal	0.22**	1.08	0.50
	int. distal	0.22**	0.83	0.19*
	int. proximal	0.44$^{\bullet}$	0.74	1.30
63.6emTwo Main Effects, No Interaction	main distal	1.00	**1.00**	1.00
	main proximal	**1.10**	0.85	**1.23**
	pair distal	0.37**	0.20***	0.23**
	pair proximal	0.24***	0.20***	0.42*
	int. distal	0.06***	0.28***	0.12***
	int. proximal	0.10***	0.05***	0.12***

proximal-group main effect comparisons for charts with no color mapping and congruent color-spatial mapping, and more distal-group main effect comparisons for charts with incongruent color-spatial mapping, see Table 6.4 bottom section row 1 and 2.

When we additionally consider the order in which the sentence conclusions were produced, for bar charts depicting *an interaction but no main effect*, viewers seemed to start making comparisons by looking at the main effect or the pair-wise comparisons, as shown in Figure 6.5 top left grid. By the second conclusion, most viewers shifted to make pair-wise comparisons. Very few participants made pair-wise proximity and edge comparisons (about 6%), so we lumped them together as ‘others’ in our figures. For those who did, these comparisons were often the second or third conclusion people drew, never the first. Notice there were few grey color patches in the top rows of Figure 6.5 (first conclusions), and more grey patches towards the bottom (third conclusions).

For bar charts depicting *one main effect and no interaction*, the majority of conclusions were main-effect comparisons. People were the most likely to make proximal-group main effect comparisons as their first conclusion, as shown in figure 6.5 annotation 3a (note that size grouping is collapsed here). In fact, participants mostly made proximal-group main effect comparisons for their second and third comparisons as well. Those who didn't make main-effect comparisons mostly made pair-wise distal-group comparisons, although not as many people made pair-wise comparisons when viewing this data compared to data depicting no main effect but an interaction.

For bar charts depicting *two main effects and no interaction*, people likely first made a proximal or distal main effect comparison. For their second conclusion, they either switched to make the other main effect comparison, or made a pair-wise comparison. Different from those who viewed bar charts depicting one main effect, participants who viewed bar charts depicting two main effects were equally likely to make distal-group and proximal-group pair-wise comparisons.

6.2.3. Effect of Size-Spatial Grouping

Because underlying data relations (whether there exists main effect/interaction in the data or not) impacted the conclusions people drew, we also investigate the influence of size-spatial grouping of the marks, which is typically determined by the similarity of the underlying data values. We suspect that size-spatial grouping would serve as a better predictor to conclusions viewers make because it captures visual differences that collapsing over data relation does not. For example, the size-spatial congruent (tall tall short short) and incongruent (tall short tall short) versions of the one main effect no interaction chart are visually different. This difference

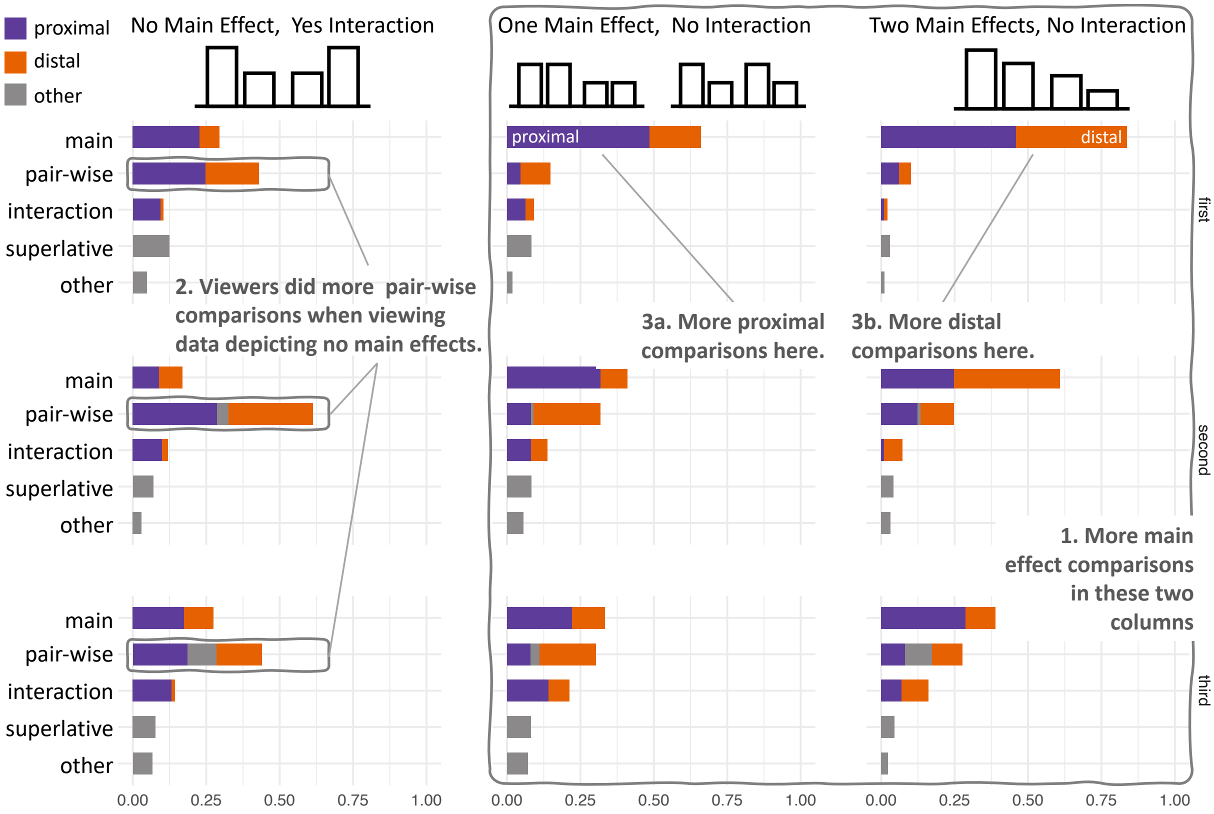


Figure 6.5. Conclusions by underlying data relations, aggregated across color mappings. The bar length presents the percentage of conclusions out of the total amount of conclusion per grid box (each grid box adds up to 100%).

is only captured when we look at size-spatial grouping, as they have the same underlying data relation.

We sorted the bar charts into four size-spatial grouping categories. As shown in Figure 6.6, size grouping could be congruent with spatial grouping such that the two similarly large bars are spatially proximate, and the two similarly small bars are spatially proximate. An example of this category is the "one main effect, no interaction" chart that is grouped by company, see Figure 6.2(b). The second category is a variation that arranges the bars in decreasing size. An example of this category is the "two main effects, no interaction" chart that is grouped by region, see Figure 6.2(d). This decreasing grouping still maintains spatial-size congruency, such that the two larger bars are spatially proximate, and the two smaller bars are spatially proximate, but the bar sizes are not as similar as those in the first category. The third category is when size and spatial grouping are incongruent, such that the spatially proximate two bars significantly differ in size (tall short tall short). An example of this is the "one main effect, no interaction" chart grouped by region, see Figure 6.2(c). The fourth category is a variation that is also size-spatial incongruent where the tallest two bars are the furthest away from each other. Examples include the "no main effect, yes interaction" charts, see Figure 6.2(e)

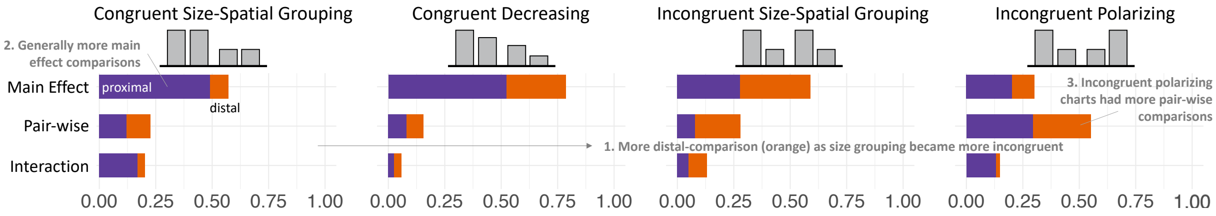


Figure 6.6. Size grouping and proximal-group and distal-group conclusions made by viewers, with aggregated color mapping. The values in each cell indicate the percentage of conclusions in that category out of the total amount of conclusions for the column.

People drew more proximal-group comparisons when size-spatial grouping is congruent, and more distal-group comparisons when size-spatial grouping is incongruent. As the size grouping become more incongruent with spatial grouping, people became more likely to draw a balanced amount of both proximal-group and distal-group conclusions. You can see that there were more orange (distal comparisons) moving from left (congruent size-spatial-grouping) to right (incongruent size-spatial grouping) in Figure 6.6 annotation 1. This shows a competition between spatial and size grouping, suggesting both drive what a viewer would conclude. If spatial grouping was always the stronger driver, you would expect participants to have always made proximal-group main comparisons. If size-grouping was the stronger driver, you would expect participants to have always made distal-group main comparisons when size-spatial grouping is incongruent. But this was not the case.

We observed similar dominance of main effect conclusions, with an increased number of pair-wise comparisons in incongruent polarizing charts, as shown in Figure 6.6 annotation 2+3. This makes sense as size-spatial grouping often depends on the underlying data relation, and incongruent polarizing charts are often depicting no main effects but an interaction, which suggests that size-spatial grouping is a better predictor than data relations as it captures the impact of data relations in a more nuanced way. Table 6.5 shows the results of a multinomial logistic regression analysis predicting conclusion type with size-spatial grouping. The likelihoods of participants drawing a particular conclusion under each size-spatial grouping were compared, using the likelihood of distal-group main comparison as reference (except for the incongruent color-spatial condition's congruent size-spatial column. No participant made distal-group main effect comparison when viewing the incongruent color-spatial chart, thus using distal-group main effect as a reference would result in exaggerated odds ratio). More details can be found in the supplementary materials.

As shown in Table 6.5 top section, for charts with *no color mapping*, viewers were most likely to make proximal-group main comparisons when they saw congruent size-spatial grouping charts, 2.44 times that of making a distal-group main comparison, and significantly more than other comparison types. Viewers were the most likely to make proximal-group pair-wise comparisons when they saw congruent decreasing charts, equally highly likely to make proximal- and distal-group main effect comparisons when they saw incongruent size-spatial

Table 6.5. Bar Odds Ratio of Color Mapping by Size-Spatial Grouping

Color	**Category**	**Congru.**	**Decreas.**	**Incongru.**	**Polariz.**
63emNo Mapping	main distal	1.00	1.00	**1.00**	1.00
	main proximal	**2.44***	1.46	**1.20**	1.20
	pair distal	0.33$^{\bullet}$	0.23*	0.80	2.00$^{\bullet}$
	pair proximal	0.11*	**3.08***	0.24**	**2.60***
	int. distal	5.00e-09	1.21e-08	0.24**	2e-07
	int. proximal	0.33$^{\bullet}$	0.08*	0.28**	1.10
63emIncongr. Color Spatial	main distal	2.29e-07***	1.00	**1.00**	1.00
	main proximal	**1.00**	1.83$^{\bullet}$	0.74	1.89
	pair distal	0.62	0.33$^{\bullet}$	0.49*	1.67
	pair proximal	0.77	0.17*	0.23***	**3.11****
	int. distal	0.31*	0.25*	0.37**	0.44
	int. proximal	0.54	0.08*	0.09***	1.11
63emCongr. Color Spatial	main distal	1.00	1.00	**1.00**	1.00
	main proximal	**26.0****	**3.29****	0.82	4.00**
	pair distal	2.00	0.29	0.71	**5.40*****
	pair proximal	4.00	0.57	0.29**	3.60*
	int. distal	4.44	0.14$^{\bullet}$	0.14***	5e-07***
	int. proximal	1.10*	0.14$^{\bullet}$	0.14***	2.20

grouping charts, and most likely too make proximal-group pair-wise comparisons when viewing incongruent polarizing charts. For charts with *incongruent color-spatial mapping*, viewers were most likely to make proximal-group main effect comparisons when viewing congruent size-spatial charts, slightly more likely to make proximal-group main effect comparisons when they saw decreasing charts, equally likely to make proximal- and distal-group main effect comparisons when they saw incongruent size-spatial grouping charts, and most likely to make proximal-group pair-wise comparisons when they saw incongruent polarizing charts. For charts with *congruent color-spatial mapping*, viewers were most likely to make proximal-group main effect comparisons when viewing congruent size-spatial charts and congruent decreasing charts, equally highly likely to make proximal- and distal-group main effect comparisons when they saw incongruent size-spatial grouping charts, and more likely to make distal-group pair-wise comparisons when viewing incongruent polarizing charts.

Although color grouping has a weak effect on comparisons, we did see an effect of of color on comparisons made when the size-grouping is *incongruent polarizing* (see right-most column, the middle and bottom sections in Table 6.5). When the chart had incongruent color-spatial grouping, participants more readily made *proximal-group* pair-wise comparisons, 3.11 times that of main distal comparisons, but when the chart had congruent color-spatial grouping, participants more readily made *distal-group* pair-wise comparisons, 5.40 times that of main distal comparisons.

Overall, the effects of size-spatial grouping seemed to be more granular and able to account for the effect of data relation, suggesting that size grouping is a more effective predictor guiding viewer conclusions. Next, we examine potential interactions between color- and size-spatial grouping on comparisons made.

6.2.4. Interaction between Color, Size, and Spatial Grouping

We compared the odds ratio of participants comparing the *two spatially proximate groups* for each color mapping, across the four possible size-spatial groupings, as shown in the top section of Table 6.6, using no color mapping as reference. Overall, besides that viewers were more 1.57 times more likely to make proximal comparisons when viewing congruent size-spatial charts with congruent color-spatial mapping, everything else was approximately equal. The middle section documents the odds ratio of participants comparing the *two distal groups*. Again, besides that viewers were only a quarter as likely to make distal comparisons when viewing congruent size-spatial charts with congruent color-spatial mapping, everything was approximately equal. This again suggests that color had negligible effects on guiding comparisons.

Table 6.6. Bar Odds Ratio Aggregated Color, Size, and Spatial Grouping

Proximal Comparisons	**Congru.**	**Decreasing**	**Incongru.**	**Polarizing**
No Color Mapping	1.00	1.00	1.00	1.00
Color-Spatial Incongru.	1.15	1.04	0.86	1.12
Color-Spatial Congru.	1.57*	1.16	0.81	1.00

Distal Comparisons	**Congru.**	**Decreasing**	**Incongru.**	**Polarizing**
No Color Mapping	1.00	1.00	1.00	1.00
Color-Spatial Incongru.	0.99	1.18	1.27	0.93
Color-Spatial Congru.	0.24*	0.63	1.01	1.07

Total	**Congru.**	**Decreasing**	**Incongru.**	**Polarizing**
proximal	2.11**	1.00	0.40***	1.70***
distal	0.47**	0.99	2.48***	0.59***

The bottom section directly compares the odds ratio of participants making a proximal- versus a distal-comparison across size-spatial groupings, collapsing over color-spatial grouping. This serves as a proxy to estimate and compare the effect of spatial and size grouping on sentence conclusion generations. For instance, larger likelihoods of making proximal comparisons suggests a stronger effect of spatial grouping on eliciting comparisons. Participants were more likely to make proximal comparisons overall. They were approximately twice as likely to make proximal comparisons, and half as likely to make distal when they viewed congruent size-spatial charts and incongruent polarizing charts. This pattern flipped for when participants viewed incongruent size-grouping charts. When they viewed congruent decreasing chart, they were equally likely to make proximal- and distal-comparisons.

6.2.5. Discussion

In general, viewers are the most likely to make main effect comparisons. They made pair-wise comparisons more readily when the underlying data is depicting an interaction but no main effect, or when the size-spatial grouping is incongruent and polarizing (these patterns tend to co-occur, see Figure 6.2(e). Viewers rarely generated interaction comparisons, even when the underlying data was showing only an interaction with no main effects. When viewers saw this type of charts, they made pair-wise comparisons between spatially adjacent data values. Surprisingly, varying color mapping in bar visualizations had only a small impact on what comparisons a viewer would make. There was one exception, where color influenced which pair-wise comparison a viewer would make when the data depicted an interaction but no main effect, such that incongruent color-spatial arrangements increased the likelihood of viewers making proximal pair-wise comparisons. On the other hand, size-spatial grouping have a much bigger impact on guiding viewer conclusions. Overall, viewers are approximately 31% more likely to generate sentence conclusions comparing spatially proximate data values than spatially separated ones, and this tendency can be doubled or halved depending on the size-spatial grouping.

6.3. Experiment 2: Scatterplot Visualizations

The previous investigation of bar charts revealed a grouping strength ranking of spatial first, size second, and color third. Experiment 2 compared two other grouping cues – color versus shape – in scatterplots. We followed the two binary factor setup in the bar visualization experiment and created scatterplots visualizing the performance of two types of car engine (straight and v-shaped) on quarter mile time and fuel economy for two transmission types (automatic and manual).

Similar to the bar experiment, participants viewed scatterplots and typed the top three conclusions they would draw. We tested four configurations of the scatter plots, as shown in Figure

6.7. We used the mtcars dataset, identified two factors - engine and transmission, which are encoded either with color (default ggplot teal 00BFC4 and red F8766D) or shape (default circle or triangles) - and plotted their corresponding fuel economy and quarter mile time. We crossed the color and shape encodings, and counterbalanced the x- and y-axes to produce the four configurations: two have engine encoded using color, and two have transmission encoded using color; two have quarter mile time plotted on the x-axis, and two have fuel economy plotted on the x-axis. Participants were excluded following the same exclusion criteria as Experiment 1. We excluded 2 participants and 11 nonsensical or empty responses, and are left with 69 participants ($MeanAge = 39.2$, $SD = 11.31$, 23 females) and 202 conclusions.

6.3.1. Response Categorization

We categorized the conclusions participants made from the scatterplots following a similar taxonomy as the bar chart experiment. Participants drew main effects and pair-wise comparisons, and identified superlatives. The scatterplot format also enabled participants to draw additional categories of conclusions, including correlations and distributions. No one compared interactions between any of the factors.

Main Effects: We categorized conclusions comparing how each factor performed on either the x- or y-axis as a main effect comparison. Comparing manual to automatic transmission or straight to v-shaped engine performance on either axis (quarter mile time or fuel economy) are main effect comparisons. These comparisons could happen either using color or shape grouping. For instance, if manual transmissions were colored red and automatic transitions were colored blue, concluding "manual transmissions have a faster quarter mile time than automatics"

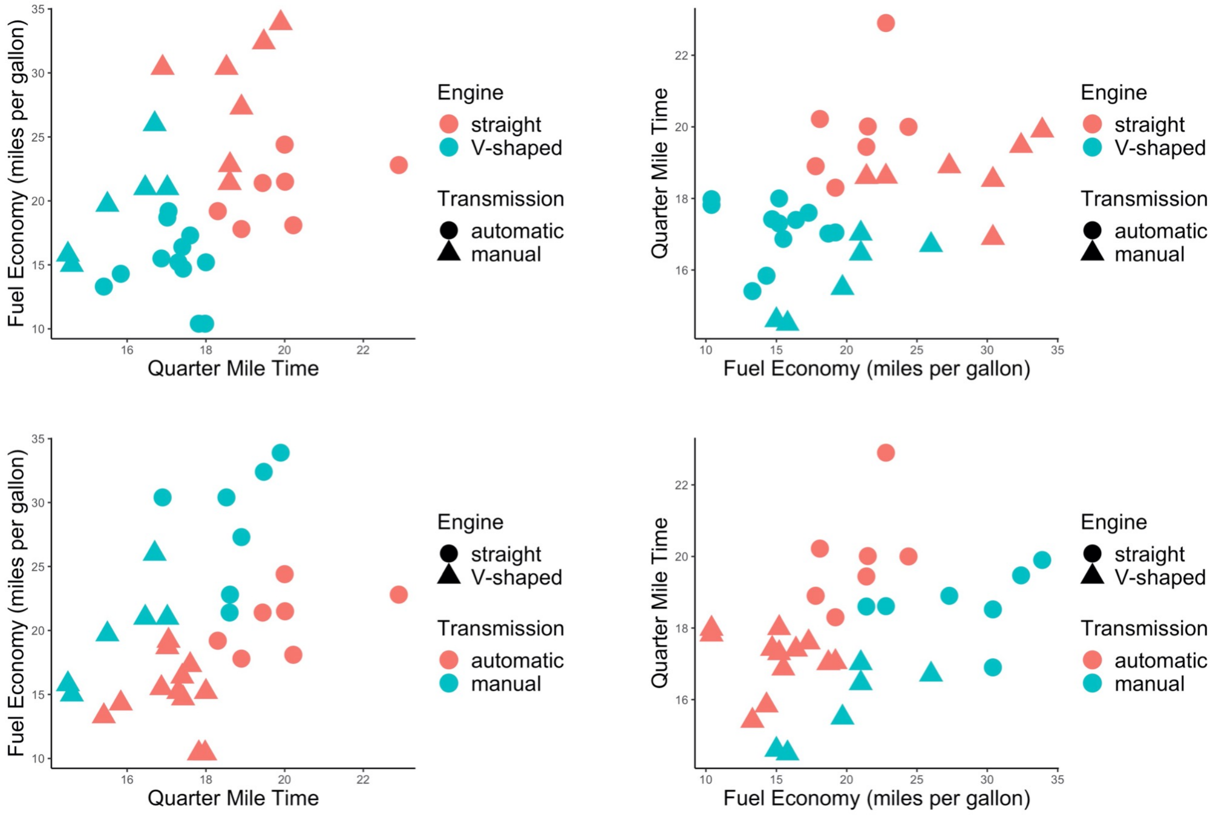


Figure 6.7. Scatterplot stimuli used in Experiment 2, depicting two factors encoded with color or shape. The same data was presented and counterbalanced with which factor was encoded to which channel (color or shape) as well as switching of the x and y variables.

would be a color comparison. If transmission were encoded by shape instead, the previous example would be a shape main effect comparison.

Pair-Wise Comparisons: Conclusions comparing one level of a factor while keeping the levels of the other factor constant is categorized as pair-wise conclusions. These comparisons can be across color or across shape. For instance, "V-shaped engines with manual transmissions are slower than straight engines with manual transmissions." For the condition in which engines were assigned shape, this would be an across-shape comparison, comparing two shapes across the same color. For the condition in which engines were assigned color, this would be an across-color comparison, comparing different colors across the same shape. Therefore, a pair-wise comparison counts towards both color and shape.

Superlatives: Some participant conclusions were about the single best performing combination of engine and transmission. For example, a participant concluded, "A straight manual transmission will give you the highest fuel economy." Unique to the scatterplot condition, some superlative codes were about performance in the x-axis, while some were about performance in the y-axis, and some were about overall performance across both axes. In this example, because the participant mentioned both engine (straight) and transmission (manual), we consider it as having mentioned both color and shape.

Correlations: Participants also commented on the relation between the x- and y-axes. Conclusions in this category often included descriptions of a correlation between fuel economy and quarter mile time. "The faster the car, the more fuel economy it gets" is an example that indicated the viewer's understanding of the general positive trend in the data.

Distributions: This category included any conclusion that centered around the spread of the data rather than the performance differences between factors. It contains three sub-categories: count, range, and variation. Count pertains to the number of each factor. For instance, "there are more V-shaped than straight engines." Range mentions the performance of the data points but makes a value statement, rather than a comparison, such as "there are no engines that are over 35 [miles per gallon] for fuel economy." While this mentions the performance of engines, there are no comparisons being made. Instead, the conclusion refers again to the spread of the data. Variation is about data variability, such as "V-Shaped manuals seem to perform consistently."

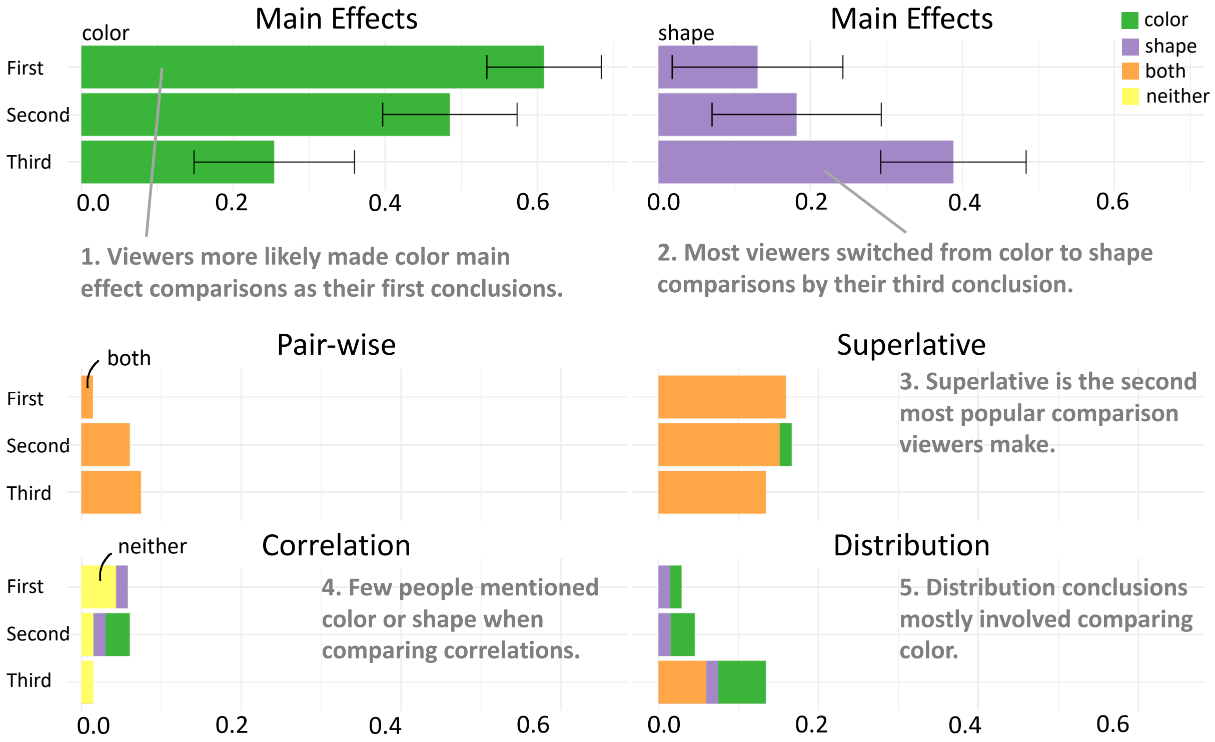


Figure 6.8. Comparison category by features compared in scatterplots. Top row shows the distribution of people comparing color versus shape for main effects. Bottom two rows show stacked distribution of features compared (color, shape, both, neither) for other categories. Bar length shows the number of comparisons made in a specific category out of all comparisons made for that rank. Error bars show standard error.

6.3.2. Sentence Conclusions for Scatterplots

Overall, viewers most likely drew main effect conclusions (68.3%), followed by superlative conclusions (15.3%), distribution conclusions (6.9%), pair-wise conclusions (5.0%), and least likely correlations (4.5%). Table 6.7 shows the odds ratio of a particular category of conclusion being drawn for each rank order, using main effect conclusion as reference. Similar to that in bar charts, participants were significantly more likely to make main effect comparisons across all three conclusions they made. For example, looking at the top table column for the 'First' rank, pair-wise comparisons were only 2% as likely to be mentioned as the first conclusion compared to main effects.

Table 6.7. Scatterplot Odds Ratio of Comparison Category and Order Compared

Conclusion	**First**	**Second**	**Third**
main effect	1.00	1.00	1.00
pair-wise	0.02***	0.09***	0.12***
superlative	0.22***	0.25***	0.21***
correlation	0.08***	0.09***	0.02***
distribution	0.04***	0.07***	0.37***

Comparison	**First**	**Second**	**Third**
compare color	**1.00**	**1.00**	1.00
compare shape	0.26***	0.38**	**1.28**
mention both	0.28***	0.38**	0.86
mention neither	0.07***	0.03***	0.05**

Participants were more likely to compare color as their first and second comparison, see Figure 6.8 annotation 1 and Table 6.7 bottom section. A majority of people shifted to compare shape or mentioned both color and shape by their third comparison, as shown in Table 6.7 and Figure 6.8 annotation 2. More specifically, as shown in the bottom of Table 6.7, viewers compared shape only 0.26 times as likely as they compared color in the first sentence conclusion. But by the third sentence conclusion they generated, they became 1.28 times more likely to compare shape relative to comparing color. In multi-class scatterplots we tested, color and shape are salient features such that very few people generated sentences without mentioning either color or shape.

As shown in Table 6.8, overall color was the most compared feature across all categories. Viewers were half as likely to compare shape when making main effect comparisons. When viewers made pair-wise comparisons[1], because this relied on selecting one dimension (either

[1] The low-N in the pair-wise comparison category created super large and super small odds ratios.

Table 6.8. Scatterplot Odds Ratio of Category and Features Compared

Conclusion	Color	Shape	Both	None
main effect	1.00	0.51***	0.00***	0.00
pair-wise[1]	1.00	0.00	0.00	1.26e+41***
superlative	1.00	0.00***	30.00***	0.00***
correlation	1.00	1.00	0.00	2.50
distribution	1.00	0.43	0.57	0.00

color or shape) and comparing how the levels of the other dimension differed on the selected dimension, both color and shape must have been mentioned. This resulted in a skewed, large odds ratio in the 'Both' column. While there were only 10 conclusions that were pair-wise comparisons in our data, eight of them compared the two groups with different shapes of the same color, while only two compared the two groups of the same shape with different colors. While there were not enough data points to make reliable statistical inferences, it seemed to support the previous analysis that color was a stronger grouping cue than shape. We expand on this observation in Section **??** to further discuss potential implications.

We also examined whether viewers more readily compared visual groups based on the x- or y-axis value. Chi-squared test for given probabilities comparing the number of times people made x- or y- axis based comparisons suggested that, with post-hoc bonferroni adjustmets Hervé and Hervé (2020), people were equally likely to compare values on the x- and y-axis ($\chi^2 = 63.72$, $p_{xy(posthoc)} = 1.00$).

6.4. Discussion

Overall, viewers were the most likely to make main effect comparisons. When viewing bar charts, they made pair-wise comparisons more readily when the underlying data depicted no main effect but an interaction, or when the data was organized with polarizing incongruent size-spatial grouping (they tend to co-occur, as shown in Figure 6.2(e)). They rarely generated interaction comparisons, even when the underlying data showed only an interaction with no main effects, corroborating findings in psychology suggesting that interaction is cognitively difficult to discern and express Halford et al. (2005); Shah and Freedman (2011). Varying color mapping in bar visualizations had a negligible impact on what comparisons a viewer would make. There was a competition between size- and spatial-grouping, such that viewers were approximately 31% more likely to generate sentence conclusions comparing spatially proximate data values than spatially separated ones, and this tendency could be doubled or halved depending on the size-spatial grouping. In multi-class scatterplots, color was more likely to be compared first than shape, but by the third conclusion, shape became more likely to be compared.

I provide some design guidelines to help inform visualization researchers and practitioners on how to arrange their data values. For a bar chart, place the values that your audience should compare (or treat as a group) next to each other. Making them different colors may not hurt, but the added benefit is likely minimal. Beware that viewers are likely to group values of similar sizes, whether or not they should do so. For a scatterplot, your viewers will prioritize comparing different colors over shapes.

Most current work on visual comparison focuses on the efficiency or precision of a given comparison. We argue that these factors are important, but irrelevant if a visualization viewer never makes that comparison in the first place. In our real-world experience, the biggest delays in understanding a new visualization are knowing how to read a visualization and knowing which patterns one should pay attention to. When communicating data, analysts tend to assume that the viewer sees what they see (Xiong et al., 2019), instead of designing the visualization to push the viewer toward seeing the ‘right’ pattern. We hope that refinement of the type of model developed here can lead to a set of guidelines or concrete tools to help them achieve that goal.

CHAPTER 7

Conclusion

I've demonstrated that visualizations are one form of ambiguous figures such two people looking at the same dataset could come to different conclusions. Visualization design can also influence the *type* of information people extract and the inferences people make from data. This corroborates existing findings in visual analytics and decision making that suggest choosing the appropriate visualization designs can improve the accuracy and efficiency of data interpretation. But this line of work is far from done. Designing a visualization is like mixing music with a soundboard — every design decision you make is a switch or a knob. With the turn of each nob and switch, youâll generate something new. There are so many decisions that can go into creating a visualizations, and I'm excited to continue exploring the perceptual and cognitive affordances of these design decisions.

I encourage future researchers to bridge work in human cognition and data visualization to shed light on the impact of information visualization on data communication and decision-making. As we increasingly rely on data to understand, communicate, and make decisions, we need to further understand how our brains work to extract critical values, statistics, and patterns needed to make decisions about data, so we can design more effective visualizations.

References

Allbritton, D. W., McKoon, G., and Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(3):714.

Alvarez, G. A. and Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4):392–398.

Ancker, J. S., Senathirajah, Y., Kukafka, R., and Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, 13(6):608–618.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2):157–162.

Attneave, F. (1971). Multistability in perception. *Scientific American*, 225(6):62–71.

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., and Brooks, C. (2010). Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2573–2582.

Bates, D. (2005). Fitting linear mixed models in r. *R news*, 5(1):27–30.

Bernstein, D. M., Atance, C., Loftus, G. R., and Meltzoff, A. (2004). We saw it all along: Visual hindsight bias in children and adults. *Psychological Science*, 15(4):264–267.

Birch, S. A. and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5):382–386.

Blank, H., Fischer, V., and Erdfelder, E. (2003). Hindsight bias in political elections. *Memory*, 11(4-5):491–504.

Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P. W., Reppa, I., Floridi, L., and Chen, M. (2012). An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768.

Bornstein, M. H. and Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, 46(3):207–222.

Bostock, M., Carter, S., Amanda, C., and Quealy, K. (2012). One report, diverging perspectives.

Boy, J., Pandey, A. V., Emerson, J., Satterthwaite, M., Nov, O., and Bertini, E. (2017). Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5462–5474. ACM.

Brady, T. F. and Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3):384–392.

Bromme, R. and Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist*, 49(2):59–69.

Camerer, C., Loewenstein, G., and Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, 97(5):1232–1254.

Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Cassar, G. and Craig, J. (2009). An investigation of hindsight bias in nascent venture activity. *Journal of Business Venturing*, 24(2):149–164.

Chong, S. C. and Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4):393–404.

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

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

Correll, M. and Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151.

Cozby, P. C. (2007). *Methods in behavioral research*. McGraw-Hill.

Croxton, F. E. and Stryker, R. E. (1927). Bar charts versus circle diagrams. *Journal of the American Statistical Association*, 22(160):473–482.

Dimara, E., Bailly, G., Bezerianos, A., and Franconeri, S. (2019). Mitigating the attraction effect with visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):850–860.

Dimara, E., Bezerianos, A., and Dragicevic, P. (2017). The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics*, 23(1):471–480.

Dixon, G. N. and Clarke, C. E. (2013). Heightening uncertainty around certain science: Media coverage, false balance, and the autism-vaccine controversy. *Science Communication*, 35(3):358–382.

Eells, W. C. (1926). The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, 21(154):119–132.

Egeth, H. E., Leonard, C. J., and Leber, A. B. (2010). Why salience is not enough: Reflections on top-down selection in vision. *Acta psychologica*, 135(2):130.

Epley, N. and Waytz, A. (2010). Mind perception. *Handbook of social psychology*.

Furnham, A. and Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42.

Gilovich, T., Savitsky, K., and Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2):332.

Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C. D., and Roberts, J. C. (2011). Visual comparison for information visualization. *Information Visualization*, 10(4):289–309.

Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics*, 19(12):2277–2286.

Grice, H. P., Cole, P., Morgan, J. L., et al. (1975). Logic and conversation. *1975*, pages 41–58.

Guo, J. (2016). Researchers have debunked one of our most basic assumptions about how the world works.

Halford, G. S., Baker, R., McCredden, J. E., and Bain, J. D. (2005). How many variables can humans process? *Psychological science*, 16(1):70–76.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American psychologist*, 53(4):449.

Harrison, L., Yang, F., Franconeri, S., and Chang, R. (2014). Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics*, 20(12):1943–1952.

Hart, G. (1996). The five w's: An old tool for the new task of task analysis. *Technical communication*, 43(2):139–145.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. (2011). Statistical inference for stochastic simulation models–theory and application. *Ecology Letters*, 14(8):816–827.

Hegarty, M. (2005). Multimedia learning about physical systems. *The Cambridge handbook of multimedia learning*, pages 447–465.

Herman, I., Melançon, G., and Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43.

Hervé, M. and Hervé, M. M. (2020). Package ârvaidememoireâ.

Hsieh, F. Y., Bloch, D. A., and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine*, 17(14):1623–1634.

Hullman, J., Adar, E., and Shah, P. (2011). The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1461–1470. ACM.

Hullman, J., Kay, M., Kim, Y.-S., and Shrestha, S. (2018). Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE transactions on visualization and computer graphics*, 24(1):446–456.

Huttenlocher, J., Hedges, L. V., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3):352.

Ibrahim, A., Seifert, C., Adar, E., and Shah, P. (2016). Using graphs to debias misinformation.

Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166.

Kahan, D. M., Peters, E., Dawson, E. C., and Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1):54–86.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kale, A., Nguyen, F., Kay, M., and Hullman, J. (2019). Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics*, 25(1):892–902.

Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. (2016). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proc. of the 2016 CHI*, pages 5092–5103.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11–IT.

Keysar, B. and Henly, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, 13(3):207–212.

Khan, M. and Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, 34(1):1–14.

Kim, J. G. and Biederman, I. (2012). Greater sensitivity to nonaccidental than metric changes in the relations between simple shapes in the lateral occipital cortex. *NeuroImage*, 63(4):1818–1826.

Kim, Y., Wongsuphasawat, K., Hullman, J., and Heer, J. (2017). GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing. *Proc. of ACM CHI 2017*.

Kim, Y.-S., Walls, L. A., Krafft, P., and Hullman, J. (2019). A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 682:1–682:14, New York, NY, USA. ACM.

Klein, N. and OâBrien, E. (2018). People use less information than they think to make up their minds. *Proceedings of the National Academy of Sciences*, 115(52):13222–13227.

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

Köhler, W. and Wallach, H. (1944). Figural after-effects. an investigation of visual processes. *Proceedings of the American Philosophical Society*, 88(4):269–357.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Mit Press.

Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., and Gabrieli, J. D. (1989). Evidence for two types of spatial representations: hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):723.

Kosslyn, S. M., Murphy, G. L., Bemesderfer, M. E., and Feinstein, K. J. (1977). Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, 106(4):341.

Kranjec, A., Lupyan, G., and Chatterjee, A. (2014). Categorical biases in perceiving spatial relations. *PloS One*, 9(5):e98604.

Leary, M. R. (2007). *The curse of the self: Self-awareness, egotism, and the quality of human life*. Oxford University Press.

Lew, T. F. and Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the weber noise of relative positions. *Journal of Vision*, 15(4):10–10.

Lovett, A. and Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, 28(10):1408–1418.

Márquez, C., Izquierdo, M., and Espinet, M. (2006). Multimodal science teachers' discourse in modeling the water cycle. *Science Education*, 90(2):202–226.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., and Barberia, I. (2015). Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6:888.

McKenzie, G., Hegarty, M., Barrett, T., and Goodchild, M. (2016). Assessing the effectiveness of different visualizations for judgments of positional uncertainty. *International Journal of Geographical Information Science*, 30(2):221–239.

Michal, A. L. and Franconeri, S. L. (2017). Visual routines are associated with specific graph interpretations. *Cognitive Research: Principles and Implications*, 2(1):1–10.

Michal, A. L., Uttal, D., Shah, P., and Franconeri, S. L. (2016). Visual routines for extracting magnitude relations. *Psychonomic bulletin & review*, 23(6):1802–1809.

Moere, A. V., Tomitsch, M., Wimmer, C., Christoph, B., and Grechenig, T. (2012). Evaluating the effect of style in information visualization. *IEEE transactions on visualization and computer graphics*, 18(12):2739–2748.

Newman, A., Bylinskii, Z., Haroz, S., Madan, S., Durand, F., and Oliva, A. (2018). Effects of title wording on memory of trends in line graphs. *Journal of Vision*, 18(10):837.

News, B. (2018). Heatwave: Is there more crime in hot weather?

Newton, E. (1990). *Overconfidence in the communication of intent: Heard and unheard melodies*. PhD thesis, Department of Psychology, Stanford University, Stanford, CA.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Nothelfer, C. and Franconeri, S. (2019). Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320.

Nyhan, B. and Reifler, J. (2019). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 29(2):222–244.

Oestermeier, U. and Hesse, F. W. (2000). Verbal and visual causal arguments. *Cognition*, 75(1):65–104.

Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M., and Bertini, E. (2014). The persuasive power of data visualization. *IEEE transactions on visualization and computer graphics*, 20(12):2211–2220.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., and Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7):739.

Parsons, P. C. (2018). Conceptual metaphor theory as a foundation for communicative visualization design. In *IEEE VIS Workshop on Visualization for Communication (VisComm 2018)*.

Picon, E. and Odic, D. (2017). Finding maximal and minimal elements in a set is capacity-unlimited and massively-parallel. *Journal of Vision*, 17(10):1284–1284.

Pohl, R. and Haracic, I. (2005). Der rückschaufehler bei kindern und erwachsenen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 37(1):46–55.

Qualtrics, I. (2013). Qualtrics. *Provo, UT, USA*.

Qualtrics, L. (2014). Qualtrics [software]. *Provo, Utah*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radio, N. P. (2011). Money buys happiness.

Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review*, 100(2):573–78.

Rhodes, R. E., Rodriguez, F., and Shah, P. (2014). Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5):1432.

Ripley, B., Venables, W., and Ripley, M. B. (2016). Package ânnetâ. *R package version*, 7:3–12.

Roberts, B., Harris, M. G., and Yates, T. A. (2005). The roles of inducer size and distance in the ebbinghaus illusion (titchener circles). *Perception*, 34(7):847–856.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Roese, N. J. and Vohs, K. D. (2012). Hindsight bias. *Perspectives on psychological science*, 7(5):411–426.

Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.

Ruginski, I. T., Boone, A. P., Padilla, L. M., Liu, L., Heydari, N., Kramer, H. S., Hegarty, M., Thompson, W. B., House, D. H., and Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172.

Saket, B., Endert, A., and Demiralp, C. (2018). Task-based effectiveness of basic visualizations. *IEEE transactions on visualization and computer graphics*.

Sarikaya, A. and Gleicher, M. (2018). Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1). to appear (InfoVis 2017).

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148.

Shah, P. and Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578.

Shah, P., Michal, A., Ibrahim, A., Rhodes, R., and Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging? In *Psychology of Learning and Motivation*, volume 66, pages 251–299. Elsevier.

Shaklee, H. and Elek, S. (1988). Cause and covariate: Development of two related concepts. *Cognitive Development*, 3(1):1–13.

Shiffrin, R. M. (2016). Drawing causal inference from big data. *Proceedings of the National Academy of Sciences*, 113(27):7308–7309.

Shtulman, A. and Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2):209–215.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Snow, J. and Mann, M. (2013). Qualtrics survey software: handbook for research professionals. *Provo, UT: Qualtrics Labs*.

Spence, I. and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., et al. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349:g7015.

Sweeny, T. D., Grabowecky, M., and Suzuki, S. (2011). Simultaneous shape repulsion and global assimilation in the perception of aspect ratio. *Journal of vision*, 11(1):16–16.

Szafir, D. A., Haroz, S., Gleicher, M., and Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11.

Tufte, E. R. (2001). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, B. (2014). Visualizing thought. In *Handbook of human centric visualization*, pages 3–40. Springer.

Valdez, A. C., Ziefle, M., and Sedlmair, M. (2018). Priming and anchoring effects in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):584–594.

Wade, N. J. (1994). A selective history of the study of visual motion aftereffects. *Perception*, 23(10):1111–1134.

Ward, A., Ross, L., Reed, E., Turiel, E., and Brown, T. (1997). *Naive realism in everyday life: Implications for social conflict and misunderstanding*. Lawrence Erlbaum Association Hillsdale, NJ.

Ware, C. (2012). *Information visualization: Perception for design*. Elsevier.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39.

Xiong, C., van Weelden, L., and Franconeri, S. (2019). The curse of knowledge in visual data communication. *IEEE transactions on visualization and computer graphics*.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.

Zacks, J. and Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079.

Zhou, H., Majka, E. A., and Epley, N. (2017). Inferring perspective versus getting perspective: Underestimating the value of being in another personâs shoes. *Psychological science*, 28(4):482–493.

Zwick, R., Pieters, R., and Baumgartner, H. (1995). On the practical significance of hindsight bias: The case of the expectancy-disconfirmation model of consumer satisfaction. *Organizational behavior and human decision processes*, 64(1):103–117.