An Investigation Into the Cognitive Processing of Magnitude in Data Visualisations

Duncan Bradley

Data visualisations are effective communication tools which leverage the power of the human visual and cognitive systems. However, those viewing visualisations are at the mercy of data visualisation design choices, which can substantially influence interpretations of presented information. Interpreting data visualisations does not just involve accurately comprehending values, but often also involves drawing inferences about data and making subjective judgements. Therefore, developing an understanding of cognitive processing of data visualisations is important for guiding the construction of effective and faithful representations of data.

Data visualisations can convey many different features of a dataset. One such feature is the absolute magnitude of plotted values: how large or small they are. This thesis investigates cognitive mechanisms involved in judging absolute magnitude in data visualisations, revealing the effects of design choices on interpretation. Three sets of experiments explore participants' subjective judgements of magnitude in a variety of visualisation formats.

The first set of experiments in this thesis (three experiments) explores the role of axis limits in informing magnitude judgements. Manipulating axis limits in dot plots causes the same data points to appear near the top or bottom of the visualisation. Participants' responses revealed an association between higher positions and higher magnitude ratings, indicating a bias in interpretation. A further experiment employing dot plots with inverted y-axes indicated that impressions of magnitude were driven primarily by the relative positions of data points within axis limits, not their absolute physical positions.

The second experiment in this thesis extends inquiry into the role of axis limits to choropleth maps. Manipulating the limits of accompanying colour legends alters the framing of presented data, without changing how plotted values appeared. Extending the colour legend's upper limit beyond the maximum value in the dataset resulted in lower magnitude ratings. This demonstrates that interpretations of absolute magnitude are informed by surrounding context, not just by the appearance of plotted values.

The final set of experiments in this thesis (two experiments) explores how additional knowledge about plotted data can inform interpretations of magnitude. Denominators provide numerical context relevant to magnitude judgements. Extending a bar chart's axis beyond plotted data to incorporate a denominator value elicited lower magnitude ratings, compared to bar charts' default settings. Omitting denominator information from accompanying text substantially increased this bias. This illustrates that additional knowledge about a dataset diminishes the roles of axis limits in informing impressions of magnitude.

This work was conducted with a focus on computational reproducibility. In addition to sharing data and analysis code, this involved facilitating the ability to reproduce the computational environment used for analysis. This approach, which increases openness and transparency in research, is also discussed in detail.

Through experimental research, this thesis reveals how the framing of values within axes informs judgements of their absolute magnitudes. The results provide insight into the cognitive

processing of magnitude in data visualisations, wherein context shapes viewers' inferences. This illustrates how inevitable subjectivity in data visualisation design can influence a data visualisation's appearance and its message. Data visualisation designers should consider the graphical representation of absolute magnitude and, where appropriate, employ axes ranges which faithfully convey this aspect of data.

Table of contents

1 Introduction **5**
- 1.1 Research Motivation and Objective . . . 5
- 1.2 Overview of the Thesis . . . 6
- 1.3 Contributions . . . 6

2 Literature Review **7**
- 2.1 Data Visualisation Formats and the Grammar of Graphics . . . 9
- 2.2 Data Visualisation Software and the Influence of Default Settings . . . 10
- 2.3 Popular Guidance on Effective Data Visualisation Design . . . 11
- 2.4 Rigorous Data Visualisation Research . . . 13
- 2.5 Perceptual Precision in Data Visualisations . . . 14
- 2.6 Beyond Perceptual Precision . . . 15
- 2.7 Manipulating Axes in Data Visualisations . . . 17
- 2.8 Misleading Data Visualisations . . . 19
- 2.9 Data Visualisation Literacy . . . 20
- 2.10 Interpreting Absolute Magnitude . . . 21
- 2.11 Structure of the Thesis . . . 22

3 Experimental and Computational Methodology **24**
- 3.1 Experimental Psychology . . . 24
- 3.2 Analysis Methodology . . . 25
- 3.3 Reproducibility . . . 25
- 3.4 Sharing Code and Data . . . 26
  - 3.4.1 The Importance of Public Sharing . . . 27
- 3.5 Effective Programming Practices . . . 28
  - 3.5.1 Literate Programming and Dynamic Documents . . . 29
- 3.6 Computational Environments . . . 29
  - 3.6.1 Capturing Computational Environments Using Containers . . . 30
  - 3.6.2 Rocker for Capturing R Environments . . . 30
  - 3.6.3 Comparing Containers with Virtual Machines . . . 31
- 3.7 Pragmatism Over Perfectionism . . . 31
- 3.8 The Approach to Reproducibility in This Thesis . . . 32
  - 3.8.1 Data, Code, and Dynamic Documents . . . 32
  - 3.8.2 Docker Containers . . . 32
  - 3.8.3 Experiment Resources . . . 32

3.9 Conclusion . . . 33

Preface to Chapter 3 **34**
Rationale For Experimental Design . . . 34

4 Magnitude Judgements Are Influenced by Data Points' Relative Positions Within Axis Limits **35**
4.1 Introduction . . . 35
4.1.1 Overview . . . 36
4.2 Related Work . . . 36
4.2.1 Effects of Axis Limits on Comparison Judgements . . . 36
4.2.2 Effects of Axis Limits on Magnitude Judgements . . . 38
4.2.3 Judgements of Event Outcomes . . . 39
4.2.4 Data Visualisation Literacy . . . 39
4.3 Experiments . . . 40
4.3.1 Experiment 1 . . . 40
4.3.2 Experiment 2 . . . 45
4.3.3 Experiment 3 . . . 51
4.4 General Discussion . . . 54
4.4.1 Relationship to Prior Work . . . 55
4.4.2 Additional Findings . . . 56
4.4.3 Limitations and Future Directions . . . 56
4.4.4 Conclusion . . . 57
4.5 References . . . 57
References . . . 57

Postscript to Chapter 3 **60**
Overview . . . 60
Rationale for Experimental Design . . . 60
Summary of Findings . . . 61

Preface to Chapter 4 **62**

5 Choropleth Maps Can Convey Absolute Magnitude Through the Range of the Accompanying Colour Legend **63**
5.1 Introduction . . . 63
5.2 Related Work . . . 65
5.2.1 Choropleth Maps . . . 65
5.2.2 Communicating Absolute Magnitude Through Data Visualisation . . . 66
5.2.3 Colour Legends . . . 68
5.3 Methodology . . . 69
5.3.1 Outline . . . 69
5.3.2 Pre-Registration . . . 71

5.3.3 Design . . . 71
5.3.4 Participants . . . 71
5.3.5 Procedure . . . 73
5.3.6 Materials . . . 75
5.4 Analysis . . . 76
5.4.1 Analysis Methods . . . 76
5.4.2 Part 1: Participants Satisfying Attention Check Criteria (N = 100) . . . 76
5.4.3 Part 2: All Participants (N = 165) . . . 78
5.4.4 Exploratory Analysis . . . 80
5.5 Discussion . . . 81
5.5.1 Additional Analyses . . . 82
5.5.2 Relationship to Prior Work . . . 83
5.5.3 Limitations and Future Research Directions . . . 84
5.5.4 Implications . . . 85
5.6 Conclusion . . . 86
5.7 References . . . 86
References . . . 86

Postscript to Chapter 4 **90**
Overview . . . 90
Rationale for Experimental Design . . . 90
Summary of Findings . . . 91
Relationship to the Study in Chapter 3 . . . 91

Preface to Chapter 5 **92**

6 Axis Limits and Denominator Information Influence Magnitude Ratings in Bar Charts **93**
6.1 Introduction . . . 93
6.1.1 Overview . . . 93
6.1.2 Related Work . . . 94
6.1.3 Open Research Statement . . . 96
6.2 Experiment 1 . . . 96
6.2.1 Introduction . . . 96
6.2.2 Method . . . 96
6.2.3 Analysis . . . 100
6.2.4 Discussion . . . 102
6.3 Experiment 2 . . . 102
6.3.1 Introduction . . . 102
6.3.2 Method . . . 103
6.3.3 Analysis . . . 104
6.3.4 Discussion . . . 111

6.4 General Discussion . . . 111
6.4.1 Relationship to Prior Work . . . 112
6.4.2 Implications . . . 114
6.4.3 Limitations and Future Work . . . 114
6.4.4 Conclusion . . . 115
6.5 References . . . 115
References . . . 115

Postscript to Chapter 5 **117**
Overview . . . 117
Rationale for Experimental Design . . . 117
Summary of Findings . . . 117
Relationship to the Studies in Chapter 3 and Chapter 5 . . . 118

7 Conclusion **120**
7.1 Research Objectives . . . 120
7.2 Main Findings . . . 120
7.2.1 Data Visualisation Literacy . . . 122
7.3 Reproducibility . . . 122
7.4 Contributions and Implications . . . 122
7.5 Limitations and Future Directions . . . 124
7.6 Closing Remarks . . . 125
7.7 New . . . 125

References **126**

1 Introduction

Data visualisations help people make sense of numbers. Whilst a list of numbers may allude to an upwards trend, presenting those numbers *as* visual phenomena can aid interpretation. A data visualisation systematically depicts the precise numerical values, facilitating mental processing of this information.

As *external* representations of data, visualisations reduce perceptual and cognitive burdens in interpretation (Scaife and Rogers, 1996). By imparting efficiency and clarity, data visualisations support pattern-recognition and reasoning. However, a single dataset can be depicted in numerous ways, and different designs can vary widely in their effectiveness (Franconeri et al., 2021). Thus, data visualisation's strength can also be its vulnerability. Outsourcing cognitive processes to a graphical depiction leaves a viewer at the mercy of the chosen method of visual representation. Thus, understanding successful design is crucial.

The effectiveness of data visualisations can be defined in many ways, encompassing their various objectives, which include informing, persuading, engaging and promoting retention (Bertini et al., 2020). However, in general, successful data visualisations will convey pertinent information in a visually- and cognitively-comprehensible manner (Macklinay, 1986; van Wijk, 2005). Failing to meet these criteria risks misleading viewers, which is antithetical to the purpose of data visualisation. Therefore, knowledge of human factors in visualisation is vital for ensuring charts, graphs, and maps achieve their potential.

1.1 Research Motivation and Objective

Interpreting data in any medium does not involve simply observing numerical values, but rather identifying patterns and making *inferences* about the data. For example, one may notice that values are decreasing rapidly, or that there is substantial variability in the dataset, or that some data is missing. One may also make inferences about how large or small values are. This is an important aspect of understanding data, because the same numerical value can be considered large or small depending on its context.

The BBC radio programme *More or Less*, which examines statistics reported in the news and elsewhere, often addresses this issue. Some figures may instinctively seem large, others small. However, asking the question 'Is it a big number?' considers whether this initial impression is appropriate. For example, a country's multi-trillion dollar national debt may seem large, but may be similar to that of other countries when accounting for its high gross domenstic

product. Similarly, understanding the context of the Richter scale is required to determine that a value less than one does indeed reflect a small earthquake magnitude. Gauging magnitude is important for understanding numbers in a wide variety of situations.

As data visualisations are used to convey numerical information, it is important to understand how viewers interpret the magnitude of presented values. Studying the cognitive processing of magnitude reveal how inferences are generated and provide insight into the effects of design choices. This, in turn, can inform recommendations for designers who may wish to represent the magnitude of values using graphical cues. Yet, this has been an underexplored topic in visualisation, with insufficient empirical research exploring this aspect of interpreting visualisations. The aim of this thesis is to generate robust empirical evidence on the interpretation of magnitude in data visualisations, using controlled experiments.

1.2 Overview of the Thesis

Research questions In this chapter, I review related research to provide context for the empirical work conducted in this project, before outlining the structure of the thesis itself.

1.3 Contributions

Through a series of empirical experiments, I demonstrate that data visualisation design choices can affect mental representation of numbers' magnitudes. These large-sample, controlled experiments strengthen and expand the evidence base on this overlooked aspect of data visualisation. This work reveals that, in various types of visualisations, graphical cues to context play a role in the processing of how large or small values are. Specifically, judgements about magnitude are informed by the relative positions of values within axis limits. Focusing on underlying cognitive mechanisms contributes findings which are applicable to a variety of visualisation formats. These findings also contribute recommendations for designers, which involve considering suitable axis ranges in order to convey magnitude appropriately. This guidance challenges a convention in data visualisation design and advocates against the use of particular default settings, where appropriate.

2 Literature Review

Throughout history, data visualisations have provided insights on the dominant topics of the day, from science and healthcare to civil rights and warfare. Identifying the first use of data visualistion is impossible, but it is clear that humans have used graphic forms to display numerical information for millennia. For example, on a clay tablet dating from 3100-3000 BC, circles and semicircles represent the quantities of the beer rations which were used to pay workers (MacGregor, 2010). Other early visualisations include geographical maps and astronomical diagrams plotting the movements of the planets. The 18th Century saw the development of many common formats used today, such as bar charts, line charts, and pie charts, all of which are typically credited to William Playfair (Friendly, 2006). However, the late 19th Century has been described as 'The Golden Age of Statistical Graphics' (Friendly, 2006, pg. 13), generating innovations in the representation of large datasets.

In 1855, John Snow produced a map showing the spatial distribution of cholera deaths in an area of London by displaying a mark at the location where each victim had lived. Deaths clustered near a contaminated water pump substantiated his radical claim that infected water sources spread this disease (Friendly, 2006). This illustrates how data visualisations can be used to demonstrate vitally important patterns and relationships that were previously overlooked. In 1857, Florence Nightingale visualised fatalities in the Crimean war, using a format known as a 'coxcomb', or 'rose diagram' (Friendly, 2006). Each month's death toll was represented by the size of a segment projecting from the chart's centre point (Speigelhalter, 1999). Crucially, the use of colour to distinguish between different causes of death reveals that unsanitary conditions in hospitals were a far bigger threat to life than the battlefield (Friendly and Andrews, 2021). This data visualisation was distributed widely to politicians, including the Prime Minister, promoting awareness of the magnitude of preventable deaths (Magnello, 2012). In 1861, Charles Joseph Minard plotted Napoleon's Russian invasion and subsequent retreat with an increasingly diminishing army. Part map, part flow diagram, and part line chart, this data visualisation is a paragon of information density, representing six variables in a single graphic whilst telling a coherent story (Tufte, 1983).

Although the above visualisations may appear to reveal major findings for the first time, none of Snow, Nightingale, or Minard used these visualisations to perform their initial analysis. Instead, these visualisations were used for the purposes of persuasion and storytelling (Kosara and Mackinlay, 2013). This is a testament to the effective use of data visualisations as rhetorical devices and instruments for storytelling, rather than their use as analytical tools. Furthermore, historically significant data visualisations have not always achieved the recognition and response they sought at the time. W.E.B. Du Bois' data visualisation exhibit on

the oppression and development of Black Americans won prizes and medals at the 1900 Paris Exposition (Du Bois, 1900), but was generally overlooked by the mainstream American press (Forrest, 2018).

It is necessary to acknowledge that the history of data visualisation is rather sparse, and to recognise *contemporary* work in this discipline (Kosara, 2016). Recent innovations in software have generated visualisations with interactive or dynamic elements (Friendly, 2006), but straightforward static visualisations have not disappeared. Indeed, one particularly successful case is the powerfully simple 'warming stripes' visualisation (Hawkins, 2018). This design uses coloured stripes to display average global temperature from 1850 to the present, highlighting the rapid increase in recent years using increasingly darker reds. By eschewing date labels, text, and a colour legend, only the fundamental message remains. Accordingly, this visualisation has been reproduced in various unlikely settings for a data visualisation (e.g., music festivals, clothing; Kinitish, 2019), earning a reputation as a recognisable symbol of the climate emergency.

When considering famous data visualisations, both historical and contemporary, it is important to avoid making unfounded conclusions about how particular design choices may have contributed to their success. The effectiveness of these designs is undeniable, on account of their documented influence. However, whilst these examples illustrate that visualisations *can* be extremely effective, case studies alone do not provide insight into *why* they are effective. The history of data visualisation reveals the power of visualisations in communication, rather than the principles of good design, and speculation about potential positive attributes is not a reliable source of knowledge. This illustrates the importance of studying data visualisations from a scientific perspective.

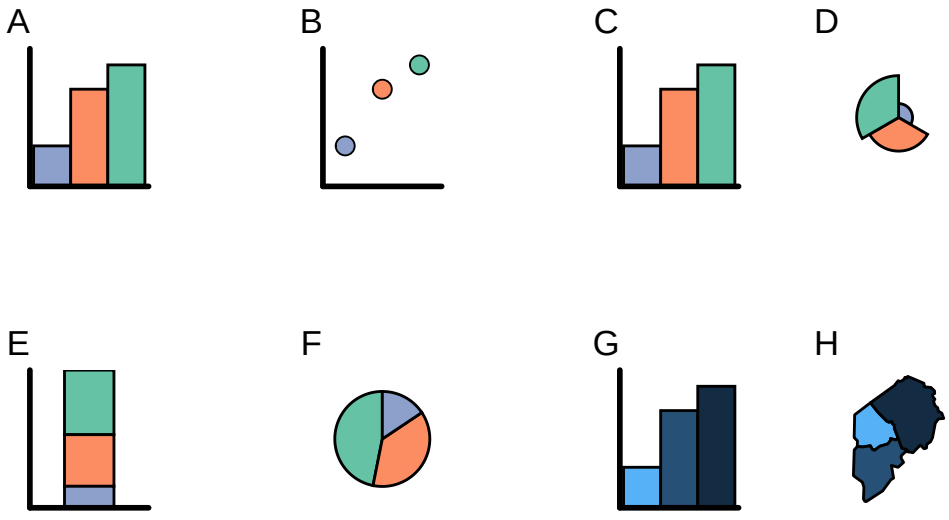
There is no guarantee that a well-received visualisation unanimously employs effective practices. Hans Rosling's presentations (e.g., Rosling, 2006), which used animated visualisations of global health data, illustrate this. Research conducted subsequently reveals variation in the value of his different techniques. These hugely popular presentations included verbal explanations of complex animated graphics, delivered with enthusiasm and a dynamic stage presence. Empirical research using Rosling's talks as stimuli demonstrates that his narration facilitates comprehension of data visualisations (Obie et al., 2019). However, the same study found that it has no effect on memory and can elicit concerns about trustworthiness. Another study found that static visualisations of this dataset improved understanding, but animated visualisations were more popular (Robertson et al., 2008). Furthermore, these designs also employed variable dot sizes in scatterplots, which can lead to perceptual biases (Anderson et al., 2021; Hong et al., 2021). With so many variables involved in these talks, more research is required to understand the components of effective storytelling with data visualisations (Kosara and Mackinlay, 2013). Rosling's contributions, in particular his pioneering use of narrative visualisation and his concern for intelligibility, should not be overlooked. Despite this, insight into the effectiveness of specific visualisation practices is best acquired through systematic study.

2.1 Data Visualisation Formats and the Grammar of Graphics

There is no *one* way to represent a dataset visually. Developing a data visualisation involves making a large number of design choices, which can culminate in vastly different results. Chart 'types' (e.g., bar chart, pie chart, line chart) offer an easy way to categorise the broad format of a visualisation. However, these categorisations do not reflect the way that data visualisations are constructed or how they function (Wilkinson, 2005). The 'Grammar of Graphics' (Wilkinson, 1999) offers an alternative approach.

The Grammar of Graphics is a system for formally defining visualisations in terms of their underlying structure. As a *grammar*, rather than a taxonomy, it was developed in order to express the *composition* of any data visualisation through six components (Wilkinson, 2005). *Elements* describe both the *aesthetic attributes* which visually encode values (e.g., position, size, hue, transparency), and the *geometries* which represent those values (e.g., bar, dot, line). *Coordinate systems* describe the canvas used for representing values. For example, Cartestian coordinates use the recognisable vertical and horizontal dimensions associated with bar charts, polar coordinates use the circular mapping associated with pie charts, map projections use a cartographic mapping associated with world maps. The other components of the Grammar of Graphics are the *data* used, *variable transformations* (e.g., mean, sum, rank), *scale transformations* (e.g., linear scale, logarithmic scale) and guides (e.g., axes, colour legends).

This system allows for efficient and consistent characterisation of different visualisation formats. For example, Figure X shows that bar charts (A) and dot plots (B) both use the same aesthetic attribute (position) to encode values, but differ in their geometry (bar versus dot). A regular bar chart (C) with polar coordinates is equivalent to a chart like Florence Nightingale's coxcomb (D). Conversely, a *stacked* bar chart (E) with polar coordinates is equivalent to a regular pie chart (F). Charts can employ more than one aesthetic attribute, for example, the above charts use hue to represent the different categorical values. However, it is possible to use lightness to represent different numerical values instead, with darker colours representing higher values (G). Using this aesthetic attribute, in combination with a map projection and geometries based on the shape of geographic regions, produces a choropleth map (H). This illustrates how the components can be combined in a flexible and modular manner, with many more possible visualisations of this dataset. The Grammar of Graphics has been influential in the development of a number of data visualisation design tools, including Polaris (Stolte et al., 2002), which became Tableau, ggplot2 in R (Wickham, 2010), D3 (Bostock et al., 2011), and Vega-Lite (Satyanarayan et al., 2017).



2.2 Data Visualisation Software and the Influence of Default Settings

Considering the *process* by which data visualisations are created is crucial for understanding this subject. Modern software has made it possible to quickly and easily produce a wide range of visualisations. However, variation across visualisation design tools affects the range visualisation formats available to users and the degree of customisation offered. For example, programming libraries, where data visualisations are created by writing lines of code (e.g., ggplot2 in R, plotly in Python) typically offer more options and greater control than simple point-and-click software (e.g., Microsoft Excel). Many visualisation design tools also provide specialised capabilites. For example, Tableau is often used for business intelligence applications, such as building dashboards (Elias et al., 2013), D3 was developed for designing visualisations for the web (Bostock et al., 2021), VegaLite was developed for generating interactive visualisations (Satyanarayan et al., 2017), and ggplot2 was developed for use within a data analysis workflow (Wickham, 2011).

What is possible and practicable using a particular piece of software will influence a designer's choices, which may in turn affect viewers' interpretations. However, visualisation software also initially imposes particular properties on aspects of a visualisation, through its default settings. For example, a pie chart, prior to customisation, will present segments in a particular order, using a particular set of colours. Even when it is *possible* to reject default designs, they can

be highly influential, because they will remain unchanged when it is unclear *how* or *why* they should be altered (Shah and Kesan, 2006).

However, existing default settings in data visualisation are not always suitable. For example, by default, software for creating line charts typically employs y-axes which are constrained to the range of the data. Therefore, the highest and lowest values presented at the chart's extremes, impeding viewers' ability to gauge the magnitude of the difference. When representing categorical values with colour, designers can improve viewers' performance by rejecting defaults in favour of colours which correspond semantically with plotted data (e.g., yellow for banana; Lin et al., 2013). When representing continuous values with colour, rainbow colour scales are a popular choice (Ware et al., 2023). However, several tools have changed their default colour scales to encourage designers to avoid the issues in perception associated with this palette (Reda and Szafir, 2021).

Default settings can also cause issues in the generation of visualisations for exploratory analysis (Correll et al., 2019). These visualisations are used to understand the characteristics of a dataset prior to formal statistical analysis. For example, histograms are used to visualise the shape of a univariate distribution. The algorithm used to produce histograms in R and D3 assumes by default that the data are normally-distributed. Consequently, abnormalities in non-normal data are 'smoothed-over', preventing viewers from identifying them. Dot plots, another format for visualising distributions, display each individual value using a dot. In R and Tableau, by default, these dots have no translucency. This can result in many overlapping dots in close proximity, impeding a viewer's ability to differentiate between areas with different densities. This study demonstrates that default settings are not always inappropriate, but when they are *agnostic* towards characteristics of plotted data, unquestioning use of default settings can conceal relevant aspects of a dataset.

Although the above research exposes issues with some default settings, they are certainly not exclusively harmful. For example, one default setting used by Microsoft Excel is 'redundant encoding', where individual data points are represented using different shapes *and* different colours (e.g., blue diamonds and green triangles). Experimental work has observed that whilst this technique does not confer benefits in some tasks (Gleicher et al. 2013), it improves viewers' performance in other tasks (Nothelfer et al., 2017). Empirical research is important in order to identify how default settings may be beneficial or detrimental. Indeed, researchers in data visualisation often suggest that their findings may inform the development of default settings (e.g., Heer and Bostock, 2010; Xiong et al., 2021, Kerns and Wilmer, 2021). Several experiments in this thesis were designed to explore the consequences of default settings.

2.3 Popular Guidance on Effective Data Visualisation Design

Our understanding of how people interpret data visualisations (and subsequent guidance) is built on shaky foundations. Some received wisdom has not been empirically tested at all, other claims have been discredited or confirmed only recently (Kosara, 2016). Consequently,

it is not always clear where evidence ends and opinion starts; intuition and unsubstantiated statements make for "visualisation folklore" (Correll, 2022, pg. 3).

Statistician Edward Tufte is a source of widely-cited advice on the design of data visualisations, which he articulates in popular books such as *The Visual Display of Quantitative Information* (Tufte, 1986). One famous contribution is the 'lie-factor', which attempts to quantify the degree of misrepresentation in charts that distort data. For example, plotting values using two dimensional images exaggerates differences between values, because perceived size is determined by an image's entire *area*, not just its *height*. Consequently, in a visualisation that appears to show a decrease of 42 percentage points, dividing by real numerical decrease of 15 percentage points, generates a lie-factor of 2.8, compared to an ideal score of 1. However, Tufte's criteria proposed for diagnosing *substantial* distortion (less than 0.95 or more than 1.05) are based on speculation, rather than scientific evidence (Beattie and Jones, 2002). In similarly arbitrary guidance, Tufte suggests that a dataset of 20 or fewer observations should be presented in a table, rather than a data visualisation. However, a subsequent empirical experiment revealed that pie charts elicited more accurate responses than tables for proportion judgements involving only three observations (Spence and Lewandowsky, 1991).

Tufte also advocated for minimalism in the design of data visualisations. His recommendation to maximise the 'data-ink ratio' involves maximising the proportion of ink (i.e., pixels) used to depict the data itself and minimising inessential elements (Tufte, 1986). However, this notion is vague and prone to excessive simplicity. Redundant features can serve to minimise error (Tversky, 1997), with 'redundant' tick marks on axes required for accurately extracting numerical values (Kosslyn, 1985). The qualifier "within reason" (Tufte, 1986, pg. 96) is an imprecise addition to this guidance, but empirical research can identify where extreme sparseness unnecessarily biases interpretations (Stock and Behrens, 1991, Gillan and Richman, 1994).

Consistent with his minimalistic approach, Tufte's recommendation to eliminate 'chartjunk' involves avoiding the use of distracting visual embellishments, which range from excessive gridlines to artistic decoration (Tufte, 1986). However, there is mixed evidence regarding the harm caused by chartjunk (Franconeri et al. 2021). However, condemning 'chartjunk' remains popular, not just on aesthetic grounds, but also due to the rhetorical qualities of minimalist designs, which imply a straightforward, unbiased presentation of data (Kosara, 2016; Kennedy et al., 2016).

Researchers argue that Tufte's recommendations for minimalistic designs do not account for human cognitive processing (Wilkinson, 2005; Chabris and Kosslyn, 2005). Furthermore, he has been criticised for failing to support his claims with empirical evidence (Feldman-Stewart et al., 2000). Instead, his guidance is underpinned by a large collection of example visualisations taken from various sources. Therefore, Tufte's principles might assist in describing common features of some successful visualisations, rather than serving as definitive rules (Kindlmann and Scheidegger, 2014). Rigorous data visualisation research is required to fill gaps in knowledge and generate a reliable evidence-base.

2.4 Rigorous Data Visualisation Research

Visualisation research takes many forms. Studies on data visualisation have employed a range of techniques, including controlled experiments, usability tests, interviews, observations, and case studies, and have focused variously on perception, cognition, exploratory data analysis, and user experience (Lam et al., 2011). Experimental psychology studies on data visualisation are particularly valuable because they generate fundamental evidence on *how* visualisations are interpreted. Considering human interpretation in visualisation research is crucial for generating generalisable knowledge. Inadequate best practice recommendations indicate insufficient understanding of psychological mechanisms. However, progress can be slow, since theories about cognitive and perceptual processes are built through cumulative work (Chen et al., 2020). Psychological research confers benefits in the form of related empirical work, alongside established methods and theories (Correll, 2022; Rensink, 2021).

Multiple studies illustrate that preferences and introspection are not a reliable source of information on effective visualisation practices. For example, an experiment exploring physicians' judgements about clinical trials found that icon arrays resulted in the most accurate judgements, compared to tables, pie charts, and bar charts (Elting et al., 1999). However, none of the 34 physicians in the sample preferred this format. In another study, medical students almost unanimously preferred visualisations with a rainbow colour scheme, but made fewer errors when using a diverging (e.g., red-blue) colour scheme (Borkin et al., 2011). Tables of values may be favoured over visualisations in certain tasks where the visualisations actually offer significant benefits (Saket et al., 2019). Similarly, participants in Burns et al.'s (2021) study estimated that pictographs took longer to understand, compared to equivalent visualisations without icons. However, this self-report measure was at odds with recorded response times, which indicated no differences between visualisations types. There is also evidence that graduate students preferred certain statistical map designs over others despite conferring no performance advantage (Mendonça and Delazari, 2014). Many authors suggest that preferences are influenced by familiarity, rather than performance advantages. Measuring preferences provides valuable insight into people's engagement with different visualisations. However, such opinions must be treated appropriately, not used to inform conclusions about effectiveness.

Rensink (2021) presents recommendations for generating useful research findings. Using a single task, and manipulating a single feature of interest, over multiple trials, assists in identifying underlying mechanisms. Integrating explanations from prior research helps ensure explanations of mental processes are sufficiently detailed. Other important but frequently overlooked matters include appropriate counterbalancing, reporting effect sizes and acknowledging individual differences.

There are a multitude of variables that can be manipulated to gain insight into visualisations. Criticisms are sometimes levelled at studies with particularly high or low levels of experimental control. However, researchers must strike an appropriate balance between ecological validity

and precision (Abdul-Rahman et al., 2020). Choosing suitable tasks for participants requires a similar trade-off (Suh et al., 2022).

Vision sciences offer a variety of paradigms for assessing various aspects of human performance in visualisation tasks. For example, experiments may evaluate accuracy (by comparing responses to a correct answer), precision (by quantifying variability in responses), or processing speed (by measuring reaction times, Elliott et al., 2020). However, chosen methods must be appropriate for a research question. Whereas methods from vision-sciences are typically concerned with performance in low-level perceptual tasks, other research focuses on decision-making (Padilla et al., 2018) or *message*-level interpretations (Pandey et al., 2015). The latter concerns broad assessments of data, such as whether a difference is large or small, rather than the ability to extract specific values. This is also referred to as *gist* (Reyna and Brainerd, 1991).

2.5 Perceptual Precision in Data Visualisations

Identifying gaps in our understanding of the psychology of data visualisations requires knowledge of prior lines of inquiry and established findings. Arguably the most influential study in the field of data visualisation is Cleveland and McGill's (1984) investigation of elementary perceptual processes involved in viewing visualisations. This study sought to establish how *precisely* viewers can represent different graphical properties used to encode data (e.g., position, length, angle, etc.). For each encoding type, participants identified which of two marks conveyed the smaller value, and estimated the difference in size as a percentage. Subsequent ranking based on the magnitude of participants' errors produced a hierarchy of visual encoding channels. Since position-encoding produced smaller errors than both length- and angle-encoding, this suggests that data will be represented most precisely when encoded using position on a common (aligned) scale.

This study's findings have endured replication (Heer and Bostock, 2010) and enthusiasm for perceptual precision has inspired a great deal of important research in this field. This research spans visual processing of proportion (Spence and Lewandovsky, 1991; Hollands and Spence, 1998), variance (Stock and Behrens, 1991), correlation (Harrison et al., 2014; Hong et al., 2021), and other basic processes, such as visual comparison (Simkin and Hastie, 1987; Zacks et al. 1998) and colour discrimination (Szafir, 2018). The study has also influenced development of software for automating visualisation design (Mackinlay, 1986) and simulating visualisation comprehension (Lohse, 1993). However, to consider perceptual precision as the *only* relevant concern in data visualisation design is unwarranted; many additional factors require consideration.

2.6 Beyond Perceptual Precision

Optimally-precise visual cues are not always employed when viewing visualisations. Viewers are sensitive to other task-irrelevant visual cues, which can lead to inaccurate judgements about plotted data (Yuan et al., 2021). Furthermore, in particular tasks, precision can actually hinder, rather than facilitate, judgements. For example, because perceptual averaging benefits from lower spatial frequencies, colour encoding offers greater efficiency than more precise position encoding in line charts (Correll et al., 2012, see Figure 2.1). Effective decision-making under uncertainty does not necessarily correspond to precision in probability estimation, because of the differences in mental processing associated with these two distinct tasks (Kale et al., 2020).

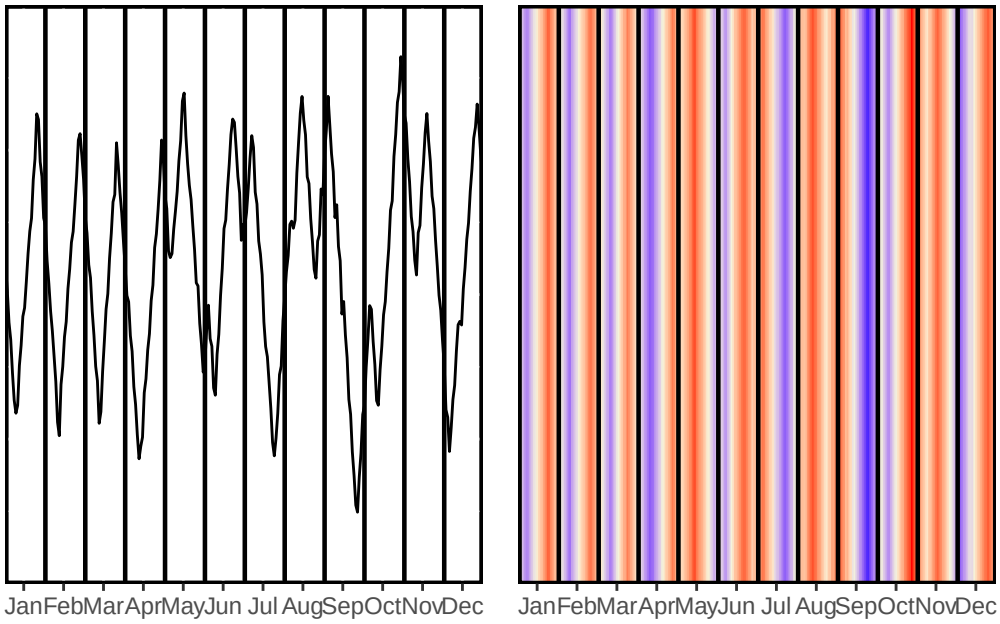


Figure 2.1: noise example similar to Correll et al., 2012

Furthermore, the choice of graphical encodings employed in a data visualisation can influence the *type* of interpretation it elicits. For example, viewers are more likely to refer to trends when describing line graphs and discrete differences when describing bar charts (see Figure 2.2). **This can occur even when the nature of the plotted data is ill-suited to this type of characterisation (Zacks and Tversky, 1999). This means that a line chart may provoke a peculiar interpretation such as 'a building becomes more secure as the alarm system becomes more active', whereas a bar chart may provoke an interpretation such as 'a building with 10 motion sensors is more secure than a building with 5 motion sensors'.** Similarly, *production* of bar charts and line charts is also influenced

by whether a discrete or continuous relationship is specified in the brief. Design choices also influence beliefs about the distribution of underlying data, when presenting average values. Compared to a data point positioned 'outside' a bar, a data point positioned 'inside' a bar is more likely to be considered part of the underlying data (Newman and Scholl, 2012). However, confidence intervals eliminate this bias (Pentoney and Berger, 2016). This accords with the notion that viewers' cognitive associations between visual features and abstract characteristics of data are important in data visualisation design. Through common metaphors (e.g., hierarchy and vertical position), aspects of a design may offer *affordances*, carrying connotations which encourage particular interpretations (Xiong et al., 2022, Ziemkiewicz and Kosara, 2008, Kindlmann and Scheidegger, 2014).

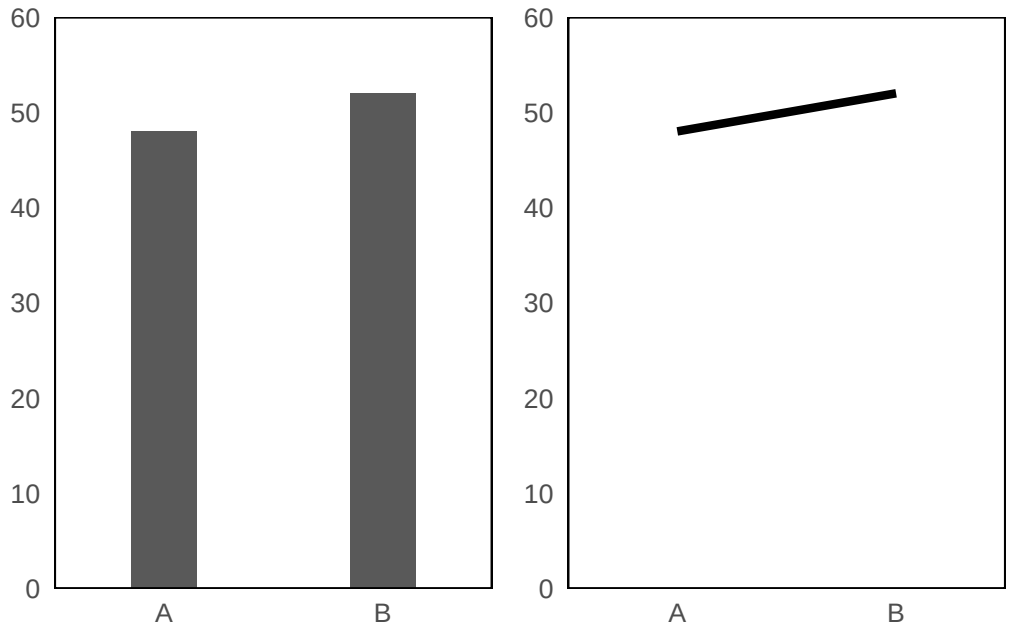


Figure 2.2: Zacks and Tversky

Attention is another important factor in comprehension of data visualisations. Complex tasks requiring selective attention can cause distinctive patterns in non-focal data to be completely overlooked (Boger et al., 2021). Features of data mentioned in textual summaries are overweighted in viewers' mental representations, causing difficulty with the ability to assume the perspective of a naïve viewer (Xiong et al., 2019). In addition, the salience of vertical bars may be responsible for incorrect reports of differences between histograms with identical distributions (Lem et al., 2014). As a solution, explicitly encoding differences between pairs of values can facilitate pattern recognition (Nothelfer and Franconeri, 2020) and highlighting particular attributes can facilitate recall (Ajani et al., 2021).

Simply conveying information is not the only purpose of data visualisations, since they also

influence recall, opinion-formation, and decision-making (Bertini et al., 2020). As illustrated above, a large number of cognitive biases affect these aspects of the mental processing of data, as well as several others, including causal reasoning and assessment of hypotheses (Dimara et al., 2020). Whilst it is necessary to consider the precision of elementary perceptual processes, that alone is not sufficient for a comprehensive understanding of how data visualisations function (Bertini et al., 2020).

2.7 Manipulating Axes in Data Visualisations

Understanding how inaccurate impressions arise provides insight into mechanisms involved in interpreting data visualisations. This, in turn, can inform recommendations for effective design. A prominent topic in the literature on misleading visualisations is axis truncation. This typically refers to the practice of employing a y-axis which commences with a non-zero value (Correll et al., 2020), though may also be considered any adjustment at either extreme of an x- or y-axis (Pandey, 2015). Figure 2.3 shows examples.

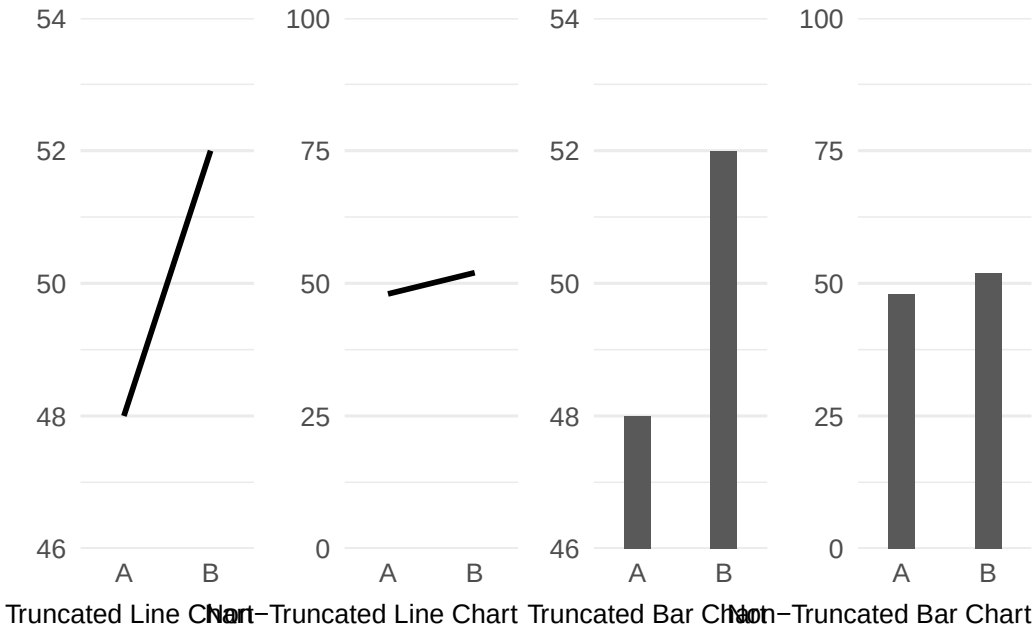


Figure 2.3: figure caption

There is considerable evidence that the range of axis values employed in charts influences interpretations of data. The majority of research on this topic has focused on how constraining the range of an axis, and thus increasing the physical distance between plotted values, increases

the perceived magnitude of the difference between those values. For example, accountants appraising financial performance using line and bar charts interpreted plotted increases as larger when they were depicted using a truncated y-axis (Taylor and Anderson, 1986). Similarly, bar charts employing truncated axes biased students' investment decisions (Arunachalam and Pei, 2002). Students were more likely to select a less-successful company when a truncated chart exaggerated that company's growth rate, compared to when a non-truncated chart was used. An online experiment also observed that differences between values were considered larger when truncated bar charts were used (Pandey et al., 2015). This experiment examined message-level representations of data by framing questions in terms of subject matter (e.g., access to safe drinking water) rather than graphical elements (e.g., difference in bar length. Other axis manipulations, such as log-scales (Romano et al., 2020), inverted scales (Woodin et al., 2021, Pandey et al., 2015), and expanded axes in scatterplots (Cleveland, 1982) also influence judgements about data.

Risk communication research has independently generated similar findings. Because many hazards cannot be completely avoided, data visualisations are often used to contrast the levels of risk associated with two scenarios (e.g., intervention versus no intervention). Thus, assessments of 'risk reduction' are essentially judgements about the magnitude of difference between two values. For example, one experiment compared stacked bar charts, which include additional information on the total number of individuals at risk, to bar charts which displayed only the number of individuals *affected* (Stone et al., 2003). The latter design increased the bars' visual disparity, and subsequently increased impressions of the magnitude of difference.

The physical distance between data points consistently biases interpretations of the magnitude of difference in spite of attention to actual numerical values and also design features intended to highlight truncation (Correll et al., 2020). Bias is diminished, but still observed, following explicit warnings about errors in judgement due to y-axis truncation. This suggests that this effect is largely automatic, and does not primarily occur due to insufficient engagement of cognitive capabilities (Yang et al., 2021).

Researchers have also explored individual differences in interpretations of data presented using truncated axes. One study observed no association between participants' susceptibility to bias due to axis truncation in bar charts and their data visualisation literacy (Yang et al., 2021). Conversely, another experiment suggests that the effect of axis truncation on subjective judgements and quantitative estimates in line charts disappears when accounting for data visualisation literacy (Driessen et al., 2022). However, low variability in observed data visualisation literacy levels in the latter experiment raised concerns about the scale used to measure data visualisation literacy.

Pandey et al. (2015) and Yang et al (2021) propose that this bias could arise due to the dominance of first impressions during translation from graphical schemata (Pinker, 1990) to a 'real-world' conceptual understanding (see also, Carpenter and Shah, 1998, Tversky and Kahneman, 1974). Additionally, Yang et al. (2021) suggest that viewers' beliefs about the communicative intent of a designer could play a role in viewers' interpretations. Under Grice's *Co-operative Principle* (Grice, 1975), communicative contributions in conversation are assumed

to be truthful, relevant, clear, and sufficiently informative. To extrapolate this to data visualisations, viewers might infer that differences between values must be genuinely large if they appear large, because they would otherwise not be presented as such.

In *How to Lie With Statistics*, Huff (1954) suggests that axis truncation creates a false impression of plotted data. This practice has been labelled 'deceptive' for both bar and line charts (Lauer and O'Brien, 2020). A tool for automatically identifying and correcting misleading line charts extends y-axes to include zero whenever this value is omitted from the original chart (Fan et al., 2022).

Recent work has presented an alternative perspective on this controversial practice. Non-truncated axes can obscure significant differences just as easily as truncated axes can exaggerate inconsequential differences. The appropriate magnitude to convey depends on what constitutes an important difference in the data at hand (Correll et al., 2020). Indeed, *failing* to truncate an axis could be considered misleading in certain circumstances (Wainer, 1984). Yang et al. (2021) suggest that effective designs will ensure that a viewer's immediate characterisation of plotted data closely corresponds to their interpretation following a detailed inspection. Acknowledging that differences must be depicted in proportion to their significance, Witt (2019) reports that axes spanning approximately 1.5 standard deviations provide a balance between sensitivity and bias in fields with standardised effect size measures, such as psychology. Unfortunately, different domains will not necessarily share the same notion of what amounts to a meaningful difference. Choices regarding axis ranges are ultimately designers' unavoidable decisions (Correll et al., 2020).

Although line charts and bar charts are equally susceptible to biases due to truncation (Correll et al., 2020; Witt et al., 2019), there may be reason to treat them differently. Truncation distorts the mapping between a bar's extent and the quantity it represents, but free-floating position-encoding used in line charts does not convey quantity in the same manner, providing immunity against such distortion (Bergstrom and West, 2017). Therefore, whilst starting an axis at zero cannot guarantee that differences between values are depicted appropriately, this does ensure adherence to a fundamental aspect of visualisation design. Alternatively, to avoid this trade-off, quantitative data with discrete categories can be plotted using position-encodings only (e.g., dot plots).

2.8 Misleading Data Visualisations

Some misleading visualisations may prevent viewers from accurately extracting numerical information. However, research on axis truncation illustrates that misleading visualisations may also interfere with subjective judgements. A line chart may avoid misrepresenting a dataset's numerical properties yet generate a distorted impression of the magnitude of a trend. The latter is revealed not by assessing the *performance* of viewers, but their *interpretations* (Stone et al., 2015).

Influencing subjective judgements may still be considered a *misleading* practice because a dishonest framing of information could elicit an unreliable interpretation that would differ from the same viewer's better-informed perspective. Not all aspects of deceptive design are *inherently* misleading, and deceptiveness can be context-dependent. Comparing examples of 'misleaders' from Ge et al.'s (2023) design space helps illustrate this distinction. 'Concealed uncertainty' and 'cherry-picking' refer to unambiguously deceptive practices, whereas 'aggregation' and 'scale range' must be preceded by the word *inappropriate* in order to convey their capacity to deceive.

2.9 Data Visualisation Literacy

Understanding individual differences in the ability to comprehend data in visualisations is important for understanding the psychology of data visualisations (Boy et al., 2014). Research on this topic requires reliable tools for measuring data visualisation literacy.

Galesic and Garcia-Retamero's 13-item test (2011) was based on Friel et al.'s (2001) hierarchy of skills for interpreting visualisations, which ranges from comprehension to extrapolation. Research has demonstrated that this scale can predict whether a graphical representation will facilitate understanding of risk information (Okan et al., 2012). A different 53-item test employs a wide range of data visualisation formats, and higher scores are positively associated with both numeracy and need for cognition (Lee et al., 2019).

Research on data visualisation literacy has tended to focus on interpretation of well-designed charts (Ge et al., 2023). However, the ability to detect (Camba et al., 2022) and make sense of (Ge et al., 2023) misleading charts should be considered an important feature of data visualisation literacy. A robust 30-item test enables assessment of an individual's ability to accurately comprehend deceptive designs (Ge et al., 2023). This work also suggests that attention and critical thinking may benefit viewers in avoiding some, but not all, biased interpretations. Using Galesic and Garcia-Retamero's 13-item test (2011), Okan et al. (2016) found that higher data visualisation literacy is associated with more time processing a visualisation's misleading features, thus promoting correct interpretations. Lower data visualisation literacy is associated with greater reliance on conventions (e.g., the relationship between vertical position and magnitude).

The empirical work presented in this thesis employs the 5-item version of Garcia-Retamero et al.'s (2016) Subjective Graph Literacy scale. Users are asked to rate their competence in working with bar charts, line charts, and pie charts, and also their ability to perform simple tasks using bar charts. This approach echoes prior work in the development of subjective numeracy scales. Despite its short completion time and use of subjective ratings, it is strongly correlated with an objective measure of data visualisation literacy (Galesic and Garcia-Retamero, 2011). The scale also produces a final score out of 30, offering greater sensitivity than a similarly brief objective scale, where tallying correct responses produces a final score out of 4. These

characteristics make for an appropriate tool for assessing participants' data visualisation literacy in experimental studies. Indeed, this measure has been used to assess variability between participants in studies on axis truncation (Yang et al., 2021), correlation (Strain et al., 2023), information synthesis (Mantri et al., 2022), and explanation of visualisations (Yang et al., 2023).

2.10 Interpreting Absolute Magnitude

Data visualisation design has the potential to impact subjective judgements of many aspects of data, such as variability, noise, and numerosity. Prior research has closely examined how axis truncation can influence judgements of *relative* magnitude (differences between values). In contrast, little is known about how axis limits may influence judgements of *absolute* magnitude: how large or small values are. Despite this, these basic judgements can be fundamental in developing a basic interpretation of quantitative data. For example, assessing the probability of rain, the number of patients on a waiting list, the amount of CO2 emitted during a journey, and the level of support for a political candidate, are all judgements of *absolute* magnitude. Limited insight into how magnitude is interpreted in data visualisations impedes understanding of how visualisations may effectively communicate magnitude. Prior research on this topic is summarised below.

In bar charts displaying data on individuals affected by a risk, perceived likelihood decreased when the total population at risk was emphasised using shaded bars, rather than blank space (Stone et al., 2017, see Figure 2.4). In bar charts violating the convention of mapping higher values to higher positions, participants frequently misinterpreted magnitudes (Okan et al., 2012). This was due to difficulty in rejecting first impressions, particularly for participants with low data visualisation literacy. Other work which has combined judgements of values' magnitudes with judgements of relative differences has impeded analysis of the former (Okan et al., 2018). Visualisations that facilitate comprehension of relative differences may fail to effectively communicate the absolute magnitudes of values depicted, illustrating a potential trade-off in design (Reyna et al., 2008).

One study has specifically focused on how axis ranges may inform impressions of absolute magnitude. Sandman et al. (1994) manipulated risk ladders, where individual probabilities are presented on vertical scales incorporating a range of probability values. Changing this range alters the position of a plotted value. Perceived threat (a composite measure made up of perceived likelihood, danger, reported concern and fear) was higher when the risk appeared near to the top of the ladder, compared to near the bottom. This is akin to framing effects described in the psychology literature (Tversky and Kahneman, 1981). However, the position of plotted values did not completely dictate magnitude judgements. A numerically higher risk plotted at the same position near the top of the ladder generated higher ratings. There was also mixed evidence regarding the effects on intentions to spend money mitigating the risk. Confidence in the robustness of these findings is limited by various factors including use of a

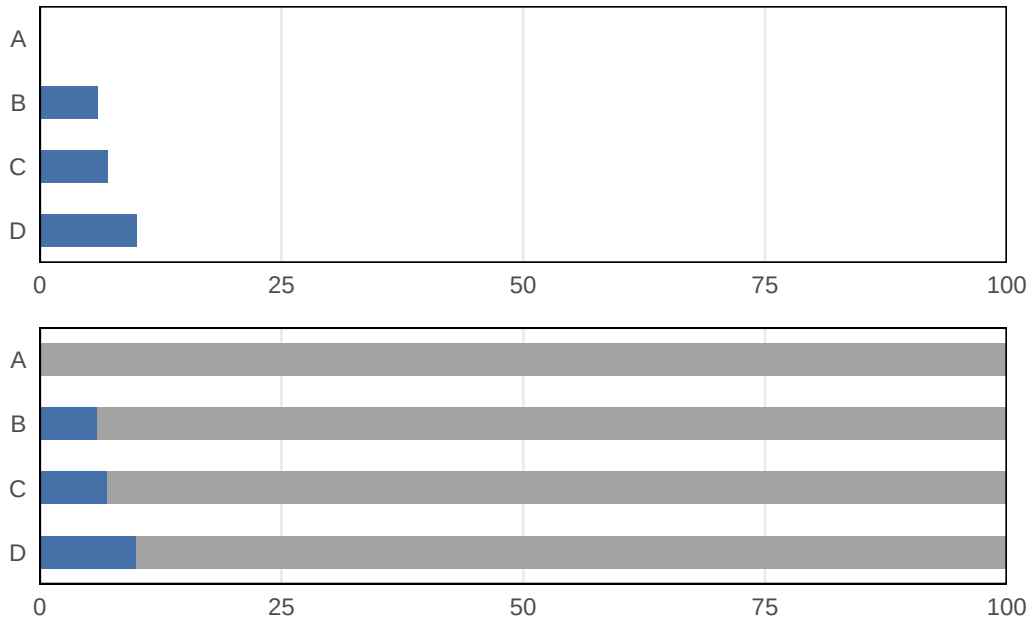


Figure 2.4: Stone et al. 2017 example

single trial per participant, a single scenario, a composite measure obscuring pure magnitude ratings, and a confounding variable of the risk ladder's range.

Comparing linear and logarithmic risk ladders, Freeman et al. (2021) did not replicate Sandman et al.'s (1994) main finding. However, in addition to a graphical cue to magnitude, they used risk ladders which employed additional symbolic number cues in their titles, labels, and accompanying descriptions. A broken scale may also have reduced the degree to which inferences were based on the value's physical position. Therefore, participants' judgements may not have been based purely on the appearance of visualisations.

This thesis presents a detailed investigation into how impressions of the magnitude of numerical values are influenced by data visualisation design.

2.11 Structure of the Thesis

The objective of this thesis is to understand how design choices influence cognitive processing of the magnitude of values presented in data visualisations. I present several empirical studies, each investigating different factors affecting viewers' interpretations and each using a different data visualisation format.

Chapter 2 discusses the methodology and epistemological approach employed in this thesis. In addition to explaining the experimental and statistical techniques chosen, I discuss how research projects can benefit from efforts to increase transparency. Approaches for increasing the reproducibility of published work are presented, in line with recommendations from a variety of disciplines. A particular emphasis is placed on computational reproducibility: the capacity to recreate the computational environment used in generating results. This provides background for the practices used in the following empirical research chapters.

Chapter 3 presents a set of three experiments which establish that interpretations of magnitude can be influenced by data visualisation designs. The first experiment demonstrated that manipulating axis limits in dot plots affected participants' judgements of overall magnitude. Two further experiments investigated whether this occurred because axis limits altered the absolute or relative positions of plotted values within axis limits. In dot plots with inverted y-axes, where higher numerical values are presented at lower positions, values near bottom were associated with higher magnitudes. This illustrates that interpretations of magnitude are informed by the relative positions of values within axis limits.

Chapter 4 presents an experiment which explores how visualisations may influence interpretations of magnitude even when the appearance of plotted values remains unchanged. This experiment demonstrated that manipulating colour legend limits in choropleth maps affected participants' judgements of overall magnitude. Participants rated magnitudes as lower when the range of values on the colour legend extended beyond the largest plotted value. This illustrates that the numerical context accompanying plotted values can influence interpretations of magnitude, without altering the physical appearance of those values.

Chapter 5 presents a set of two experiments which investigate the role of contextual information on interpretations of magnitude. The first experiment demonstrated that participants' judgements of overall magnitude were affected by extension of bar charts' upper axis limits which incorporated a denominator value. A second experiment revealed that participants' bias was increased when this denominator information was excluded from the text accompanying the chart. This illustrates that knowledge about a dataset's characteristics (e.g., denominator value) can influence the extent to which design choices affect interpretations of magnitude.

Finally, Chapter 6 presents a synthesis of this empirical work, alongside a discussion of implications and future directions.

3 Experimental and Computational Methodology

The knowledge generated in a research project is necessarily shaped by the methods of inquiry. A recent survey of visualisation researchers revealed variation in conceptions of how progress is made in the field, with multiple approaches for generating knowledge (Correll et al., 2022). In this chapter, I discuss the epistemological approach which underlies this thesis. This provides a backdrop to the subsequent empirical work and justification for my choices. In addition, by explicitly discussing these decisions, I recognise that they inevitably influence my findings. It is necessary to acknowledge that this methodology is one of many, each carrying their own implications. This reflects the fact that an epistemological approach imposes a particular perspective, unavoidably generating a somewhat narrow view on the topic of interest.

3.1 Experimental Psychology

Selecting a research method involves considering the most suitable type of data for addressing the research question. To understand how data visualisation design choices affect interpretations of absolute magnitude, testing hypotheses using controlled experiments is highly appropriate. This allows for systematic measurement of viewers' judgements and isolates the graphical features of interest from extraneous features. Controlling for the influence of other variables helps establish a causal link between the manipulation and cognition (Barbosa et al., 2021). Experimental methods are well-established in visualisation research for generating robust empirical evidence on the effects of design choices (Abdul-Rahman et al., 2019).

The purpose of psychological studies on data visualisation is to understand viewers' interpretations. However, latent variables cannot be interrogated directly, and must be 'operationalised' to enable analysis. That is, interpretations of the magnitude of numerical values are captured through measurable responses which correspond to underlying mental representations. Thus, experimental methods rely on dependent variables which faithfully reflect actual cognition. Designing experiments also requires compromise between ecological validity (the degree to which the experiment reflects a realistic scenario) and experimental control (the degree to which the researcher dictates aspects of the experiment). In this thesis, I have strived for realism where possible, but have prioritised experimental control in order to ensure the robustness of findings. In visualisation research in particular, it is often necessary to control for differences in participants' knowledge by presenting artificial or abstract data (Lam et al., 2012). Whilst

qualitative studies (e.g., in-depth interviews), may produce richer data than experiments, they do not provide the precision required to systematically evaulate biases in interpretation.

All experiments in this thesis were conducted online using Prolific.co, a website for recruiting research participants. This provided access to a diverse group of participants, which contrasts with the relative homogeniety of a student population. Furthermore, online experiments provide the ability to easily collect data from a large number of participants, which reduces the chance of generating false positives during analysis. In addition to using large participant samples, employing multiple trials per condition helps establish robust effects which are not vulnerable to the particular characteristics a single trial. Similarly, generating generalisable knowledge about mental processing of visualisations often requires multiple experiments. A single experiment is typically not sufficient for an understanding of cognitive factors in interpretation (Chen et al., 2020).

This is a largely positivist approach, concerned with verifiable results which can be generalised beyond the experiment to describe a cognitive mechanism. However, there is also arguably a *postmodern* quality to highly controlled experiments (Mayrhofer et al., 2021). That is, a controlled experiment can be considered a constructed, stimulated setting, with contrived tasks and stimuli that do not precisely reflect the 'reality' under investigation (i.e., spontaneous judgements of authentic data visualisations). Recognising this does not invalidate conclusions from experimental studies, but requires that generalisation of results is treated with caution.

3.2 Analysis Methodology

Large quantative datasets from controlled experiments require appropriate statistical analysis. Judd 2017 - Singmann and kellen (2019) - Meteyard and Davis (2020) - modelling - Barr vs. Bates buildmer as a solution to this debate

Effect sizes to indiciate the size of the effect - Wilkinson and Task Force on Statistical Inference(1999).

3.3 Reproducibility

In recent years, the typical model for conducting and publishing scientific research has been intensely scrutinised. This has prompted serious concern about whether reported findings can be trusted. For example, Ioannidis (2005) estimated that published research may consist of more falsehoods than true assertions. Researchers also report that in Psychology, many studies are not equipped to generate reliable results (Fraley and Vazire, 2014) and the literature is afflicted with a high rate of false-positive findings (Simmons et al., 2011). A large-scale project performing replications of psychology experiments revealed that the evidence for many established conclusions was not as strong as initially reported (Open Science Collaboration,

2015). A survey of over 1500 researchers found widespread perception that science was facing a 'crisis' (Baker, 2016). However, this recognition also has provoked concerted efforts to address these problems in research, through the Open Science movement (Cruwell et al., 2019).

Recommendations for improving scientific research focus on different aspects of the research lifecycle. Improving how studies are conducted, reported, and evaluated requires targeted solutions. For example, rigorous methods and statistical analysis, facilitate researchers in generating valid conclusions. Other practices, such as openly sharing data and code, increase transparency, providing crucial insight into how these conclusions were generated (Munafo et al., 2017). Peng (2011) suggests that the ultimate test of scientific claims is *replication*. This involves independently repeating an entire empirical investigation, thus generating new data to assess consistency with an existing finding. However, this is resource-intensive. A different, albeit less rigorous, approach to evaluating scientific claims involves using a project's original data and code to validate reported findings. If this is possible, the work is *reproducible*. By reusing existing resources, this is simpler than conducting a replication study, yet still facilitates assessment of whether reported results are reliable. *However, if researchers do not make relevant resources available, this undertaking is impossible. Research that cannot be evaluated in this way is not reproducible.*

The empirical work presented in this thesis has been conducted with a focus on ensuring reproducibility. This chapter will review published work on best practices for sharing code, data, and computational environments, and outline the approach to reproducibility employed in this thesis.

3.4 Sharing Code and Data

There are many convincing arguments for openly sharing code and data. *Scientific approaches require that researchers can properly assess the credibility of published work (Klein et al., 2018) and can independently authenticate other researchers' conclusions (Blischak et al., 2019). Thus, supporting third parties in reproducing research can increase perceptions of its robustness and reliability (Sandve et al., 2013).* This can also facilitate identification of errors in analysis (Klein et al., 2018). In addition to these motivating factors, authors may even appreciate the advantages of reproducible practices more than their peers (Piccolo and Frampton, 2016). For example, these practices can save time and effort (Sandve et al., 2013), and permanently sharing resources provides insurance against the loss of those resources (Klein et al., 2018).

A textual description of analysis in a manuscript presents an incomplete and vague account of the analytical process (Piccolo and Frampton, 2016). Sharing code helps detail the journey from the original dataset to inferential statistics (Klein et al., 2018), otherwise their software is a 'black box' (Morin et al., 2012). In the past, the possibility of issues or inconsistencies arising from computer code was overlooked (Plesser, 2018). However, it is now widely recognised that a computational analysis pipeline can present opportunities for error. Making code openly available permits *independent* reproduction of all computational processes (Stodden et al.,

2016). This, in turn, can engender trust, promote collaboration, and facilitate new applications (Jiménez et al., 2017). *Each stage of processing must be included (Sandve et al., 2013) and any files produced using the analytical pipeline should be expendable, since reproducing them using the code supplied should be trivial (Marwick et al., 2018). For full transparency, data should be supplied in a raw, unprocessed form (White et al., 2013).* Keeping raw data separate from other files ensures that the original file is not altered and the stages of processing are clear (Marwick et al., 2018). *Other resources, such as stimuli and experiment scripts should also be shared alongside data and code (Klein et al., 2018).*

The FAIR principles (Wilkinson et al., 2016) propose that data (and metadata) should be Findable (easily discovered), Accessible (easily obtained), Interoperable (easily integrated with other tools), and Reusable (easily employed beyond their original use). FAIR principles are also relevant to other computational tools (Lamprecht et al., 2020), with similarities to Open Source Software, which does not place limits on who may examine, adapt and extend the underlying code (Jiménez et al., 2017).

When sharing resources, a researcher's choices can either assist or obstruct re-use (Chen et al., 2019). For example, using non-proprietary file types ensures that third parties can readily access resources (White et al. 2013). Rather than personal or institutional websites, independent providers (e.g., Open Science Framework) are recommended for depositing these resources (Chen et al., 2019, Klein et al., 2018). Effective documentation is also valuable. A 'codebook' or 'data dictionary' can be used to explain the contents of a data file (Klein et al., 2018), inline comments can be used to explain code (Rule et al., 2019), and a README can be used to cover elementary information such as setup instructions (Lee et al., 2018). Documentation can also provide details on data collection and known issues (White et al. 2013). Finally, licences contribute to a research project's longevity, and provide a clear statement for third parties, ensuring that their use of resources is appropriate (Jiménez et al., 2017). Where possible, lenient licences should be employed to avoid unnecessary restrictions (White et al., 2013).

3.4.1 The Importance of Public Sharing

It is fallacious to assert that if authors consistently shared data and code *on request*, freely available access would be unnecessary. To begin with, papers outlive their authors, and requests obviously cannot be fulfilled by an author after they die (Klein et al., 2018). *Empirical research further demonstrates why it is important to share resources publicly.* In a study of 204 papers from a journal which *required* authors to provide data and code on request, only 44% delivered on this promise (Stodden et al., 2018). Where research code is not publicly available, various issues preclude procurement. These include local storage failures, restrictive institutional licences, concern about potential use, and concern about labour involved in providing support (Collberg & Proebsting, 2016). *Provision of data and code on request simply cannot be guaranteed, necessitating public sharing.* In the field of data visualisation research, public sharing has historically been uncommon. Of papers submitted to the VIS 2017 conference,

15% shared materials openly and 6% shared data openly (Haroz, 2018). *Greater transparency would increase the credibility of data visualisation research and facilitate identification and rectification of issues in published work (Kosara and Haroz, 2018).*

Researchers' working practices and technological solutions both contribute to reproducibility. Whilst it has been suggested that behaviour and technology play *equal* roles (Sandve et al. 2013), *others argue that innovations have been so effective that researchers' engagement with these tools is now the primary challenge* (Grüning et al., 2018). *Researchers report that several factors impede or deter their sharing of research data, including lack of expertise, lack of precedent, and lack of time (Houtkoop et al., 2018).*

3.5 Effective Programming Practices

Conducting analysis using an automated approach has three main benefits over manual processing: increased reproducibility, increased efficiency, and reduced error (Sandve et al., 2013). Writing functions in a modular style can avoid redundant repetition, promotes comprehension and supports reuse of code (Wilson et al., 2015). *Researchers should also split code into appropriate chunks which each achieve a clearly-defined goal (Rule et al., 2019).* These techniques share many similarities with the Unix philosophy (Gancarz, 2003). This approach to computer programming emphasises simplicity, modularity, and reusability.

The task of preparing data prior to analysis is an important aspect of working with data. Wickham (2014) presents a set of tools, and underlying theory for this task, arguing that analysis can be facilitated by ensuring that data is in the correct structure. This structure is known as 'tidy' data, which consists of a column for each variable (each type of measurement) and a row for each observation (each unit measured). A principled approach simplifies the process of creating a tidy dataset using Wickham's functions. Because each function treats data in a standardised manner, various functions can be employed in concert. The collection of R packages containing these functions (the 'Tidyverse') was designed with a concern for *humans*, not just computational performance (Wickham et al., 2019), so Tidyverse-style code is likely to promote comprehension (Bertin and Baumer, 2020).

Several other coding behaviours can facilitate reproducibility. For example, *absolute* file paths refer to a specific directory on a user's machine, which will not be replicated on other users' machines. Using *relative* file paths, which locate files in relation to the project directory, ensure code is *portable* and can be used on any machine (Bertin and Baumer, 2020). *Additionally, independent researchers cannot successfully verify findings if only an approximate resemblance is achieved.* Therefore, for any process involving random number generation, a random seed must be specified within the script, to ensure exact reproduction of results (Sandve et al., 2013).

3.5.1 Literate Programming and Dynamic Documents

Knuth (1984) presented a novel perspective on comprehensibility in computer programming which has been influential in the literature on computational reproducibility. Knuth's premise is that a programming script should not be regarded primarily as a set of instructions for a computer to follow, but a tool to assist humans in understanding those instructions. This approach, known as 'literate programming', involves pairing code with corresponding text, such that reporting and documentation are closely linked to underlying code (Sandve et al., 2013; Piccolo and Frampton, 2016). Dynamic documents allow authors to mix code and narrative within a single file, with results updated whenever the document is rendered. Producing (and re-producing) an entire manuscript using a dynamic document offers opportunities to easily observe the implementation of code used for each aspect of analysis (Peikert and Brandmeier, 2021). In addition to descriptive and inferential statistics, data visualisations may also be rendered dynamically (FitzJohn et al., 2014). This efficient format enhances transparency (Holmes et al., 2021), supports interactivity (Rule et al., 2019) and avoids errors which can occur when manually collating results (Peikert and Brandmeier, 2021). Including computationally-expensive code (e.g., complex statistical models) within a dynamic document can be problematic since this code is executed every time the document is rendered (FitzJohn et al., 2014). However, capacity for model caching provides a convenient antidote. This facilitates access to results by storing the output from models, which is then only updated when relevant data and code are updated.

3.6 Computational Environments

Providing data and code is necessary, but not sufficient, for guaranteeing reproducibility. For example, research has found that even when the nominally required resources are available, it is not always possible to reproduce results exactly (Stodden et al., 2018), or even to execute the code (Collberg & Proebsting, 2016). In a high-profile case, a publicly-accessible Python script for processing organic chemistry data relied on the ordering of files by the Windows operating system, producing erroneous results for Linux users (Neupane et al., 2019). A study using an automated approach to test the execution of 379 Python scripts from academic research found that success depended in part on the Python version used and the presence of files capturing dependencies (Trisovic et al., 2021). Another study used a similar approach to test over 9000 R scripts (Trisovic et al., 2022). Approximately three in four scripts produced errors when executed. Implementing a code-cleaning algorithm reduced this number, but the majority (56%) still failed to run successfully. This indicates that good programming practices can improve code but cannot totally eliminate issues. Another source of error was incompatibility of R software versions and required packages. *Thus, a failure to recreate the computational environment used when originally running the script prevented successful execution.*

Peng (2011) argues that reproducibility can be characterised as a spectrum. Sharing code offers some benefits over a standalone publication, providing data increases reproducibility

further, but ensuring that the code can be precisely executed is even better. Each researcher's unique preferences and proficiencies result in roughly the same number of computational environments as individual researchers, illustrating the benefit of recording one's computational environment (Nüst et al., 2017). Additionally, software under continuous development, such as the Tidyverse collection of packages, is frequently updated, meaning code can stop functioning unless specific versions are recorded (Holmes et al., 2021). Other software dependencies and parameter settings also complicate reproduction, requiring precision and comprehensiveness in documentation in order to achieve full *computational reproducibility* (Piccolo and Frampton, 2016).

3.6.1 Capturing Computational Environments Using Containers

Like many other aspects of reproducibility, innovations in software have made it possible for researchers to capture their computational environments. R package managers, such as *renv* (Ushey, 2020) conveniently load specific package versions for individual projects. However, they do not guarantee computational reproducibility, because they do not preserve the version of R in the same way (Holmes et al., 2021) or support additional dependencies (Peikert and Brandmeier, 2021, Nüst et al., 2017). Containerisation technology offers an effective solution. A 'container' can capture a much greater extent of the computational environment than a package manager (Grüning et al., 2018). This technology also provides an efficient and principled approach for recreating the environment, compared to a list of instructions for manual execution (Marwick et al., 2018).

Docker (Merkel, 2014) is a popular tool for generating containers. This process begins with a Dockerfile: a text-based file which provides instructions for installing specific package versions and loading other dependencies and resources. The Dockerfile is used to build a Docker image, which captures the computational environment. When this image is running, the environment is activated, and users may interact with this environment (Nüst et al., 2020b, Boettiger and Eddelbuettel, 2017).

Collating all dependency information in a single Dockerfile provides simplicity, and ensures that the original computational environment can be reproduced even after updating the software. Since the primary objective is ensuring reproducibility, this approach prioritises openness and human readability over optimising performance (Nüst et al., 2020b, Boettiger, 2015). As well as simple implementations, complex arrangements can be accommodated, but present additional challenges. For example, dynamic document generation may also require specifying LaTeX dependencies (Boettiger, 2015).

3.6.2 Rocker for Capturing R Environments

Researchers can save time and ensure consistency by using pre-existing Docker images (Nüst et al., 2020b). One particularly valuable example of this is Rocker which captures R environments

for use in Docker. This tool provides portable R environments for use with a variety of systems, facilitating computational reproducibility (Boettiger and Eddelbuettel, 2014). Consequently, any researcher can execute, edit, and extend R code in a replica of the environment originally used for its development. Developing Rocker images involves a trade-off between generalisability and specificity. An image designed to be too widely applicable would be cumbersome, but images with overly-specific use cases would be hard to find (Boettiger and Eddelbuettel, 2017). The solution involves providing base images that are easily expanded for specific requirements, with various Rocker images 'stacked' together as required, avoiding unnecessary complexity (Nüst et al., 2020a).

3.6.3 Comparing Containers with Virtual Machines

Virtual machines perform a similar function to containers. However a notable difference is that virtual machines are large, whilst containers are comparatively lightweight (Piccolo and Frampton, 2016). This difference is due to the fact that virtual machines use their own kernel, whereas containers use the operating system kernel provided by the local machine. This reduces the relative size of a container, and enhances its computational power (Cito et al., 2016). Thus, virtual machines may be considered more comprehensive than containers, offering a greater degree of separation from the characteristics of the host machine (Grüning et al., 2018, Piccolo and Frampton, 2016). However, containers are typically compatible with version control systems (Piccolo and Frampton, 2016) and offer greater transparency (Nüst et al., 2020). Furthermore, due to their modular features, making minor adaptations is trivial with a container but comparatively prolonged with a virtual machine.

3.7 Pragmatism Over Perfectionism

Despite the myriad recommendations for best practice, a principle often endorsed in the literature on reproducibility concerns the merits of small efforts. *Taking some steps to increase reproducibility still enhances a project's quality compared to neglecting this aspect altogether (Piccolo and Frampton, 2016).* Withholding resources in pursuit of continuous refinement risks never sharing them at all. This fallacy is captured by the maxim 'the best is the enemy of the good'. Analysis code does not need to be perfect in order to be useful to others (Klein et al., 2018), and it is impossible to benefit from external inquiry if the code is not shared (Barnes, 2010). Barnes (2010) argues that perceived limitations simply reflect that the code works only for the specific scenario at hand; inessential improvements are by definition not required for basic functioning. *Researchers ought to accept these limitations and share their code anyway.* In addition to code, this notion has also been applied to metadata (White et al., 2013) and containerisation (Nüst et al., 2020).

3.8 The Approach to Reproducibility in This Thesis

The following describes the different aspects of reproducibility for the subsequent empirical studies presented in this thesis. Whilst this work does not follow a pre-defined workflow, the approach closely resembles published workflows (e.g., van Lissa et al., 2020; Peikert and Brandmeier, 2021).

3.8.1 Data, Code, and Dynamic Documents

For each study, raw data is provided. The only pre-processing of this data was the essential removal of sensitive information (transparently documented in corresponding scripts). All subsequent processing, from data cleaning to data wrangling, is included in a Quarto dynamic document (Allaire, et al., 2022), which also includes all data analysis, visualisation, and accompanying text. Therefore, consistent with the principles of literate programming, textual descriptions are presented in conjunction with corresponding code (Sandve et al., 2013).

The process of selecting appropriate statistical model specifications can be opaque, involving random effects, convergence issues, and additional parameters. In the interest of transparency, consistency, and statistical rigour, I use the *buildmer* package (Voeten, 2022) to automatically determine appropriate statistical model structures. This provides a reproducible account of the steps preceding identification of each statistical model employed in analysis, reducing human error and documenting a process as well as its outcome (Rule et al., 2019). As this package is available CRAN (Comprehensible R Archive Network), its source code is archived and transparent.

3.8.2 Docker Containers

Capturing dependencies requires reproduction of the computational environment used (Boettiger, 2015). Each study in this thesis is associated with a Dockerfile, which can be used to build a Docker container with the appropriate R version and package versions used during analysis. Employing Rocker images provides an Integrated Development Environment (RStudio), and speeds up construction of the Docker image. In each container, an entire manuscript can be generated from scratch. The Dockerfiles also provide important project metadata in a human- and machine-readable format (Leipzig et al., 2015).

3.8.3 Experiment Resources

In experimental psychology, sharing stimuli and experiment scripts is another important aspect of transparent research practice (Klein et al., 2018). All data visualisations shown to participants, along with all code used to generate those visualisations, has been made available. Experiments were programmed using PsychoPy, which developed as a tool for conducting

open and reproducible research (Peirce et al., 2019). The underlying technology is open source, the experiment scripts use non-proprietary file formats, and the ability to specify particular software versions avoids new releases breaking older code. Its integration with GitLab version control software means that each experiment is packaged in a public online repository. An entire project's resources can be downloaded to a local machine, and an interactive version of the experiment can be run online.

3.9 Conclusion

This chapter has discussed *how a lack of reproducibility in published research can reduce credibility, and has revealed how various approaches can increase reproducibility.* At the heart of these recommendations is the need to comprehensively share resources and embrace technological solutions. Making research code and raw data openly available helps an opaque analysis process to become transparent. When an entire paper's results can be fully reproduced by an independent third party, they can be thoroughly verified.

For each empirical study in this thesis, I share raw data alongside code packaged in a dynamic document. This provides transparency, illustrating exactly how the study's findings were generated. In addition, creating Docker containers for each study allows the analyses to be reproduced in their original computational environment. This comprehensive approach is uncommon in research on data visualisation, therefore this work serves as an example of how research in this field may be made more reproducible.

Transparent about epistemological approach.

Preface to Chapter 3

This thesis presents a series of empirical studies which investigate the cognitive processing of the absolute magnitude of values presented in data visualisations. Conceptions of absolute magnitude refer to how large or small numerical values are. The first set of experiments in this thesis provides the foundation for this research project.

Prior research has reported that values presented at higher positions were judged as higher in magnitude than values presented at lower positions (Sandman et al., 1994). The first set of experiments in this thesis also explores the role of *physical position* in informing magnitude judgements. I attempt to replicate Sandman et al.'s (1994) general finding, then expand upon these results to examine the underlying cognitive mechanism.

Rationale For Experimental Design

In these experiments, participants observed a series of simple dot plot visualisations of fictitious data, with numerical values encoded through their position on the vertical axis. Systematically manipulating the axis limits surrounding plotted values causes the same values to appear at high or low positions. Therefore, participants' ratings of the magnitude of plotted values reveals the effect of this design choice on interpretations. The simplicity of the stimuli provides high experimental control, but the use of meaningful scenarios concerning risk-related events increases ecological validity.

In order to gain further insight into how absolute magnitude is processed, these experiments also examine whether the *physical* positions or *relative* positions of plotted values contribute to magnitude judgements. This distinction concerns whether interpretations are influenced primarly by the association between high physical positions and high values, or the numerical context within which values are plotted. Dot plots with *inverted* axes are presented to test these two competing explanations.

4 Magnitude Judgements Are Influenced by Data Points' Relative Positions Within Axis Limits

When visualising data, chart designers have the freedom to choose the upper and lower limits of charts' numerical axes. Axis limits can determine the physical characteristics of plotted values, such as the physical position of data points in dot plots. In three experiments (total N=420), we demonstrate that axis limits affect viewers' interpretations of the magnitudes of plotted values. Participants did not simply associate values presented at higher vertical positions with greater magnitudes. Instead, participants considered data points' relative numerical positions within the axis limits. Data points were considered to represent larger values when they were closer to the end of the axis associated with greater values, even when they were presented at the *bottom* of a chart. This provides further evidence of framing effects in the display of data, and offers insight into the cognitive mechanisms involved in assessing magnitude in data visualisations.

4.1 Introduction

Context is crucial for effectively judging the magnitude of numbers. A 10% probability is twice as great as a 5% probability, but in the absence of context, it is unclear whether this value should be considered large or small. When referring to the chance of losing one's job, a 10% probability may be considered large, but when referring to the chance of losing a sports bet, a 10% probability may be considered small.

Contextual cues may influence interpretation of magnitude in data visualisations. One such cue is the range of values on an axis, which can serve as a frame of reference for assessing whether a data point represents a large or small number. Figure 4.1 (a reproduction of a similar bar chart from the New York Times), which plots over time the number of Black members of the U.S. senate, provides a striking illustration. Unusually, the y-axis does not terminate just above the highest plotted value. Instead, the y-axis extends all the way to the maximum possible number of senators: 100. As a result, bars representing Black senators are confined to the very bottom, visible just above the x-axis, and a significant expanse of blank space looms above them. This framing situates plotted data points in their numerical context, thus conveying small magnitude.

It is unclear exactly how a viewer's inferences about magnitude might be influenced by axis range. Different axis limits present data points at different positions, so one possible explanation is that viewers interpret the magnitude of data points at higher positions as 'high' and those at lower positions as 'low'. Alternatively, axis limits may provide numerical context: plotted values may be judged as small in magnitude when the potential for larger values is clearly displayed. The present pair of experiments demonstrates the influence of axis limits on viewers' interpretations and explores which of these two accounts best explains how axis limits contribute to the communication of magnitude.

4.1.1 Overview

In three experiments, we manipulated the axis limits surrounding plotted data. The same data points either appeared close to the upper end of an axis range, or close to the lower end. Likert scale ratings of values' magnitudes were higher when data points were positioned close to the end of the axis which was associated with higher numbers. By employing charts with conventional and inverted y-axis orientations to distinguish between possible explanations, we reveal that magnitude judgements are influenced by data points' relative positions within the axis limits.

4.2 Related Work

4.2.1 Effects of Axis Limits on Comparison Judgements

Several studies have explored the role of axis limits in data visualisation. Research has typically focused on how axis limits can alter impressions of the *difference between* presented values. For example, when axis ranges are expanded to create blank space around a cluster of data points, correlation between those points is judged as stronger (Cleveland et al., 1982). Participants also rate the differences between values in bar charts as greater when the vertical gap between bars is larger due to a truncated y-axis (Pandey et al., 2015).

Correll et al.'s (Correll et al., 2020) experiments found that greater truncation resulted in higher effect-size judgements in both line charts and bar charts. They found no reduction in effect size judgements when truncation was communicated using graphical techniques (e.g., axis breaks and gradients). Truncation effects also persisted even when participants estimated the values of specific data points. This suggests the bias is driven by initial impressions, rather than by a misinterpretation of the values portrayed by graphical markings. The unavoidable consequence, Correll et al. suggest, is that designers' choices will influence viewers' interpretations whether axes are truncated or not.

Choosing an appropriate axis range involves a trade-off between participants' bias (over-reliance on the visual appearance of differences) and their sensitivity (capacity to identify

actual differences) (Witt, 2019). Just as a highly truncated y-axis can exaggerate trivial differences between values, an axis spanning the entire range of possible values can conceal important differences. Based on participants' judgements of effect size, Witt (2019) found that bias was reduced and sensitivity increased when using an axis range of approximately 1.5 standard deviations of the plotted data, compared to axes which spanned only the range of the data, or the full range of possible values. This provides further evidence of a powerful association between the appearance of data, when plotted, and subjective interpretations of differences between data points.

Further evidence of truncation effects, provided by Yang et al. (2021) improves on the design of previous studies which employed only a few observations per condition (Pandey et al., 2015) or small sample sizes (Witt, 2019). Participants' ratings of the difference between two bars consistently provided evidence of the exaggerating effects of y-axis truncation. Yang et al. (2021) noted that increasing awareness does not eliminate this effect, which may function like an anchoring bias, in which numerical judgements are influenced by reference points (Tversky and Kahneman, 1974). Another potential explanation discussed draws upon Grice's cooperative principle (Grice, 1975). According to this account of effective communication, speakers are assumed to be in cooperation, and so will communicate in a manner that is informative, truthful, relevant, and straightforward. Analogously, a viewer will assume that a numerical difference in a chart must be genuinely large if it appears large, else it would not be presented that way. Effective visualisations should be designed so a viewer's instinctive characterisation of the data corresponds closely to their interpretation following a more detailed inspection (Yang et al., 2021).

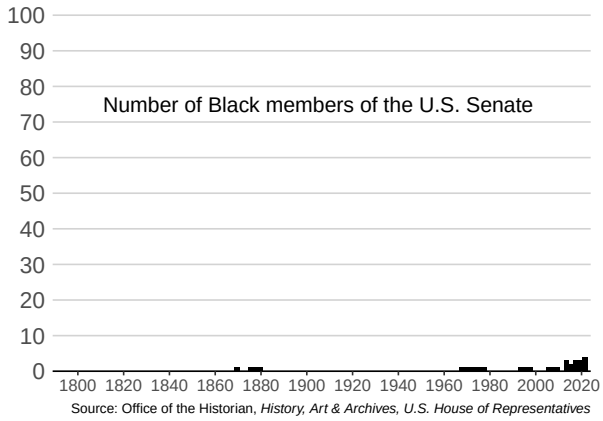


Figure 4.1: A reproduction of a bar chart from the New York Times. The y-axis limit is defined by the largest possible value, rather than the largest observed value, thus the magnitude of plotted values appears particularly small.

4.2.2 Effects of Axis Limits on Magnitude Judgements

The above research consistently demonstrates that the magnitude of *the difference between values* is interpreted differently depending on the axis limits employed. The present investigation is concerned with how interpretations of the magnitude of *the values themselves* are affected by a chart's design.

Empirical evidence demonstrates that judgement of a value's magnitude can depend on its relationship to a grand total or to surrounding values. This can influence interpretation of verbal approximations, and also absolute values. For example, participants instructed to take 'a few' marbles picked up more when the total number available was larger (Borges and Sawyers, 1974) and rated satisfaction with the same salary as higher when it appeared in the upper end of a range, compared to the lower end (Brown et al., 2008). As well as context, vertical position also plays a role in magnitude judgements. For example, children appear to intuitively understand the relationship between height and value (Gattis, 2002) Both the physical world, and language (e.g., spatial metaphors), provide countless examples where 'higher' is associated with 'more', and 'lower' with 'less', and this principle has been adopted as a convention in data visualisation (Tversky, 1997).

In charts, inversions of the typical mapping between magnitude and vertical position charts can lead to misinterpretations (Okan et al., 2012; Pandey et al., 2015; Woodin et al., 2022). Furthermore, when a company's financial performance was displayed entirely in the bottom fifth of a line chart, the company was perceived as less successful, compared to when the axis did not extend above the maximum value (Taylor and Anderson, 1986). Sandman et al. (1994) investigated assessments of magnitude in risk ladders, where greater risks are presented at physically higher positions on a vertical scale. Participants rated asbestos exposure as a greater threat when it was plotted at a higher position, compared to a lower position.

The above findings can be regarded as preliminary evidence that changing axis limits may affect appraisals of data points' magnitudes. However, the evidence is not substantial. Taylor and Anderson (1986) did not disclose how judgements were elicited, or provide details of their sample size. Sandman et al. (1994) only explored responses to one specific risk (asbestos), and each participant only took part in a single trial. The perceived threat measure was a composite of several separate ratings, preventing diagnosis of whether manipulations affected interpretations of the plotted information in particular, or just related concepts. Further, both studies introduced a confounding variable by adjusting the difference between the minimum and maximum y-axis values across conditions. Stronger evidence is required regarding how axis limits may bias inferences about magnitude, and the cognitive mechanisms involved in generating these inferences.

4.2.3 Judgements of Event Outcomes

In the present study, participants viewed charts showing fictitious data on the chance of particular events occurring. This provided participants with a purpose; evaluating information about event outcomes is a more meaningful task than assessing how 'large' an abstract value is. Each value was represented using a single dot on a percentage probability scale. Our use of dot plots for conveying percentages was motivated by their simplicity and use of a single encoding channel (position), thus avoiding confounding variables from other encoding channels.

Presenting data about events with negative consequences warranted consideration of the cognitive processing of this information. These events are composed of two core components: 1) chance of occurrence and 2) outcome magnitude (severity). Individuals' assessments of chance and severity are not necessarily independent. Events are perceived as more likely when they are described as having more severe consequences (Harris et al., 2009; Harris and Corner, 2011). In a similar manner, events are associated with more substantial consequences when they are described as more likely (Kupor and Laurin, 2020).

One account suggests that perceptions of probability and outcome magnitude are related because they are both assumed to reflect the potency of the event's cause (probability-outcome correspondence principle; (Keren and Teigen, 2001)). According to this account, probabilities can occasionally provide meaningful indications of outcome magnitude (e.g., rainfall), but it is inappropriate to apply this perspective to all situations (e.g., volcanic eruptions). Therefore, even though charts in the present study only display the *chance* of events occurring, assessments of the *severity* of events' consequences may also differ between conditions. Collecting separate judgements for chance and severity of consequences provides a clearer picture of how the manipulation affects distinct aspects of participants' representations. Our use of Likert scales (with discrete options) rather than visual analogue scales (with continuous options; Sung and Wu (2018)) prevents participants from simply mapping probability percentages directly onto a linear scale.

4.2.4 Data Visualisation Literacy

When faced with charts that violate graphical conventions by using atypical scales, individuals with low data visualisation literacy are more likely to draw on data points' physical positions when making inferences about their magnitudes (Okan et al., 2016, 2012). We administered Garcia-Retamero et al.'s (Garcia-Retamero et al., 2016) subjective graph literacy measure to determine whether responses to our manipulation of axis limits were associated with data visualisation literacy.

4.3 Experiments

We conducted three experiments manipulating y-axis limits in visualisations of fictitious data. This manipulation altered the physical positions of data points in a chart, but crucially the numerical values themselves remained the same.

Experiment 1 sought to establish whether y-axis limits affected magnitude judgements. To provide context for participants, text accompanying the charts outlined (fictitious) scenarios involving a specific negative outcome (e.g., loss on financial investment, delayed flights, etc.). Three plotted data points in each chart represented the chance of the negative outcome occurring (%) for three instances associated with the scenario (e.g., three investment opportunities, three airlines, etc.).

Experiment 2 introduced another factor in addition to the manipulation of y-axis limits. Half of the visualisations presented employed inverted y-axis orientations, where data points at lower physical positions represented greater values. This 2x2 experiment allowed us to investigate whether magnitude judgements were driven by data points' absolute positions, or their relative positions within the context of the axis limits.

Experiment 3 manipulated y-axis limits in inverted charts only, providing clarity on the ambiguous results of the previous experiment. Importantly, the use of inverted charts should not be considered an endorsement (see issues above). However, they serve to distinguish between two possible explanations, since they reverse the typical associations between physical position and magnitude.

Ethical approval was granted by The University of Manchester's Division of Neuroscience & Experimental Psychology Ethics Committee (Experiment 1: Ref 2021-11115-18258; Experiment 2: Ref 2021-11115-20464; Experiment 3: Ref. 2021-11115-20745). Data, analysis code, experimental scripts, materials and a link to run the experiments are available at https://osf.io/3epm2/. We also provide all necessary resources for running a Docker container, within which the computational environment used for analysis is recreated, meaning a fully-reproducible version of this paper can be generated.

4.3.1 Experiment 1

4.3.1.1 Method

4.3.1.1.1 Materials

4.3.1.1.1.1 Datasets

For each dataset, we generated three values from a normal distribution. Population means were specified manually in order to represent plausible values for the probability of the event occurring (28% - 72%). All datasets had a population standard deviation of 0.5. The same dataset was employed for both of the experimental conditions associated with a given event scenario.

4.3.1.1.1.2 Charts

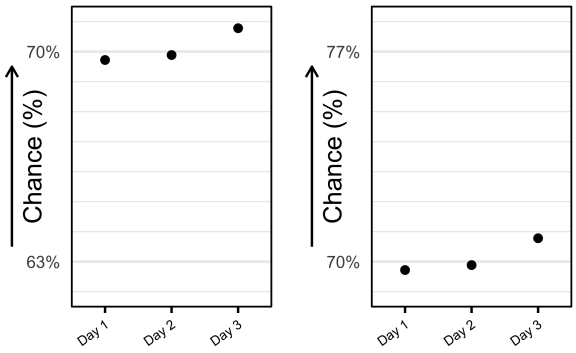


Figure 4.2: Example charts, taken from Experiment 1. The *high physical position* condition (left) presents data points near the top of the chart; the *low physical position* condition (right) presents the same data points near the bottom of the chart.

Datasets were displayed using dot plots. In experimental trials (n = 40), upper and lower axis limits were manipulated such that data points either appeared in the top third of the chart (high physical position: Figure 6.1 , left) or in the bottom third (low physical position: Figure 6.1, right).

The y-axis range in each chart was 10 percentage points. Horizontal gridlines appeared at one-unit increments. The horizontal gridlines 1.5 units from the extremes were labelled with numerical values.

Filler trials (n = 15) and attention check trials (n = 5) presented data points in the middle third of the chart. Filler trials employed this additional variation to prevent participants from identifying the purpose of the study.

4.3.1.1.2 Procedure

The experiment was programmed in PsychoPy (version 2021.1.4, (Peirce et al., 2019)) and hosted on pavlovia.org. Participants were instructed to complete the experiment on a desktop computer or laptop, not a tablet or mobile phone. Instructions explained that their task involved assessing the chance and severity of negative outcomes in various scenarios involving

risks and noted that some scenarios might appear similar to other scenarios. Participants were asked to complete the task as quickly and accurately as possible. Two practice trials preceded the experiment proper.

An example of a single trial is shown in Figure 5.3. Participants provided two responses in each trial: a rating of the chance of the negative event occurring; and a rating of the severity of the consequences if that negative event occurred. Both 7-point Likert scales had two anchors at their extremes: *'Very unlikely'* and *'Very likely'*; for the 'Chance' scale and *'Very mild'* to *'Very severe'*. for the 'Severity' scale. All other points were unlabelled. Text specified that answers should be given in response to the plotted data (e.g., *"If you camp on one of these days…"*). The term 'chance' was used instead of 'probability' to avoid confusion with the standard 0-1 scale for probabilities, and to reflect casual usage.

Participants could change their responses as many times as they wished before proceeding to the next trial, but could not return to previous trials. In attention check trials, participants were instructed not to attend to the chart, and instead to provide specified responses on the Likert scales.

Before exiting the experiment, participants were informed that all presented data were fictitious and guidance was provided in case of distress.

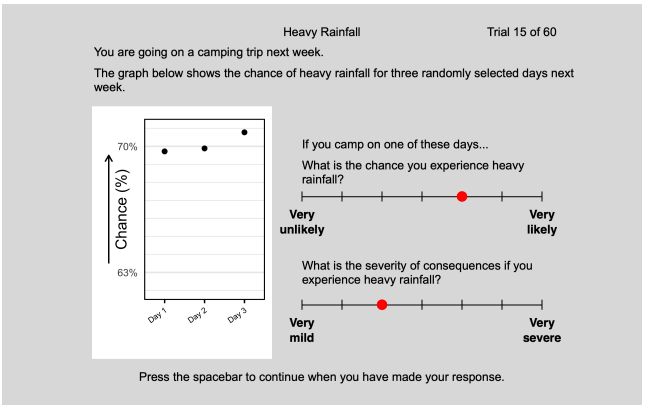


Figure 4.3: An example trial, taken from Experiment 1. Participants provided two ratings in each trial: the chance of an event occurring (magnitude rating), and the severity of consequences.

4.3.1.1.3 Design

We employed a repeated-measures, within-participants design. Participants encountered scenarios from experimental trials twice: once with data presented at a high physical position and once with data presented at a low physical position.

Materials were divided into two lists to minimise the likelihood of different versions of the same scenario appearing in close succession. One list contained half of the high-condition items and

half of the low-condition items for the experimental scenarios. The other list contained the alternate versions of each of the experimental scenarios. Fillers and attention check questions were split between the two lists, and did not appear more than once. The order of the two lists was counterbalanced across participants. Within each list, scenarios were presented in a random order.

4.3.1.1.4 Participants

The experiments were advertised on Prolific.co, a platform for recruiting participants for online studies. Normal or corrected-to-normal vision and English fluency were required for participation.

Data were returned by 160 participants. Ten participants' submissions were rejected because they answered more than two of 10 attention check questions incorrectly. This left a total of 150 participants whose submissions were used for analysis (52% male, 45% female, 3% non-binary). Mean age was 31.49 ($SD = 12.47$)[1]. The mean data visualisation literacy score was 21.28 ($SD = 4.58$), out of a maximum of 30. Participants whose submissions were approved were paid £3.55. Average completion time was 25 minutes.[2]

4.3.1.1.5 Analysis Technique

Analyses were conducted using R (version 4.2.1; (R Core Team, 2022)).

Likert scales express granularity at the level of ordinal data. They record whether one rating is higher or lower than another, but not the magnitude of this difference. Therefore, Likert scales do not necessarily capture values from latent distributions (mental representations) in a linear manner. The distance between one pair of points and another pair may appear equal, but may represent different distances on the latent distribution. Therefore, it is inappropriate to analyse Likert scale data with metric models, such as linear regression (Liddell and Kruschke, 2018). Throughout this paper, we construct cumulative link mixed-effects models, using the *ordinal* package (version 2019.12-10, (Christensen, 2019)) to analyse Likert scale ratings. Odds ratio effect sizes were converted to Cohen' d values using the *effectsize* package (version 0.8.2, (Ben-Shachar et al., 2020)).

Selection of model random effects structures was automated using the *buildmer* package in R (version 2.3, (Voeten, 2022)). The maximal random effects structure included random intercepts for participants and scenarios, plus corresponding slopes for the position variable (Barr et al., 2013). *buildmer* initially identified the most complex model which could successfully converge. It subsequently removed terms which did not contribute substantially to explaining variance in ratings.

[1] Age data was unavailable for 2 participants

[2] Timing data were unavailable for two participants.

4.3.1.2 Results

4.3.1.2.1 Magnitude Ratings

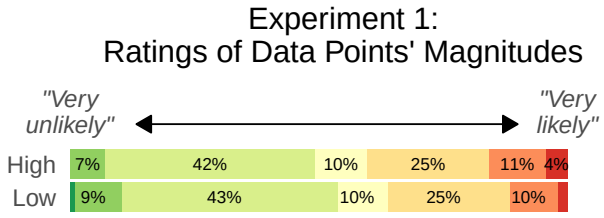


Figure 4.4: The distribution of Likert scale ratings of data points' magnitudes. The width of each response option represents the proportion of ratings recorded for that option. Note that data points presented at high physical positions (top) elicited a larger proportion of ratings on the right-hand side (representing greater magnitudes), compared to data points at low physical positions (bottom), which elicited a larger proportion of ratings on the left-hand side (representing smaller magnitudes).

Figure 4.4 plots the distribution of participants' ratings of data points' magnitudes, showing that values presented at high physical positions elicited a greater proportion of responses at the higher end of the rating scale than values presented at low physical positions.

A likelihood ratio test reveals that a model including physical position as a fixed effect explains significantly more variability in ratings than a model which does not include physical position as a fixed effect ($\chi^2(1) = 74.21$, $p < .001$). Data points' magnitudes were rated as greater when those data points were presented at high physical positions, compared to when the same data points were presented at low physical positions ($z = 8.57$, $p < .001$).

The odds ratio for the difference between conditions is 1.61 (95% CI [1.44, 1.79]). Participants were 1.61 times more likely to respond with a higher magnitude rating to data points presented at high positions than data points presented at low positions. This is equivalent to a Cohen's d value of 0.26, a small effect size.

This model included random intercepts for each participant and each scenario.

4.3.1.2.2 Severity Ratings

For ratings of the severity of consequences, a likelihood ratio test reveals that a model including physical position as a fixed effect explains significantly more variability in ratings than a model which does not include condition as a fixed effect: ($\chi^2(1) = 6.16$, $p = .013$). The severity of consequences was rated as greater when data points representing the chance of an event occurring were presented at high physical positions, compared to when the same data points were presented at low physical positions ($z = 2.50$, $p = .012$).

The odds ratio for the difference between conditions is 1.21, 95% CI [1.04, 1.41]. Participants were 1.21 times more likely to respond with a higher severity rating to data points presented at high positions than data points presented at low positions. This is equivalent to a Cohen's d value of 0.11, a very small effect size.

This model employed random intercepts for each scenario, plus random intercepts and slopes for each participant. The slopes modelled, for each participant, the average difference between responses to data presented at different positions (henceforth referred to as *by-position slopes*).

4.3.1.2.3 Data Visualisation Literacy

We also generated two additional models, to test whether or not the above results could be explained by differences in data visualisation literacy. These models were identical to the above models except for the inclusion of participants' subjective data visualisation literacy scores as an additional fixed effect. Adjusting for participants' data visualisation literacy scores did not eliminate the effects of data points' positions on ratings of the magnitude of data points themselves ($z = 8.57$, $p < .001$, odds ratio = 1.61, 95% CI [1.44, 1.79]) or severity of consequences ($z = 2.51$, $p = .012$, odds ratio = 1.21, 95% CI [1.04, 1.41]).

4.3.1.3 Discussion

This experiment demonstrates that axis limits, which determine the position of plotted values, influence inferences about data points' magnitudes. Participants rated *the same values* as greater when these values were plotted at high positions, compared to low positions. Even though the charts only displayed data on the chance of negative outcomes occurring, ratings of severity of consequences were also greater when data points were presented at high positions. Accounting for differences in participants' data visualisation literacy did not alter the pattern of results.

4.3.2 Experiment 2

4.3.2.1 Introduction

Experiment 1 (E1) found that participants associated data points with greater magnitudes when those data points were positioned near the *top* of a chart, compared to when the same data points were positioned near the *bottom* of a chart.

One possible explanation for this finding is that participants made simple associations between absolute position and magnitude, equating physically higher data points with larger magnitudes and physically lower data points with smaller magnitudes. This is congruent with

well-established conceptual metaphors for magnitude, where greater vertical positions denote greater magnitudes (Tversky, 1997).

An alternative explanation is that participants used the y-axis as a frame of reference for assessing the magnitude of plotted values. For example, when considering data points near the bottom of the axis, participants may have recognised the potential for values larger than those observed, consequently associating plotted values with smaller magnitudes.

E1 does not provide a means of differentiating these competing explanations. Drawing inferences from data points' absolute positions would bias magnitude judgements in the same direction as drawing inferences from their relative positions. A high magnitude is implied by a data point's high physical position *and* its superior position in the context other of presented values. Therefore, further investigation is required in order to distinguish between the two competing explanations.

Plotting numerical values along the x-axis would not assist in answering this question, since values that are large in the context of the x-axis limits would be positioned on the right-hand side, which is also typically associated with larger magnitudes (Woodin and Winter, 2018). However, inverting a vertical axis changes the typical relationship between physical position and numerical value: increasingly *lower* positions represent increasingly *higher* numerical values. This means data points presented near the *bottom* of a chart are numerically *larger* than the accompanying y-axis values. Therefore, inferences invoking relative numerical position would bias magnitude judgements in the opposite direction compared to inferences invoking data points' physical positions. This is illustrated in Figure 4.5.

In E2, we manipulate data points' physical positions by changing axis limits (as in E1), but *also* manipulate axis orientation, by employing conventional and inverted axes (in a 2 x 2 design). This allows us to identify the mechanism responsible for the previously observed bias in magnitude judgements. Use of absolute position would be indicated by higher magnitude ratings for data points at *high* physical positions (regardless of axis orientation). Alternatively, use of relative position would be indicated by higher magnitude ratings for data points at *high* physical positions in conventional charts and *low* physical positions in inverted charts.

Previous research suggests that charts with inverted axes can be prone to misinterpretation when viewers are not informed about the inversion (Pandey et al., 2015; Woodin et al., 2022). Therefore, we provided explicit instruction to ensure participants were aware that inverted charts were presented.

4.3.2.2 Method

4.3.2.2.1 Materials

For this experiment, we used a Latin-squared design where participants only viewed one chart per scenario. In response to this, we increased the number of scenarios. This provided some

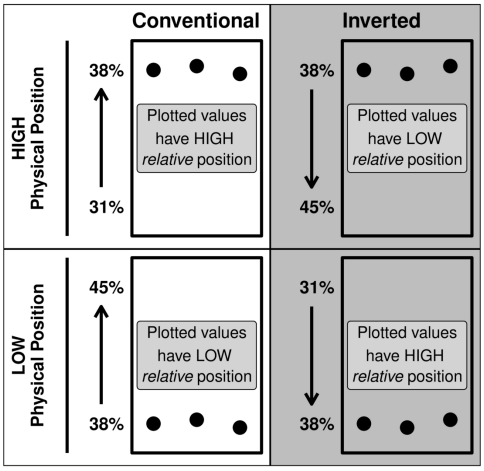


Figure 4.5: Rationale for Experiment 2: distinguishing the roles of absolute and relative position. In charts with conventional axis orientations (left column), there is congruence between data points' physical positions and their relative numerical positions in the chart. In charts with inverted axis orientations (right column), there is incongruence between data points' physical positions and their relative numerical positions in the chart. This allows us to test whether physical positions or relative numerical positions influence magnitude judgements.

compensation for the reduced experimental power caused by a reduction the number of observations per participant (as well as a reduction in participant numbers).

Two scenarios which were fillers in E1 were used as experimental scenarios[3] and three additional scenarios were created. One filler scenario was removed due to a concern about its quality (it concerned the risk to others as well as the risk to oneself). This resulted in a total of 24 experimental scenarios, 12 filler scenarios, and 5 attention check questions (41 trials in total).

4.3.2.2.2 Procedure

The experiment used PsychoPy version 2021.2.3. Participants specified the highest level of education they had received, in addition to answering demographic questions on age and gender. An additional slide in the instructions explained how to identify and interpret the different axis orientations, and encouraged participants to pay attention to this aspect of the charts:

> *You should pay attention to the direction of the arrow on the 'Chance' axis. If the arrow points upwards, the numbers in the graph get bigger as the axis goes up.*

[3]For one of these scenarios, the mean of the plotted data was also modified.

Alternatively, if the arrow points downwards, the numbers get bigger as the axis goes down.

Otherwise, the procedure was identical to E1.

4.3.2.2.3 Design

We employed a Latin-squared, within-participants design. Participants encountered each individual scenario only once, but were exposed to all combinations of physical plotting position and axis orientation throughout the experiment.

4.3.2.2.4 Participants

A viral social media post on 24th July 2021 endorsing the Prolific.co platform attracted many new users from a narrow demographic, skewing studies' participant distributions (Charalambides, 2021). Therefore, the experiment was not advertised to users who signed-up to Prolific.co after 24th July 2021. The experiment was also not advertised to those who had participated in E1.

Data were returned by 129 participants. Per pre-registered exclusion criteria, five participants' submissions were rejected because they answered more than two of 10 attention check questions incorrectly. Submissions from four other participants were excluded from the final dataset for the following reasons: maximum completion time (67 minutes) was exceeded (two participants); the submission constituted second attempt following a saving error on first attempt (one participant); data were collected prior to pre-registration (one participant). This left a total of 120 participants whose submissions were used in the analysis (49% male, 51% female). Mean age was 30.73 (SD = 17.83). 100% had completed at least secondary education. The mean data visualisation literacy score was 21.72 (SD = 4.70), out of a maximum of 30. Participants whose submissions were approved were paid £3.55. Average completion time was 21 minutes.

4.3.2.3 Results

4.3.2.3.1 Magnitude Ratings

Figure 4.6 plots the distribution of participants' ratings of data points' magnitudes, for data points presented at high and low physical positions, in charts with conventional axis orientations and inverted axis orientations.

A likelihood ratio test reveals that a model including the interaction between physical position and axis orientation as a fixed effect explains significantly more variability in ratings than a model without this interaction as a fixed effect ($\chi^2(1) = 8.22$, p = .004). There was a significant interaction between physical position and y-axis orientation (z = 2.91, p = .004).

Figure 4.6: Participants rated the chance of each negative event occurring on a 7-point Likert scale. The distribution of ratings, ranging from "Very unlikely" (far left, dark green) to "Very likely" (far right, red) is shown separately for each combination of the levels of each condition (axis orientation: conventional, inverted; data points' physical position: high, low). Note that the pattern of responses to data presented at different positions in the Conventional Axis condition appears to be the opposite to the pattern for Inverted Axis condition. When charts used conventional axes, greater magnitude ratings were more common for data presented at high physical positions, whereas when charts used inverted axes, greater magnitude ratings were more common for data presented at low physical positions.

The odds ratio associated with this interaction is 0.30 (95% CI [0.13, 0.68]). This is equivalent to a Cohen's d value of 0.66, a medium effect size.

This model employed random intercepts and by-position slopes for each scenario. Random intercepts were included for each participant, as well as slopes capturing differences in participants' responses to data presented at different positions, different orientations, and the interaction between these.

Pairwise comparisons (with Sidak adjustment) reveal that the effect of position in charts with conventional y-axis orientations (E1) was replicated (z = 3.56, p = .001). Data points' magnitudes were rated as greater when they were presented at high physical positions, compared to when they were presented at low physical positions. There was no significant difference between magnitude ratings for data points plotted at different positions when inverted axes were used (z = -1.39, p = .512). Therefore, we observe a different pattern of results when an inverted axis is used, compared to when a conventional axis is used. This suggests that differences in ratings for data points at different absolute positions in physical space are not due to simple associations between vertical position and magnitude.

4.3.2.3.2 Severity Ratings

For ratings of the severity of consequences, a likelihood ratio test reveals that a model including the interaction between physical position and axis orientation as a fixed effect explains significantly more variability in ratings than a model without this interaction as a fixed effect ($\chi^2(1) = 5.13$, p = .024). There was a significant interaction between physical position and y-axis orientation (z = 2.28), p = .022).

The odds ratio associated with this interaction is 0.69 (95% CI [0.50, 0.95]). This is equivalent to a Cohen's d value of 0.20, a small effect size.

This model employed random intercepts for each scenario. Random intercepts were included for each participant, as well as slopes capturing differences in participants' responses to data presented at different positions, different orientations, and the interaction between these.

Despite the interaction, the main effect in severity ratings from E1, different responses to data points at different positions in conventional charts, was not replicated (z = 1.53, p = .414). There was also no evidence of different responses to data points at different positions in inverted charts (z = -1.54, p = .412). This interaction appears to be driven by a weak and likely spurious difference between ratings for data points at high physical positions in inverted and conventional charts (z = -2.52, p = .047).

4.3.2.3.3 Data Visualisation Literacy

Adjusting for participants' data visualisation literacy scores did not change the pattern of results regarding ratings of the magnitude of data points themselves (z = 2.91, p = .004, odds ratio = 3.31, 95% CI [1.48, 7.42]) or the severity of consequences (z = 2.29, p = .022, odds ratio = 1.44, 95% CI [1.05, 1.98].

4.3.2.4 Discussion

In E1, when using conventional charts only, we found that displaying data within different axis limits affected magnitude judgements. However, it was unclear whether judgements were based on data points' absolute physical positions, or their relative positions within axis limits, because both would generate similar interpretations. Therefore, in E2, for half of trials, we reversed the mapping of values in physical space, so these two aspects would imply different magnitudes for a given value. *Ratings of the severity of consequences were not significantly affected by the position of data points representing the chance of negative events occurring. Accounting for differences in participants' data visualisation literacy did not alter the pattern of results.*

In E2, we replicated the primary finding from E1. In charts with conventional axis orientations, the same data points elicited different magnitude judgements when presented at different positions. These differences were consistent with magnitudes implied by data points' absolute physical positions *and* their relative positions within axis limits. However, in charts with inverted axis orientations, we did not observe the same pattern. Therefore, we can conclude that the interpreting the magnitude of data points presented at distinct physical positions depends on how axes are oriented.

Figure @ref(fig:r2-int-plot) suggests that the pattern of results for inverted charts is the reverse of the pattern for conventional charts. However, our analysis indicates that the same data points did not elicit significantly different magnitude judgements when presented at different

positions in *inverted* charts. Therefore, we cannot conclude from this analysis that magnitude judgements are driven solely by the relative positions of data points within axis limits. This lack of significant difference is likely due to a lack of experimental power. An additional experiment is required to confirm whether there is a genuine difference.

4.3.3 Experiment 3

4.3.3.1 Introduction

The interaction in E2 revealed that the influence of position on magnitude judgements depends on how different numerical values are arranged in a chart (axis orientation). The pattern of responses in inverted charts appeared to be the inverse of the pattern for conventional charts. This suggests that participants did not generate inferences about magnitude based on data points' absolute physical positions. However, the absence of a significant difference between ratings for data points at different positions in *inverted* charts prohibits the conclusion that interpretations are influenced by data points' relative positions within axis limits.

It is possible that no significant effect was detected for inverted charts due to insufficient experimental power. Unlike E1, which recruited 150 participants for a single-factor design, E2 recruited 120 participants for a 2x2 Latin-squared design. Despite an increase in the number of experimental scenarios (from 20 to 24), there were still fewer observations for each unique condition (3000 in E1 vs. 720 in E2).

In E3, we increase the experimental power and present inverted charts only, using the same experimental design as E1. This provides a clearer account of how magnitude is interpreted in inverted charts, furthering understanding of the mechanism by which axis limits influence interpretations of magnitude. Inferences based on data points' absolute physical positions would be indicated by higher magnitude ratings for data points at *high* physical positions (mirroring the finding for conventional charts). Alternatively, inferences based on data points' relative positions within axis limits would be indicated by higher magnitude ratings for data points at *low* physical positions (the reverse of the finding for conventional charts).

4.3.3.2 Method

4.3.3.2.1 Materials

Materials were identical to E1, except for the inversion of the y-axis in all charts, including practice trials.

4.3.3.2.2 Procedure

The experiment used PsychoPy version 2021.2.3. One slide in the instructions explained to participants how charts with inverted axes function: *"In all graphs in this experiment, the arrow on the 'Chance' axis points downwards, meaning the numbers get bigger as the axis goes down."*. Otherwise, the procedure was identical to E1.

4.3.3.2.3 Design

As in E1, we employed a repeated-measures, within-participants design.

4.3.3.2.4 Participants

The experiment was not advertised on Prolific.co to those who had participated in E1 or E2, or those who signed-up to Prolific.co after 24th July 2021 (due to the shift in participant demographics). Otherwise, the inclusion criteria were the same as the previous experiments.

Data were returned by 161 participants. Ten participants' submissions were rejected because they answered more than two of 10 attention check questions incorrectly. One additional participant was excluded from the final dataset because they exceeded the maximum completion time (87 minutes). This left a total of 150 participants whose submissions were used for analysis: (60% male, 40% female). Mean age was 29.64 (SD = 9.56)[4]. 100% had completed at least secondary education. The mean data visualisation literacy score was 21.87 (SD = 4.28). Participants whose submissions were approved were paid £3.45, and average completion time was 24 minutes.

4.3.3.3 Results

4.3.3.3.1 Magnitude Ratings

Figure 4.7 plots the distribution of participants' ratings of data points' magnitudes, showing that values presented at *low* physical positions elicited a greater proportion of responses at the higher end of the rating scale than values presented at *high* physical positions.

A likelihood ratio test reveals that a model including physical position as a fixed effect explains significantly more variability in ratings than a model which does not include physical position as a fixed effect ($\chi^2(1) = 46.45$, $p < .001$). Data points' magnitudes were rated as larger when those data points were presented at *low* physical positions, compared to when the same data points were presented at high physical positions, in contrast to the findings in Experiment 1 ($z = 6.80$, $p < .001$).

The odds ratio for the difference between conditions is 1.39 (95% CI [1.27, 1.53]). Participants were 1.39 times more likely to respond with a higher magnitude rating to data points presented

[4]Age data were unavailable for two participants.

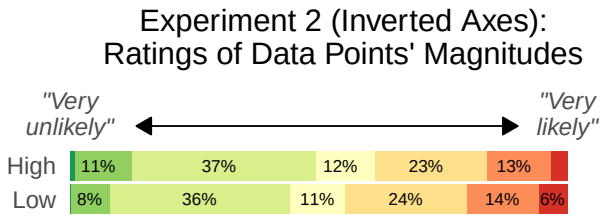


Figure 4.7: The distribution of Likert scale ratings of data points' magnitudes. The width of each response option represents the proportion of ratings recorded for that option. Note that data points presented at high physical positions (top) elicited a larger proportion of ratings on the left-hand side (representing *smaller* magnitudes), compared to data points at low physical positions (bottom), which elicited a larger proportion of ratings on the right-hand side (representing *larger* magnitudes).

at *low* positions than data points presented at high positions. This is equivalent to a Cohen's d value of 0.18, a very small effect size.

This model employed random intercepts for each scenario.

4.3.3.3.2 Severity Ratings

For ratings of the severity of consequences, a likelihood ratio test reveals that a model including physical position as a fixed effect explained numerically more variability in ratings than a model which does not include condition as a fixed effect, although this failed to reach significance: ($\chi^2(1) = 3.40$, $p = .065$). The severity of consequences was rated as numerically greater when data points representing the chance of an event occurring were presented at low physical positions, compared to when the same data points were presented at high physical positions, although this failed to reach significance ($z = 1.85$, $p = .064$).

The odds ratio for the difference between conditions is 1.13 (95% CI [0.99, 1.28]). Participants were 1.13 times more likely to respond with a higher severity rating to data points presented at low positions than data points presented at high positions. This is equivalent to a Cohen's d value of 0.07, a very small effect size.

This model employed random intercepts for each scenario, plus random intercepts and by-position slopes for each participant.

4.3.3.3.3 Data Visualisation Literacy

Adjusting for participants' data visualisation literacy scores did not change the pattern of results regarding ratings of the magnitude of data points themselves ($z = 6.83$, $p < .001$, odds ratio = 1.39, 95% CI [1.27, 1.53]) or the severity of consequences ($z = 1.85$, $p = .064$, odds ratio = 1.13, 95% CI [0.99, 1.28]).

4.3.3.4 Discussion

This experiment demonstrates that inferences about data points' magnitudes are influenced by their *relative* positions on a chart's axis, rather than their absolute physical positions. Viewing data in charts with inverted y-axes, participants rated the same values as greater when these values were plotted at *low* positions, compared to high positions. Ratings of the severity of consequences were not significantly affected by the position of data points representing the chance of negative events occurring. Accounting for differences in participants' data visualisation literacy did not alter the pattern of results.

In the previous experiment (E2), we did not observe a *significant* difference between magnitude ratings for data points at different physical positions in inverted charts. However, the pattern was consistent with the inferences based on data points' relative positions within axis limits. This experiment, with increased experimental power, provides evidence for statistically significant difference. Furthermore, the differences in estimated marginal means in inverted charts are almost identical across the two experiments (E2: r printnum(abs(r2_c_emm.hi-r2_c_emm.li)); E3: r printnum(abs(r3_c_emm.h-r3_c_emm.l))). In light of this new experiment, there is stronger evidence for the claim that axis limits influence interpretations of magnitude by determining data points' relative numerical positions.

4.4 General Discussion

Given the use of data visualisation for reliable communication of numerical information, understanding how design choices affect interpretations is an important matter. In a pair of experiments, we demonstrate that judgements of data points' magnitudes are influenced by a chart's axis limits. These experiments provide insight into the cognitive processes involved in assessing magnitudes in data visualisations.

We manipulated the axis limits accompanying plotted data, which affected the numerical context in which data appeared *and* the physical positions of data points.

Mention 2x2 here

Although a 2x2 experiment produced ambiguous results regarding interpretation of magnitude in charts with inverted y-axes, a followup

However, regardless of their physical positions, data points were associated with greater magnitudes when they appeared close to the end of the axis associated with higher values. Interpretation of the same absolute value is biased by its relative numerical position. This highlights viewers' sensitivity to surrounding information when assessing data. We illustrate that this framing effect occurs even when no contrasting data points are present to provide context: axis values are sufficient for informing magnitude judgements.

Our findings suggest that axis limits influence, but do not wholly *dictate*, impressions of magnitude. The distribution of magnitude judgements approximately followed the distribution of plotted numerical values, suggesting numerical values *also* contributed to magnitude judgements. In addition, the effect size associated with manipulating axis limits was larger for charts with conventional axis orientations, compared to charts with inverted y-axes. This suggests that the absolute physical position of data points partially contributed to participants' assessments. However, it is evident from the pattern of results for inverted charts that relative numerical position exerts a greater influence on magnitude judgements than absolute physical position. Whilst we cannot conclude that viewers interpret the axis range as the complete context for assessing plotted data, it is clear that axes inform magnitude judgements by defining plotted values' relative positions.

4.4.1 Relationship to Prior Work

The present data complement findings from research on y-axis truncation, which has observed that axis limits accompanying plotted values can influence viewers' impressions of those values. While previous investigations have shown that *comparisons* of plotted values are affected by y-axis limits (Correll et al., 2020; Pandey et al., 2015; Witt, 2019; Yang et al., 2021), the present findings show that they also influence *magnitude judgements*. This finding supports the notion that viewers are sensitive to visualisation rhetoric, which involves provoking a specific interpretation through a particular presentation of numerical information (Hullman and Diakopoulos, 2011).

A previous study addressing a similar question also concluded that a data point's location within a range of values affects interpretation of its magnitude (Sandman et al., 1994). The present study builds upon this research by identifying the mechanism behind this effect and removing the confound of variable axes ranges. It also extends the finding beyond a single scenario (asbestos) to a wider range of situations. By analysing different types of judgement separately, rather than using a combined measure, we verify that axis limits affect interpretations of the specific variable displayed in a chart.

In addition to the conceptual metaphor for magnitude, physical positions are also linked to *emotional valence* (where high positions are associated with positive valence). (Woodin et al., 2022) found that physical arrangements of data consistent with the conceptual metaphor for valence somewhat facilitate comprehension, but that associations between position and magnitude affect interpretations more strongly. Visualisations in the present experiments displayed data on negative events, so data were aligned with the conceptual metaphor for valence in inverted charts, and misaligned in conventional charts. Participants evidently did not use valence metaphors to interpret values in conventional charts; this would have produced the opposite pattern of results to those observed. The simplest explanation for our results suggests that participants relied on relative position when interpreting both conventional and inverted charts, rather than sometimes generating inferences based on a conceptual metaphor for valence.

4.4.2 Additional Findings

Prior research has observed positive correlations between participants' perceptions of event probability and outcome magnitude (Harris et al., 2009; Harris and Corner, 2011; Kupor and Laurin, 2020). We did not find robust evidence that assessments of the severity of consequences were affected by our manipulation of data points representing the chance of events occurring. However, whereas prior work substantially manipulated underlying scenarios, our subtler manipulation retained the same probability values, changing only the surrounding context. In addition, participants evaluated the severity of an event's *consequences*, which is one step removed from the property explored in prior research: the potency of the event itself. The effects of axis limits on interpretation of data about the incidence of events do not reliably extend to judgements about their consequences.

Adjusting for data visualisation literacy did not eliminate the influence of axis limits on interpretations. Yang et al. (2021) also observed that data visualisation literacy could not sufficiently explain variance in the degree of bias caused by y-axis truncation. This measure reflects comprehension of the conventions of data visualisation, indicating receipt of elementary instruction (Okan et al., 2016). Therefore, it is perhaps better suited for capturing viewers' application of basic knowledge in interpretation (Yang et al., 2021), whereas Ge et al.'s CALVI test (Ge et al., 2023) may be more appropriate for predicting susceptibility to differences in presentation format.

4.4.3 Limitations and Future Directions

We employed inverted y-axes solely for the purpose of distinguishing competing explanations; their use should not be considered an endorsement. To avoid misinterpretations, participants were given instruction on how to read inverted charts. With this explicit instruction, our data provide evidence *contrary* to the typical finding of misinterpretation resulting from associating higher positions with higher values (Pandey et al., 2015; Woodin et al., 2022). However, this instruction may have suppressed a spontaneous interpretation of magnitude, based on physical position, in favour of a learned interpretation. Our investigation therefore only explores the cognitive processing associated with assessing magnitude in charts which viewers know how to read.

This study was designed to explore one factor involved in assessing magnitude. The influence of axis limits on interpretations was relatively small, raising the question of how much this factor influences real-world decision-making and behaviour. A forced choice measure, or response scale with concrete values, would be suitable for capturing these outcomes in future work. Addressing this question will help to quantify how much a designer's choice of axis limits affects a viewer's choices and actions. An appreciable effect would have implications for visualisation design, suggesting use of axis limits which convey magnitude appropriately to avoid misleading users. Suitable axis limits cannot be objectively determined, but must be informed by the designer, based on their assessment of the data (Correll et al., 2020). The effects of axis range

on discrimination ability would also warrant consideration, taking account of the intended application.

4.4.4 Conclusion

We conducted two experiments investigating how axis limits inform interpretations of plotted values' magnitudes. Participants' subjective judgements were affected by data points' positions in relation to accompanying axis limits. The association between data points' positions and magnitude judgements critically depends on whether plotted data appear closer to the axis limit associated with higher or lower values. The cognitive processes associated with assessing magnitude in data visualisations involve taking into account the numerical context in which the data appear.

4.5 References

References

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**:255–278. doi:10.1016/j.jml.2012.11.001

Ben-Shachar M, Lüdecke D, Makowski D. 2020. Effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software* **5**:2815. doi:10.21105/joss.02815

Borges MA, Sawyers BK. 1974. Common verbal quantifiers: Usage and interpretation. *Journal of Experimental Psychology* **102**:335–338. doi:10.1037/h0036023

Brown GDA, Gardner J, Oswald AJ, Qian J. 2008. Does Wage Rank Affect Employees' Well-being? *Industrial Relations* **47**:355–389. doi:10.1111/j.1468-232X.2008.00525.x

Charalambides N. 2021. We recently went viral on TikTok - here's what we learned.

Christensen RHB. 2019. Ordinal—Regression Models for Ordinal Data.

Cleveland WS, Diaconis P, McGill R. 1982. Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. *Science* **216**:1138–1141. doi:10.1126/science.216.4550.1138

Correll M, Bertini E, Franconeri S. 2020. Truncating the Y-Axis: Threat or Menace?Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM. pp. 1–12. doi:10.1145/3313831.3376222

Garcia-Retamero R, Cokely ET, Ghazal S, Joeris A. 2016. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making* **36**:854–867. doi:10.1177/0272989X16655334

Gattis M. 2002. Structure mapping in spatial reasoning. *Cognitive Development* **17**:1157–1183. doi:10.1016/S0885-2014(02)00095-3

Ge LW, Cui Y, Kay M. 2023. CALVI: Critical Thinking Assessment for Literacy in VisualizationsProceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg Germany: ACM. pp. 1–18. doi:10.1145/3544548.3581406

Grice P. 1975. Logic and Conversation In: Cole P, Morgan JL, editors. Syntax and Semantics Vol.3: Speech Acts. New York: Academic Press. pp. 41–58.

Harris AJL, Corner A. 2011. Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**:1571–1578. doi:10.1037/a0024195

Harris AJL, Corner A, Hahn U. 2009. Estimating the probability of negative events. *Cognition* **110**:51–64. doi:10.1016/j.cognition.2008.10.006

Hullman J, Diakopoulos N. 2011. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics* **17**:2231–2240. doi:10.1109/TVCG.2011.255

Keren G, Teigen KH. 2001. The probability—outcome correspondence principle: A dispositional view of the interpretation of probability statements. *Memory and Cognition* **29**:1010–1021. doi:10.3758/BF03195763

Kupor D, Laurin K. 2020. Probable Cause: The Influence of Prior Probabilities on Forecasts and Perceptions of Magnitude. *Journal of Consumer Research* **46**:833–852. doi:10.1093/jcr/ucz025

Liddell TM, Kruschke JK. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* **79**:328–348. doi:10.1016/j.jesp.2018.08.009

Okan Y, Galesic M, Garcia-Retamero R. 2016. How People with Low and High Graph Literacy Process Health Graphs: Evidence from Eye-tracking: Graph Literacy and Health Graph Processing. *Journal of Behavioral Decision Making* **29**:271–294. doi:10.1002/bdm.1891

Okan Y, Garcia-Retamero R, Galesic M, Cokely ET. 2012. When Higher Bars Are Not Larger Quantities: On Individual Differences in the Use of Spatial Information in Graph Comprehension. *Spatial Cognition and Computation* **12**:195–218. doi:10.1080/13875868.2012.659302

Pandey AV, Rall K, Satterthwaite ML, Nov O, Bertini E. 2015. How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion TechniquesProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. Seoul, Republic of Korea: ACM Press. pp. 1469–1478. doi:10.1145/2702123.2702608

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* **51**:195–203. doi:10.3758/s13428-018-01193-y

R Core Team. 2022. R: A Language and Environment for Statistical Computing.

Sandman PM, Weinstein ND, Miller P. 1994. High Risk or Low: How Location on a "Risk Ladder" Affects Perceived Risk. *Risk Analysis* **14**:35–45. doi:10.1111/j.1539-6924.1994.tb00026.x

Sung Y-T, Wu J-S. 2018. The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods* **50**:1694–1715. doi:10.3758/s13428-018-1041-8

Taylor BG, Anderson LK. 1986. Misleading Graphs: Guidelines for the Accountant. *Journal of Accountancy* **162**:126–135.

Tversky A, Kahneman D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* **185**:1124–1131. doi:10.1126/science.185.4157.1124

Tversky B. 1997. Cognitive Principles of Graphic DisplaysProceedings of the AAAI 1997 Fall Symposium on Reasoning with Diagrammatic Representations. Menlo Park, CA: AAAI Press,. pp. 116–124.

Voeten CC. 2022. Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects.

Witt JK. 2019. Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis. *Meta-Psychology* **2**:1–20. doi:10.15626/MP.2018.895

Woodin G, Winter B. 2018. Placing Abstract Concepts in Space: Quantity, Time and Emotional Valence. *Frontiers in Psychology* **9**:1–14. doi:10.3389/fpsyg.2018.02169

Woodin G, Winter B, Padilla L. 2022. Conceptual Metaphor and Graphical Convention Influence the Interpretation of Line Graphs. *IEEE Transactions on Visualization and Computer Graphics* **28**:1209–1221. doi:10.1109/TVCG.2021.3088343

Yang BW, Vargas Restrepo C, Stanley ML, Marsh EJ. 2021. Truncating Bar Graphs Persistently Misleads Viewers. *Journal of Applied Research in Memory and Cognition* S2211368120300978. doi:10.1016/j.jarmac.2020.10.002

Postscript to Chapter 3

Overview

The first study in this thesis was designed to examine whether axis range influences interpretations of absolute magnitude in data visualisations. Axis range was initially identified as a plausible factor influencing magnitude judgements because it dictates the physical positions of data points, which have conventional associations with magnitude (e.g., higher positions are associated with higher values). However, this study revealed that the principal way axis limits influence interpretations is by dictating the *relative positions* of data points. Thus, an axis provides a frame of reference which informs subjective judgements of magnitude.

Rationale for Experimental Design

Investigating the influence of axis range on magnitude judgements required appropriately designed experimental stimuli. It was important to employ a suitable visualisation format for the axis manipulations in this study. Dot plots, which were used throughout, employ a position-only encoding. This means that, unlike in bar charts, the encoding is not distorted by axes that do not start at zero. This afforded greater freedom when generating datasets and choosing axis limits. The visualisations could present a wide range of values and also maintain the same difference between the upper and lower limits. Furthermore, unlike bar charts, the physical characteristics of the data points themselves did not change due to the manipulation or inversion of axes, preventing unwanted confounds. The minimalistic design contributes to tight experimental control.

In seeking to understand how axis ranges influence cognitive processing of absolute magnitude, this study adopted some necessarily contrived design choices. In particular, the use of singular probability point estimates does not conform to the recommendations for visualising probabilistic information, which involve explicitly encoding uncertainty and displaying discrete outcomes (Fernandes et al. 2018; Kay et al. 2016). In addition, the use of inverted axes violates conventions and is associated with difficulties in comprehension (Pandey et al., 2015, Woodin et al., 2021). However, there is no evidence to suggest that these design choices significantly hindered comprehension in this study. Indeed, participants' responses to inverted axes were consistent with the understanding that higher values were presented at lower physical positions. This behaviour would not be expected if these visualisations resulted in fundamental

misunderstandings. As required for systematic experimental study, stimuli were developed in order to facilitate investigation of underlying cognitive mechanisms.

This study also required careful consideration of participants' task. Participants were asked to consider various risks, with the chance of negative outcomes presented as percentage probabilities. This topic was chosen to emulate the type of data displayed in everyday settings where absolute magnitude would be a relevant concern. The task was not designed to assess the comprehension of specific numerical values. Instead, the task was designed to interrogate *message*-level interpretations, pertaining to the prevailing impressions resulting from the visualisations. Thus, participants were instructed to consider *overall* magnitude (and severity of consequences), and respond with a subjective judgement, rather than a precise numerical estimate. Collecting these measures is important, because the *gist* of information, rather than *verbatim* knowledge, typically informs behaviour (Reyna and Brainerd, 2008).

Summary of Findings

This work demonstrated that axis limits contribute to interpretations of absolute magnitude in data visualisations. When different axis limits accompanied the same data points, magnitude judgements were affected in a subtle yet systematic manner. Participants generated inferences about magnitude by using axis limits as reference points. Despite an inconclusive 2x2 experiment, a high-powered experiment involving inverted charts revealed that magnitude ratings were higher when data points were numerically high within the range of axis values. Severity ratings were not consistently affected by axis limits. It is also important to recognise that interpretations of magnitude were not *dictated* by the numerical context provided by axes. That is, magnitude ratings were influenced by the absolute values plotted, not just their relative positions.

Prior research on axis limits has largely focused on interpretation of *relative magnitude*: the magnitude of *difference between plotted values* (Driessen, Correll, Yang, Witt, Pandey, Stone et al., 2003). However, the first study in this thesis provides evidence that axis limits also affect interpretations of *absolute magnitude*: the magnitude of *the values themselves*. Consequently, it highlights an additional consideration for designers when choosing axis limits. These results accord with prior findings on this topic, which found that the upper and lower limits in risk ladder visualisations affected impressions of risk (Sandman et al., 1994). The first study in this thesis used a range of scenarios, specific measures, and highly-controlled stimuli, thus building upon Sandman et al's (1994) study to construct a robust evidence base.

Preface to Chapter 4

The overall objective of this thesis is to explore the role of axis limits in the cognitive processing of absolute magnitude in data visualisations. My aim is to address a general cognitive mechanism, therefore experiments should not be limited to one type of data or visualisation format. Rigorous evidence requires examination of different stimuli. Therefore, the second investigation in this thesis explores interpretations of magnitude in choropleth maps.

Choropleth maps are a type of data visualisation conveying numerical data associated with discrete geographical areas. For example, they may be used to display literacy rates or the prevalence of an infection in different regions. Values for each region are encoded using colour, often with darker colours symbolising higher values. This contrasts with other popular visualisation formats where numerical values are encoded through geometric attributes. For example, scatterplots, bar charts, and pie charts, which employ position, length, and angle, respectively.

A colour legend often accompanies choropleth maps, displaying the mapping between colours and data points. This serves the same purpose as a typical numerical axis (e.g., in a bar chart), showing how visual features correspond to numerical values. Therefore, a colour legend's limits are functionally equivalent to axis limits. However, one important difference between a colour legend and a typical axis makes the choropleth map a valuable instrument in this investigation into cognitive processing. Crucially, it is possible to manipulate a colour legend's limits without changing the appearance of plotted values in the map or the correspondence between colours and numerical values. This contrasts with the manipulation of the axes of dot plots in the first study in this thesis, where manipulating axis limits necessarily altered the appearance of plotted values. Therefore, the second study investigates whether interpretations of magnitude are informed by a colour legend's limits alone, in the absence of any change to the appearance of plotted values.

5 Choropleth Maps Can Convey Absolute Magnitude Through the Range of the Accompanying Colour Legend

Data visualisation software provides the ability to create highly customizable choropleth maps. This presents an abundance of design choices. The colour legend, one particular aspect of choropleth map design, has the potential to effectively convey data points' absolute magnitudes (how large or small they are). Colour legends present the mapping between a specific range of colours and a specific range of numerical values. In this experiment, we demonstrate that manipulating this range affects interpretations of the plotted values' absolute magnitudes. Participants (N = 100) judged the urgency of addressing pollution levels as greater when the colour legend's upper bound was equal to the maximum plotted value, compared to when it was significantly larger than the maximum plotted value. This provides insight into the cognitive processing of plotted data in choropleth maps that are designed to promote inferences about overall magnitude.

5.1 Introduction

To make sense of statistics presented in newspaper articles or scientific reports, it is often important to interpret their meaning in context. This may involve determining whether the presented values represent large or small numbers. Data visualisations are often used to convey statistics, so understanding how these tools may communicate data points' magnitudes is crucial.

Numerical values in choropleth maps are often encoded using the entire range of the chosen colour palette, in order to aid discrimination and facilitate identification of spatial patterns. Thus, the range of values in the accompanying colour legend typically consists of only those values which were observed. However, this is not the only application for a choropleth map. In certain cases, displaying values' *absolute* magnitudes may be considered more pertinent than displaying their *relative* magnitudes. This would allow a viewer to gauge, on the whole, how large or small presented values are, in context. To communicate this, the range of values in the accompanying colour legend may include values which were not observed but remain relevant nonetheless. Designers may wish to sacrifice discrimination ability for an overt display of magnitude, in order to convey their intended message.

Indeed, choropleth maps displaying overall magnitudes have been used in practice. Figure 5.1 depicts data concerning public support for a federal ban on abortion in the U.S. The accompanying colour legend presents the entire range of possible values: from 0% to 100% support. Since plotted values do not exceed 30%, their magnitudes appear small, in context. In addition, whereas a typical colour scale would amplify differences between regions, this design presents variability between states as low. This lends credibility to the notion that, for this aspect of a divisive issue, public support is consistently low across the U.S.

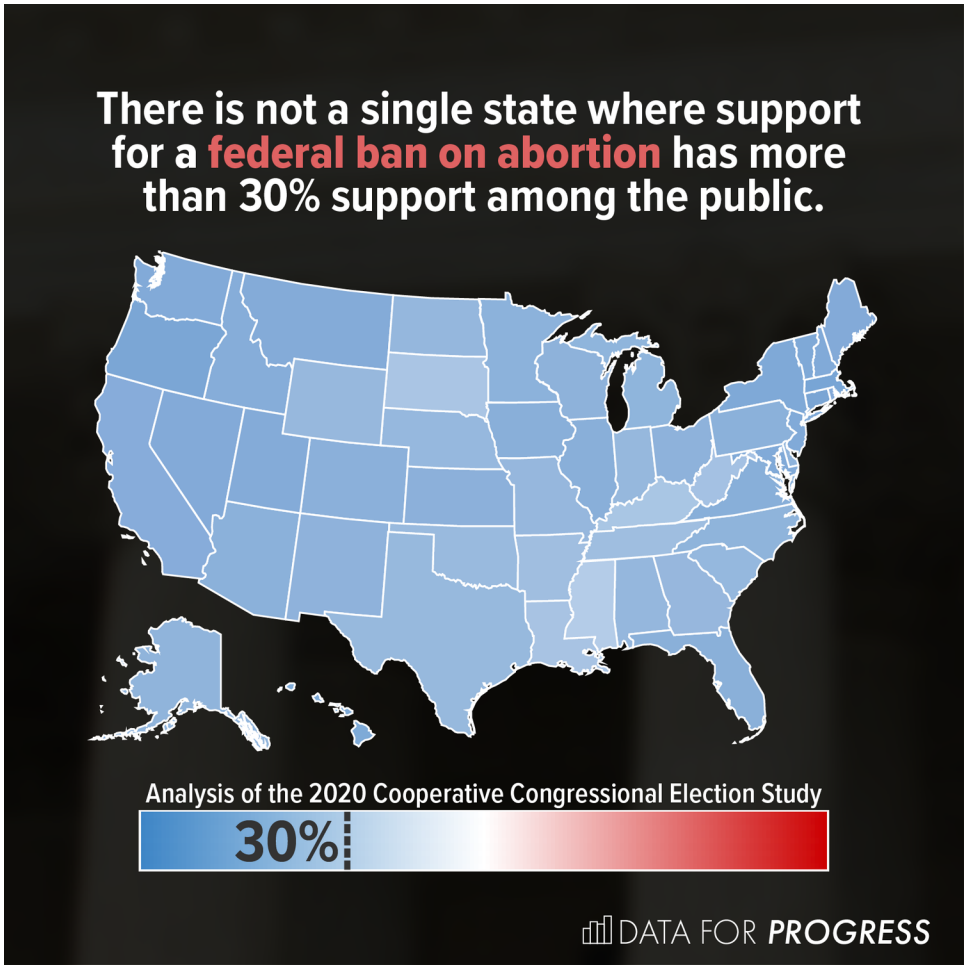


Figure 5.1: A choropleth map displaying data from an analysis of state-level public support for a federal ban on abortion in the U.S (Fischer and Ali, 2021). The colour legend employs a diverging blue-red colour palette, with white in the centre, showing the full range of possible values. The 30% point is marked with a dotted line and labelled to indicate that no state exceeds this level of support. Reproduced with permission.

The map may appear homogeneous, but choropleth maps present opportunities for conveying

information *beyond* relative geographical differences, just as line charts may show stagnant wages. By presenting a wider numerical context, the accompanying legend imbues the map with meaning, illustrating low variability and small magnitudes. The simplicity of this message does not preclude its visualisation; as well as illuminating complex patterns, data visualisations are also designed to improve retention and engagement (Bertini et al., 2020), and support cognition (Hegarty, 2011).

This paper explores cognitive processing of overall magnitude in choropleth maps. Through an empirical study, we demonstrate that colour legends, which depict the mapping between colours and numerical values, can imply how large or small plotted values' absolute magnitudes are. Even when the mapping between colour and numerical value remains the same, the range of the colour legend provides a crucial source of context. The relationship between this range and the plotted data influences viewers' interpretations of magnitude.

5.2 Related Work

5.2.1 Choropleth Maps

Choropleth maps are thematic maps which employ colour to symbolise numerical values, conveying quantitative data in a spatial manner. Choropleth mapping uses datasets where each data point corresponds to a discrete area, typically defined by administrative boundaries (e.g., national or local government regions). Ratios, proportions and averages are plotted to enable appropriate comparisons between regions (Dent et al., 2009).

Dent et al. (2009) discuss several considerations for choropleth map design, including data preprocessing, spatial resolution, and appropriate accompanying text. However, data classification is a particularly prominent theme in guidance on choropleth mapping. To better convey patterns in the spatial distribution of data, values can be classified into discrete classes (Kraak and Ormeling, 2013). Decisions around classification involve trade-offs between clarity of patterns in the map and clarity of the legend. Natural Breaks methods [e.g. Jenks optimization; Jenks and Caspall (1971)] identify class boundaries according to the distribution of data, ensuring clusters of similar values appear homogeneous. The Equal Frequency method ensures uniform prevalence of each class within the map, whereas the Equal Interval method simply employs the same numerical range for each class (Dent et al., 2009). Unclassed choropleth maps (Tobler, 2010), which do not employ discrete groups at all, are an alternative option. Legends are not divided into classes, meaning each unique value is represented distinctly. This may increase estimation error, yet avoids the impression that similar values either side of a class boundary are substantially different (Kraak and Ormeling, 2013).

Regarding the minimum and maximum values used in the labels for each class, two options are available. Continuous class ranges include non-observed values to create a continuous sequence of numbers. This provides consistency when re-using a legend for multiple maps. However, this may increase the chance that viewers make imprecise estimations of specific values, compared

to non-continuous class ranges which include only observed values (Dent et al., 2009). The use of open-ended categories at a legend's extremes (Paul, 1993) is an additional consideration, generating a similar generalisability-precision trade-off.

Dykes et al. (2010) explored several creative approaches to map legend design, providing alternatives to conventional implementations. One such design, displaying statistical information within a legend, has been implemented in several forms, for communicating distributions (Cromley and Ye, 2006; Kumar, 2004) and uncertainty (Retchless and Brewer, 2016) in choropleth maps. Several studies have illustrated the influence of legend design on cognitive processing for a range of maps. Proximity between icons and corresponding text within a legend was found to be the most influential aspect of spacing on visual search (Li and Qin, 2014). For thematic maps showing several geographical features, overall task performance was found to be similar across three different legend arrangements (list legend, grouped legend, natural legend), but user preferences depended on legends' suitability for specific tasks (Gołębiowska, 2015). Consistent with left-hemisphere specialised language processing, legends presented on the right of a map were processed faster than those presented on the left (Edler et al., 2020). Eye-movement tracking has revealed that fixation on map legends decreases with repeated exposure, illustrating the role of legends for developing initial cognitive representations (Hepburn et al., 2021). This body of research provides evidence that legend design influences various aspects of map interpretation.

5.2.2 Communicating Absolute Magnitude Through Data Visualisation

Empirical studies in various scientific fields have explored how interpretations of magnitude are influenced by data visualisation design choices.

Recently, the practice of y-axis truncation has enjoyed attention in experiments at the intersection of the disciplines of data visualisation and psychology. Y-axis truncation refers to the practice of minimising the range of values that appear on the y-axis. This typically involves starting the y-axis at a value greater than zero (Correll et al., 2020). However, some experiments on y-axis truncation have employed axes that are roughly symmetrical about the plotted data (Witt, 2019). Truncation effects are therefore not just associated with the exclusion of a zero value, but also the exclusion of values *above* the observed data, which make differences appear smaller. Thus, more generally, truncation effects illustrate people's treatment of axes as implicit scales for making qualitative judgements about presented data.

Research on the effects of y-axis truncation has focused on how this practice can alter people's interpretations of the magnitude of the difference between plotted values. Demonstrating the effect of y-axis truncation with a large online sample, Pandey et al. (2015) found that ratings of the magnitude of the difference between values were greater when a truncated axis was used to display the difference between safe drinking water levels in two towns. In both bar charts and line charts, increasing the degree of truncation produces increasing estimations of the severity of the difference between values (Correll et al., 2020). Encouraging careful attention

to plotted data (by ensuring that numerical values are read precisely) does not eliminate this effect (Correll et al., 2020). Warnings somewhat reduce, but do not eradicate, the difference between interpretations of truncated and non-truncated charts (Yang et al., 2021). Visual indicators of truncation are also ineffective (Correll et al., 2020).

Witt (2019) demonstrated that using the widest possible y-axis range diminishes a viewer's sensitivity, which is the ability to distinguish between different degrees of separation between values. On the other hand, using the smallest possible y-axis range increases bias in interpretation (i.e., the extent to which judgements of the magnitude of difference deviate from actual effect sizes). To maximise sensitivity and minimise bias, and to ensure correspondence between the appearance of the difference and the reality, Witt suggests using a range of 1-2 standard deviations for y-axis limits.

Witt's (2019) recommendations are prescribed for disciplines which use standardised effect sizes (e.g., Cohen's d) in the reporting of data and statistics. Correll et al. (2020) provide more general advice relevant to those in all disciplines: the appearance of differences in a visualisation should be appropriate for the specific data. Therefore the decision whether or not to truncate an axis depends on the real-world magnitude of the difference, and ultimately designers should ensure they represent this faithfully. Evidence suggests that viewers interpret the axis range as a representation of the relevant numerical context within which plotted data should be assessed. When an axis only just contains a pair of values, they will generally be considered to be highly divergent. When an axis easily contains these values, they will generally be considered similar, because the difference between values will be dwarfed by the vastness of the scale. Arbitrary rules will not absolve a chart designer's responsibility to consider what their visualisation implies (Correll et al., 2020).

As Yang et al. (2021) discuss, one explanation for these effects draws on Grice's co-operative principle (Grice, 1975). This theory, originally concerning linguistic utterances, would suggest that components of a chart, such as axes, will be considered to communicate relevant information about plotted data. Thus, a viewer will derive a designer's intended message from the features of the visualisation. Changing one's interpretation of magnitude in accordance with changes to axis range could therefore be considered a coherent response.

Research on risk communication has also explored how visualisation design choices affect interpretations of presented information. A set of experiments relevant to the present investigation originated with empirical data which suggested that icon arrays were more effective than text at promoting risk-averse behaviour (Stone et al., 1997). Further research (Stone et al., 2003) suggested that this occurred because the data visualisations only displayed the number of people affected by the negative outcome. Therefore, unlike the text, the icon arrays made the numerator more salient than the denominator (the total number of people in the sample). This was demonstrated empirically in the same study, using bar charts: the difference between numerators (15 vs. 30) appeared much bigger when the larger numerator (30) was used for the upper axis limits, compared to when the denominator (5000) was used for the upper axis limits. Risk reduction (the degree of difference between plotted values) was perceived as smaller when bar charts were extended to incorporate the denominator. Unlike studies on y-axis truncation

(Correll et al., 2020; Driessen et al., 2022; Pandey et al., 2015; Witt, 2019; Yang et al., 2021), the lower axis limit was not manipulated, and remained fixed at zero. This pattern of results has been replicated using icon arrays (Garcia-Retamero and Galesic, 2010) and pie charts (Hu et al., 2014), and a similar effect has been reported for line charts (Taylor and Anderson, 1986) suggesting this phenomenon is driven by a common mechanism independent of chart type.

Stone et al.'s (2003) experiment demonstrated that extending the upper limit caused participants to interpret the *difference between* values as smaller. Unfortunately, the design of this experiment leaves uncertainty as to whether this extension affected interpretations of the magnitude of *the values themselves*, because participants only compared risks between charts in the same condition, not across conditions. However, this issue was addressed by Okan et al. (2020), who found that icon arrays which *did not* display the denominator increased perceived risk relative to those which did (with larger increases at smaller probabilities). Including the denominator also resulted in more accurate estimates of the underlying risk probabilities. This accords with the finding that the apparent magnitude of risk decreases when the upper limit is extended in a risk ladder visualisation (Sandman et al., 1994). This implies that interpretations of magnitude are informed, in part, by the data point's position within the risk ladder's limits.

5.2.3 Colour Legends

In data visualisations employing geometric encodings (e.g., position, extent), axes are the dimensions along which data are plotted. In colourmap visualisations, a different type of axis is present, which is not used to display data directly, but presents the mapping between colours and numerical values, henceforth referred to as a 'colour legend'. Default settings in popular visualisation tools, such as ggplot2 (Wickham, 2016) and Matplotlib (Hunter, 2007) tend to employ colour legends which use the minimum and maximum values in the data at their extremes. Thus, the potential for values smaller than the minimum, or larger than the maximum, is not encoded by these colour legends. This facilitates comparison between values, since using a wide range of colours improves discrimination ability. Crucially, however, it does not facilitate magnitude judgements. Consider, for example, a heatmap showing profits for each quarter over the course of five years. Using the darkest colour on the colour legend to represent the highest profits could conceal the fact that profits in general have been poor for the entirety of this period, because the colour legend is agnostic towards real-world magnitude.

Research involving colour legends has often focused on assessing the appropriateness of different colour scales and capturing colour discriminability through colour difference models. Harrower and Brewer (2003) developed a tool for selecting suitable colour scales for particular forms of data: sequential scales for ordinal or numerical data, qualitative scales for categorical data, and diverging scales for highlighting midpoints. Using choropleth maps, Brychtova and Coltekin (2015) determined the minimum colour distance required for reliably detecting differences between two regions. Other work has identified specific features which make for an effective colour scheme, from low-level properties such as uniform luminance [Dasgupta et al. (2020)]

to high-level properties such as consistency with semantic colour associations (Lin et al., 2013). Researchers have also modelled the impact of mark size on colour discriminability (Stone et al., 2014) and demonstrated adaptation of colour difference models to specific viewing conditions (Szafir et al., 2014).

Choropleth maps are one of several types of colourmap visualisation which map colour to numerical data (see also, heatmaps and neuroimaging visualisations). Schiewe (2019) illustrates that impressions of quantity are positively associated with the proportion of a choropleth map occupied by darker colours. The size of geographical regions and the *classification* of values can both influence the extent to which a map displays colours on the darker end of the chosen colour scale, which impacts judgements of presented data. Whilst this study manipulated the appearance of plotted data in maps, other research has held the appearance of plotted data constant in order to study how the context surrounding a colour legend affects viewers' inferences. Schloss et al. (2019) observed that viewers' spontaneous interpretations of the relationship between colour and quantity can depend on which background colour is used. Their experiment attempted to reconcile contrasting theories about which aspects of a colour stimulus are associated with greater quantities ('dark-is-more'; 'contrast-is-more'; 'opaque-is-more'). They found that viewers associate darker colours with greater quantities when there is no apparent variation in the colour scale's opacity. However, when the colour scale does appear to have varying degrees of opacity, an 'opaque-is-more' association prevails. For example, black-white colour scales appear to have low opacity against a blue background (so lighter greys are more readily associated with smaller quantities), but high opacity against a black background (so lighter greys are more readily associated with larger quantities).

Different interpretations of the same dataset can also arise through modified displays of the same colour scale. Empirical research has compared colour legends which only indicate uncertainty using colour features (e.g., increasing luminance and decreasing saturation), to colour legends which also signal uncertainty through increasing reduction in the range of possible colours, termed Value-Suppressing Uncertainty Palettes (VSUPs, Correll et al. (2018)). In Correll et al.'s study, participants played a 'Battleship' style game which involved reducing risk by balancing danger and uncertainty. Participants were more likely to favour riskier but more certain options over uncertain options when using VSUPs. Constraining the range of colours at higher uncertainty levels may have reduced the impression that these data points could represent desirable low-danger magnitudes. The experiment we report below examines directly how the range of values in a colour legend affects interpretations of magnitude.

5.3 Methodology

5.3.1 Outline

The present experiment investigates the influence of colour legend range on the cognitive processing of magnitude. We manipulated the colour legend's upper bound, such that it was

equal to the maximum plotted value (*truncated range*) or it was equal to double the maximum plotted value (*extended range*). We employ the term 'truncated' in a broad sense, referring to a scale that is constrained such that potentially relevant values are omitted, not simply a scale that excludes a zero value. Using a lower bound of zero reduced the number of differences between the two conditions, so that only the upper bound was manipulated. This also meant that plotted values' variability appeared smaller, assisting participants in judging the *overall* magnitude of these values. For each item, the colour palette, geographic regions, and the mapping between colours and numerical values, were identical across conditions. Therefore, the only difference between versions of a given item was the range of the colour legend: the map itself remained unchanged.

Rather than asking participants to make abstract judgements about the size of abstract values, we presented fictitious pollution data, and asked how urgently action should be taken to address the pollution levels displayed in each data visualisation. This captures participants' assessments of magnitude through the type of judgements which can drive behaviour. In addition to increased ecological validity, we also anticipated that pollution data might be able to generate a balanced set of responses to the question of urgency. A variable evoking an extreme negative reaction may have elicited responses at ceiling and one too trivial may have elicited responses at floor. We expected participants to recognize that a sufficient degree of pollution would require action, but also understand that low levels may require less urgent action. We did not provide a specific definition of urgency for participants to use when making their responses. Therefore, different participants' responses may reflect different notions of urgency. However, the within-participants design accounts for individual variation. Each participant's ratings are compared against their own ratings for the alternative condition, allowing for meaningful comparison between conditions.

Pollution levels were displayed in choropleth maps, which use colour encoding to display data aggregated at the level of geographic areas. Note that we do not consider the designs of choropleth maps in this experiment to reflect best practice for plotting pollution statistics. Rather, these designs were motivated by the desire to examine the role of colour legends in the interpretation of magnitude. Previous research has illustrated that the size of geographical regions can influence ensemble coding in choropleth maps (Schiewe, 2019). However, we did not control for this aspect, instead we prioritised ecological validity by using maps with real geographical regions. These maps appeared identical across conditions in order to avoid this bias confounding results.

To control for the possibility that participants used the colour legend's numerical labels, rather than the range of values displayed, as a reference for their magnitude judgements, we omitted the colour legend's numerical labels in half of trials. This allowed us to test whether the presence of numerical labels affected the degree to which magnitude judgements were influenced by the colour legend's upper bound.

5.3.2 Pre-Registration

We predicted that urgency ratings would be higher for truncated legends, compared to extended legends. In addition, we planned to compare whether any difference between these two conditions was moderated by the presence or absence of numerical labels, but made no predictions about existence or direction of any main effect or interaction. Participants completed Garcia-Retamero et al.'s (Garcia-Retamero et al., 2016) Subjective Graph Literacy scale, therefore we also planned to test whether any observed effects (or lack of) could be explained by differences in data visualisation literacy. This five-item scale is a quick, reliable measure that is correlated with scores on Galesic and Garcia-Retamero's (Galesic and Garcia-Retamero, 2011) test-based measure of data visualisation literacy. The pre-registration, plus materials, experiment script, data and analysis code are available at https://osf.io/qe9hf/?view_only=32c420d6ef6c45b1ae2d3dc42dc6fe69. This repository contains the requisite resources to generate a fully-reproducible version of this paper.

5.3.3 Design

In each trial, we independently manipulated two aspects of the choropleth map. When the colour legend had a *truncated range*, its upper bound was equal to the maximum value displayed in the map. When the colour legend had an *extended range*, its upper bound was equal to double the maximum value (and the maximum value displayed in the map appeared at the legend's halfway point). Numerical labels on the colour legend were either *present* or *absent*. This resulted in four unique combinations of conditions. We employed a Latin-squared design, ensuring that each participant was exposed to each combination of conditions throughout the experiment, but only saw one combination for each given map. There were a total of 54 trials (48 experimental trials, six attention check trials). Example stimuli are shown in Figure 5.2.

5.3.4 Participants

We recruited participants using prolific.co. The experiment was advertised to users with English language fluency, normal or corrected-to-normal vision, and no experience of colour deficiency, who had previously participated in more than 100 studies on Prolific. Participants were paid £3.50. Ethical approval was granted by The University of Manchester's Division of Neuroscience and Experimental Psychology Ethics Committee (Ref. 2022-11115-23778).

In our pre-registration, we planned to exclude participants who failed more than one attention check question, in order to exclude those who were not sufficiently engaged in the task. However, when many more participants than expected failed more than one attention check question, this criteria was deemed too stringent and we instead awarded payment to all participants who returned data, regardless of their responses to attention check questions. Consequently, due to practical constraints, we were unable to obtain a sample which met our originally-specified sample size (N = 160) and our pre-registered inclusion criteria. Therefore,

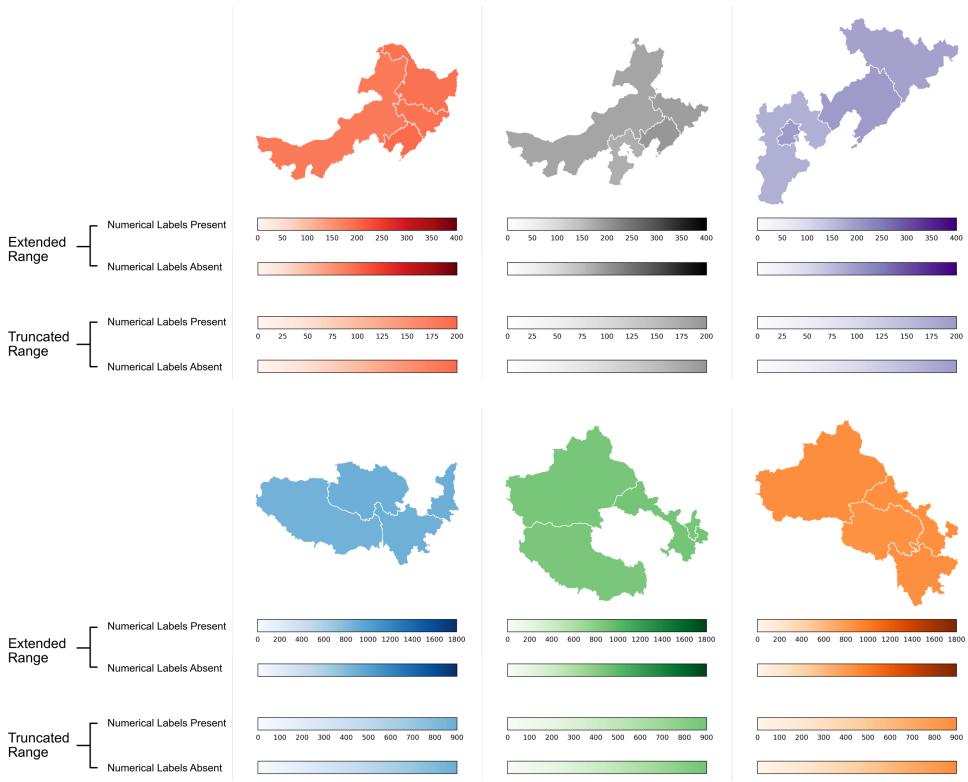


Figure 5.2: Example stimuli: six choropleth maps showing fictitious pollution data. Four colour legends are displayed below each map, but only one colour legend accompanied the map in each trial. Colour legends with extended ranges have a maximum value equal to double the maximum plotted value (top row: 400; bottom row: 1800). Colour legends with truncated ranges have a maximum value equal to the maximum plotted value in the map (top row: 200; bottom row: 900). During the experiment, all six colour scales were used in conjunction with all maximum values.

we terminated data collection once the sample of those who satisfied the attention check criteria was balanced across all four Latin-squaring lists (N = 100; 25 participants per list). We used this sample for our main analysis. As a compromise for the reduction in experimental power, we also demonstrate below that the pattern of effects is largely the same when analysing the entire dataset (those who satisfied attention check criteria and those who did not; N = 165). In Section 5, we discuss a possible reason for the higher-than-expected rate of incorrect responses to attention check questions. Demographic information is shown in Table 5.1.

Table 5.1: Demographic Information

	Gender			Age		Graph Literacy		Education
Sample	Male (%)	Female (%)	Prefer not to say (%)	Mean	SD	Mean	SD	High School or Above (%)
N = 100	59.0	40.0	1.0	30.8	8.8	21.6	4.5	98.0
N = 165	53.9	45.5	0.6	31.8	10.1	21.8	4.5	98.8

5.3.5 Procedure

The experiment was programmed using PsychoPy (Peirce et al., 2019, version 2022.1.4) and hosted on pavlovia.org. A link to an interactive version of this experiment has been excluded from this manuscript for anonymization purposes. Participants were instructed to use laptop or desktop computers, rather than another type of device and were told that the experiment was about using information to make decisions. We did not calibrate or measure colour display on participants' own screens, but using a within-participants design prevents this from influencing our results. Each participant was exposed to both experimental conditions under the same display conditions. Participants were informed that in each map, each region's colour reflected its pollution level, and that data on different types of pollution were shown throughout the experiment, with pollution levels presented using standardised units.

In every experimental trial, the text above the map read '*This map shows the levels of a certain type of pollution, in four regions*'. Participants were advised to read the question, which was presented below the map: '*How urgently should pollution levels in these regions be addressed?*' This question was used in all experimental trials, where the left anchor on the visual analogue response scale was labelled '*Not very urgently*' and the right anchor was labelled '*Very urgently*'. The instructions stated that higher pollution levels need to be addressed more urgently than

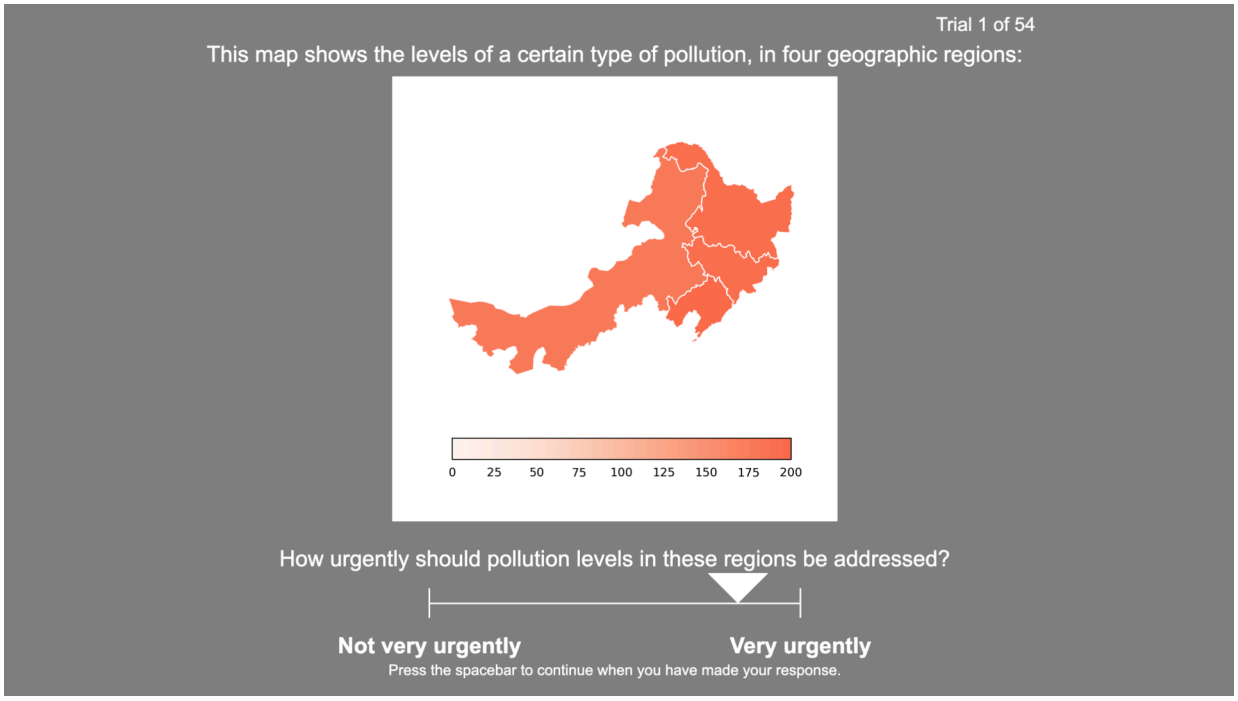


Figure 5.3: An example of a single experiment trial, showing a choropleth map with a truncated colour legend, plus a response marker on the visual analogue scale.

lower pollution levels. Participants were permitted to move the response scale marker as many times as they wished before continuing to the next trial. An example trial is shown in Figure 5.3.

Attention check items resembled normal trials except for the text displayed. Participants were asked to move the marker to one of three locations: 'to the middle of the scale', 'all the way to the *'Not very urgently'* end of the scale' or 'all the way to the *'Very urgently'* end of the scale'. In experimental trials, response scale granularity was set to 0, which permitted participants to place the marker at any location along the response scale. In attention check trials, response scale granularity was set to 0.5, so participants were only permitted to place the marker at one of three locations specified in the question: the leftmost point, the centre of the scale, or the rightmost point.

Following the final trial, participants were informed that both the data presented, and the standardised units used, were fictitious. Finally, participants were presented with a text box and the prompt '*What strategies did you use during the study? Do you have any comments about the study? (optional)*'. Average completion time was 13.57 minutes (SD = 6.24 minutes) for those who satisfied the pre-registered attention check criteria and 12.56 minutes (SD = 6.20 minutes) for the full sample.

5.3.6 Materials

Materials were generated using Python (version 3.9.12). Matplotlib (version 3.5.1) was used to generate colour legends and geoplot (version 0.5.1) was used for plotting geospatial data.

Each visualisation contained a unique combination of four neighbouring Chinese provinces (except the six attention check items, which employed six existing combinations used in the experimental items). China was chosen to reduce the potential impact of prior knowledge, as Prolific's participants tend to be located outside China. However, the choice of country was not disclosed to participants and regions were not labelled. The pollution data used were entirely fictitious, as were the 'standardised units' used to present the data.

The maximum value in the plotted data ranged from 200 to 900 (in multiples of 100), and the values for the other three provinces were between 10 and 30 units below this maximum value. Six Matplotlib colour scales ('Reds', 'Greys', 'Purples', 'Blues', 'Greens', 'Oranges') were each used once per maximum value. These scales exhibited monotonic and approximately linear variation in lightness (L^*). Monochromatic sequential scales were used for simplicity, avoiding additional differences between conditions, such as the relative amounts of different hues (multi-hue scales) or midpoints' positions (diverging scales). Table 5.2 shows the start and end colours in CIEL*a*b* space, using CIE standard illuminant D65.

For each item, a 'mappable' object defined the mapping between numerical values and colours for both truncated and extended colour legends. The lightest colour in the scale was mapped to zero and the darkest colour to double the maximum value. This range was employed in

Table 5.2: CIELab Values for Colour Legends' Start and End Colours

		Start Colour			End Colour		
Colour Scale	Range	L*	a*	b*	L*	a*	b*
Reds	Truncated	97.17	2.50	3.58	62.89	53.29	45.11
Reds	Extended	97.17	2.50	3.58	19.69	41.49	24.73
Greys	Truncated	100.00	0.01	-0.01	62.31	0.01	-0.01
Greys	Extended	100.00	0.01	-0.01	0.00	0.00	0.00
Purples	Truncated	98.74	0.69	-0.85	65.59	11.10	-22.97
Purples	Extended	98.74	0.69	-0.85	18.09	49.37	-54.13
Blues	Truncated	98.43	-0.59	-2.38	68.37	-10.35	-26.64
Blues	Extended	98.43	-0.59	-2.38	20.93	11.95	-38.06
Greens	Truncated	98.43	-2.87	2.76	72.70	-40.34	31.54
Greens	Extended	98.43	-2.87	2.76	24.36	-30.10	19.31
Oranges	Truncated	97.05	1.68	5.99	69.77	36.47	59.19
Oranges	Extended	97.05	1.68	5.99	29.34	36.61	39.62

the extended colour legend. The truncated colour legend, on the other hand, terminated at the maximum value in the data, so the range was halved (but the mapping between numerical values and colours was retained). No classification was employed in the legends, for maximum consistency across conditions. Where numerical labels were present, an identical number of labels (between six and ten) appeared on both versions of a colour legend. Tick marks were absent from all colour legends.

5.4 Analysis

5.4.1 Analysis Methods

Analysis was conducted in R (R Core Team, 2022, version 4.2.1).

Linear mixed-effects models were constructed using lme4 (Bates et al., 2015, version 1.1.31). Random effects structures were determined using buildmer (Voeten, 2022, version 2.7), which after identifying the most complex random effects structure that could successfully converge (see Barr et al., 2013), then removed random effects terms which did not significantly contribute towards explaining variance. In a diversion from the pre-registered analysis plan, we excluded the interaction term from the models used to test the main effects of colour legend range and numerical label presence.

5.4.2 Part 1: Participants Satisfying Attention Check Criteria (N = 100)

5.4.2.1 Colour Legend Ranges and Numerical Labels

Figure 5.4 shows the distribution of responses for colour legends with truncated and extended ranges.

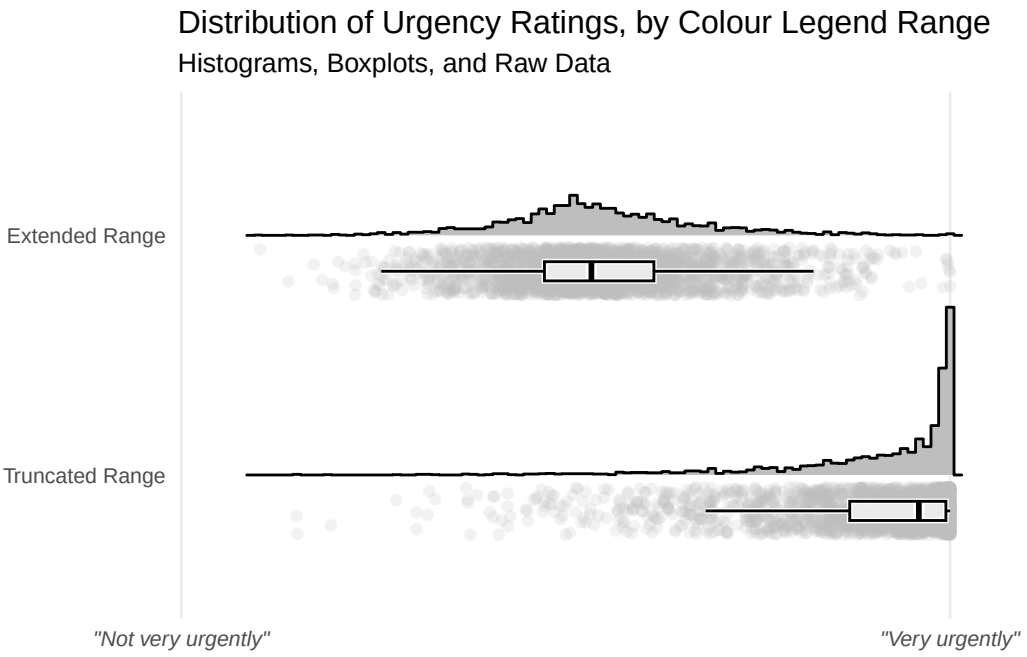


Figure 5.4: Visual analogue scale responses to the question *"How urgently should pollution levels in these regions be addressed?"*. Distributions for the two conditions are shown using histograms, boxplots, and raw data points representing individual observations. In the 'Extended Range' condition, the colour legend's upper bound was equal to double the maximum plotted value. In the 'Truncated Range' condition, the colour legend's upper bound was equal to the maximum plotted value.

Linear mixed-effects modelling revealed that urgency was rated as significantly higher when the colour legend had a truncated range (its upper bound was equal to the maximum value in the dataset) compared to when the colour legend had an extended range (its upper bound was equal to double the maximum value): $\chi^2(1) = 225.41$, $p < .001$, $\eta_p^2 = 0.90$.

Ratings were not significantly different when numerical labels were present, compared to when they were absent: $\chi^2(1) = 0.35$, $p = .556$, $\eta_p^2 < 0.01$.

There was no interaction between colour legend range and numerical labels: $\chi^2(1) = 1.73$, $p = .189$, $\eta_p^2 = 0.02$. These models all employed random intercepts for participants with random slopes for colour legend range, numerical label presence, and the interaction between these terms, plus random intercepts for items.

5.4.2.2 Data Visualisation Literacy

Adding participants' data visualisation literacy as an additional fixed effect did not remove the significant effect of colour legend range: $\chi^2(1) = 260.93$, $p < .001$, $\eta_p^2 = 0.89$. This indicates that differences in data visualisation literacy cannot explain this effect. The numerical label manipulation remained non-significant when accounting for literacy ($\chi^2(1) = 0.30$, $p = .586$, $\eta_p^2 < 0.01$). The interaction remained non-significant when accounting for literacy ($\chi^2(1) = 3.21$, $p = .073$, $\eta_p^2 < 0.01$). These models employed random intercepts for participants with random slopes for colour legend range and numerical label presence, plus random intercepts for items with random slopes for colour legend range.

5.4.3 Part 2: All Participants (N = 165)

5.4.3.1 Colour Legend Ranges and Numerical Labels

The above analysis was conducted using data from the 100 participants who satisfied the pre-registered attention check criteria. However, smaller samples are associated with lower statistical power. Below, we conduct the same analysis on the full sample of 165 participants (those who satisfied the pre-registered attention check criteria and those who did not).

Urgency was rated as significantly higher when a truncated colour legend range was used, compared to when an extended colour legend range was used: $\chi^2(1) = 272.40$, $p < .001$, $\eta_p^2 = 0.87$. Ratings were not significantly different when numerical labels were present, compared to when they were absent: $\chi^2(1) = 1.95$, $p = .163$, $\eta_p^2 = 0.01$. These models employed random intercepts for participants with random slopes for colour legend range, numerical label presence, and the interaction between these terms, plus random intercepts for items with random slopes for colour legend range.

There was a significant interaction between colour legend range and numerical label presence: $\chi^2(1) = 6.41$, $p = .011$, $\eta_p^2 < 0.01$. This model employed random intercepts for participants with

random slopes for colour legend range and numerical label presence, plus random intercepts for items with random slopes for colour legend range. We conducted pairwise comparisons with Sidak adjustment using the emmeans package (version 1.8.2, Lenth, 2021). For choropleth maps with extended colour legend ranges, there was no difference between ratings for labelled and unlabelled colour legends: z = 0.59, p = .962, Cohen's d = 0.02. For choropleth maps with truncated colour legend ranges, higher ratings were awarded when numerical labels were absent, compared to when they were present: z = 2.99, p = .011, Cohen's d = 0.10. Figure 5.5 displays the means and 95% confidence intervals for each combination of conditions, for both samples of participants: those who satisfied the pre-registered attention check criteria, and the full sample.

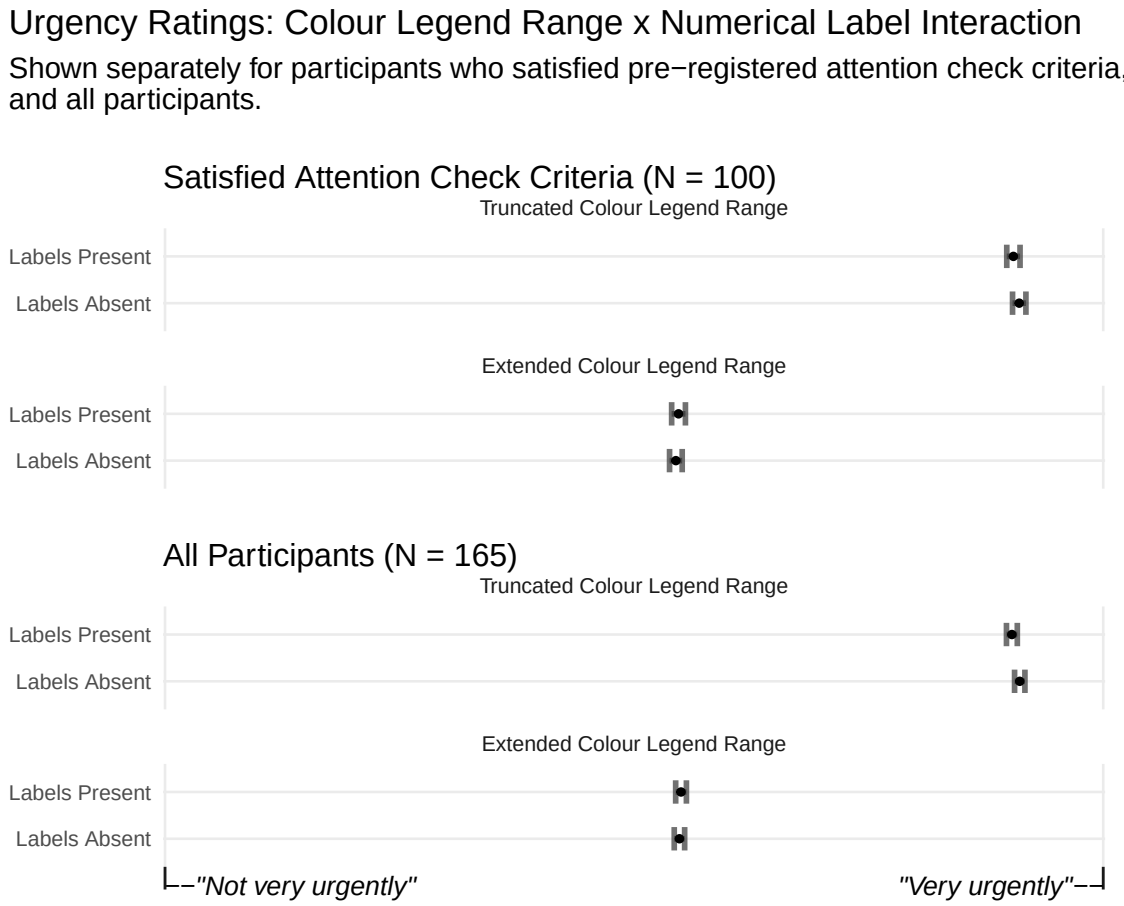


Figure 5.5: Mean urgency ratings showing the interaction between colour legend range and numerical label presence, displayed separately for the different samples of participants. Error bars show 95% confidence intervals around the means.

5.4.3.2 Data Visualisation Literacy

The same pattern of results was observed when accounting for differences in data visualisation literacy. There was a significant effect of colour legend range ($\chi^2(1) = 272.45$, $p < .001$, $\eta_p^2 = 0.87$) and no effect of numerical label presence ($\chi^2(1) = 2.09$, $p = .148$), $\eta_p^2 = 0.01$. The interaction between colour legend range and numerical label presence remained: $\chi^2(1) = 6.47$, $p = .011$, $\eta_p^2 < 0.01$. These models employed random intercepts for participants with random slopes for colour legend range and numerical label presence, plus random intercepts for items with random slopes for colour legend range.

5.4.4 Exploratory Analysis

Our pre-registered analysis did not detect an effect of the presence of numerical values on urgency ratings. However, a more fine-grained analysis can explore the role of numerical labels with greater sensitivity. This exploratory analysis examines whether urgency ratings are influenced by the actual numerical values displayed. We systematically varied the maximum value displayed in each map, which ranged from 200 to 900. Other plotted values were defined in relation to this value: between 10 and 30 units less than the maximum value. Modelling the effect of different maximum values on ratings will reveal whether judgements were informed by the numerical values displayed.

When considering only maps with numerical labels present, ratings increased as a function of maximum value ($\chi^2(1) = 27.90$, $p < .001$, $\eta_p^2 = 0.48$). This model employed random intercepts for participants with random slopes for colour legend range, plus random intercepts for items with random slopes for colour legend range. However, ratings also increased as a function of maximum value even when numerical labels were absent ($\chi^2(1) = 16.85$, $p < .001$, $\eta_p^2 = 0.32$). This model employed random intercepts for participants with random slopes for colour legend range, plus random intercepts for items. There was no significant interaction between maximum value and numerical label presence ($\chi^2(1) = 2.22$, $p = .137$, $\eta_p^2 < 0.01$). This model employed random intercepts for participants with random slopes for colour legend range and numerical label presence, plus random intercepts for items with random slopes for colour legend range.

This suggests that the numerical labels themselves were not responsible for the effect of maximum value. Instead, this effect may have been driven by the appearance of the choropleth map. The colour for the maximum value was identical in each map with the same colour palette, but the three *accompanying* values in each map were always between 10 and 30 units less than the maximum value. Consequently, these values were represented by darker colours when the maximum value was higher, thus conveying greater overall magnitude. Colour legend range ($\eta_p^2 = 0.89$) remains a greater influence than maximum value ($\eta_p^2 = 0.44$).

In the models for participants who satisfied the pre-registered attention check criteria *and* those who did not ($N = 165$), there were significant effects of maximum value, for both maps

with labelled colour legends ($\chi^2(1) = 28.55$, $p < .001$, $\eta_p^2 = 0.50$) and also maps with unlabelled colour legends ($\chi^2(1) = 16.27$, $p < .001$, $\eta_p^2 = 0.32$). These models employed random intercepts for participants with random slopes for colour legend range, plus random intercepts for items with random slopes for colour legend range. There was no significant interaction between maximum value and numerical label presence ($\chi^2(1) = 3.51$, $p = .061$, $\eta_p^2 < 0.01$). This model employed random intercepts for participants with random slopes for colour legend range and numerical label presence, plus random intercepts for items with random slopes for colour legend range. Colour legend range ($\eta_p^2 = 0.87$) remains a greater influence than maximum value ($\eta_p^2 = 0.45$).

5.5 Discussion

Choropleth maps are typically used to convey spatial variability, but may alternatively be employed to convey overall magnitude. This experiment clearly demonstrated that the range of the accompanying colour legend influences interpretations of absolute magnitude in such choropleth maps. When the colour legend's upper bound was equivalent to the maximum plotted value, participants rated the urgency of addressing pollution levels as higher, compared to when the colour legend's upper bound was equal to double the maximum plotted value. This illustrates that viewers use colour legends to put numbers' magnitudes into perspective, interpreting magnitude with respect to the range of the colour legend. A colour legend does not only provide a mapping between numerical values and colours, it also provides a range of values relevant for considering the absolute magnitude of presented data.

Crucially, the colours used to display the data in the maps, as well as the underlying numerical values, were identical across conditions. Therefore, differences in participants' judgements between conditions were not due to these factors. Instead, participants formed different impressions of these data based on the context in which they were presented. We do not suggest that one colour legend arrangement used in this experiment was misleading and the other truthful. Rather, we suggest that, under certain circumstances, either could be characterised as misleading. Thus, doctored data and deliberate deception are not the only practices behind problematic visualisations.

Colour legends simultaneously encode changes in number through both colour and physical position. Different values are represented by different colours *and* occupy different positions on the colour legend. In the present experiment, plotted values' analogous positions in the truncated colour legend were on the far right hand side, and their corresponding colours were among the darkest in the legend. On the other hand, plotted values' analogous positions were in the middle of the extended colour legend, and their corresponding colours were neither the darkest nor the lightest in the legend. This experiment cannot determine whether the location of plotted values on the legend, the range of colours included in the legend, or both of these factors, influenced processing of magnitude. The manipulation of numerical labels does not assist in answering this question because colour legends still encode changes in number even

when these changes are not labelled. However, this question may have little practical relevance since these aspects are intrinsically linked in a typical colour legend.

In this experiment, the width of truncated and extended colour legends was identical. In the truncated colour legend, a smaller range of colours spanned the same distance: there was less variation in colour over the same amount of space. We have not identified any way in which this could explain the present set of results.

5.5.1 Additional Analyses

Accounting for subjective data visualisation literacy did not change the pattern of results. This suggests that data visualisation literacy is not responsible for the observed effect of colour legend range on interpretations of magnitude. This accords with the finding that data visualisation literacy levels did not explain the bias in judgements caused by truncated axes (Yang et al., 2021). Yang et al. (2021) suggest that data visualisation literacy measures capture whether an individual has the skills required for comprehending typical chart formats. However, they do not appear to extend to aspects of visualisation comprehension which are informed by intuitive judgements rather than basic training.

Our results demonstrate that numerical labels did not influence judgements. Our pre-registered analysis found that there was no difference between ratings for maps with and without numerical labels on the colour legend. An exploratory analysis examining this further also indicates that increases in the numerical values displayed on the colour legend were not responsible for greater urgency ratings. Instead, it is likely that increased urgency ratings associated with higher maximum values were related to the presence of darker colours in the maps. This was a consequence of accompanying data points' increased proximity to the maximum value at higher maximum values (see Figure 5.2).

For data quality reasons, we conducted our main analysis on a sample of 100 participants who met our pre-registered attention check threshold (no more than one of six attention check questions answered incorrectly). However, we also conducted the same analysis on the full sample of 165 participants, in the interest of validity. The pattern of results in the two samples was extremely similar, indicating similar levels of engagement with the task regardless of attention check scores. Participants may have withdrawn attention from the accompanying text and question once they were aware that these did not change across experimental trials, consequently failing to notice attention-check trials.

The only difference between the pattern of results for these two samples was the interaction between colour legend range and numerical label presence. This interaction was not observed in the more selective sample but observed in the full sample. However, Figure 5.5 illustrates that the pattern of responses was remarkably similar. In both samples, the difference between ratings for the labelled and unlabelled versions of the truncated colour legend was very small, which suggests the significant result was driven by low variance within conditions and increased

statistical power in the larger sample. The inconsistency in inferential statistics between samples suggests that this interaction, if not spurious, is not particularly robust.

5.5.2 Relationship to Prior Work

Recommendations for best practice in choropleth map design are focused on conveying plotted values' relative magnitudes (Dent et al., 2009; Kraak and Ormeling, 2013). In this work, we suggest that efficiently conveying relative magnitudes is a *sufficient* condition for choropleth mapping, but not a *necessary* condition. We demonstrate that encoding plotted values with a smaller range of colours, and including a wider range in the accompanying legend, informs judgements about *absolute* magnitude. This is consistent with other experiments demonstrating legend design can affect cognitive processing of an accompanying map (Edler et al., 2020; Gołębiowska, 2015; Hepburn et al., 2021; Li and Qin, 2014).

Investigations into chart design have revealed that the range of values surrounding plotted data influences interpretations. Several experiments have observed that participants use axes as a source of context for assessing the magnitude of difference between values (Correll et al., 2020; Pandey et al., 2015; Witt, 2019; Yang et al., 2021). The present experiment provides further evidence for a less-frequently explored phenomenon: that design choices can affect judgements of *the magnitude of values themselves*. Like Stone et al. (2003) and Sandman et al. (1994), we demonstrate that plotted values seem greater when they are closer to a data visualisation's upper bound. However, this experiment also demonstrates that these types of effects are not unique to data visualisations using geometric encodings. Choropleth maps, where the range of values is presented in a colour legend, can also elicit this bias. Arguably, the manipulation in choropleth maps is even more subtle, because of the unique way that choropleth maps separate encoded data from the colour legend. In data visualisations such as bar charts, changing the range of values alters the appearance of the data itself (an extended y-axis results in a compressed bar). The present experiment's findings are particularly striking given that the appearance of data remained consistent despite changes to the colour legend's upper bound. This suggests differences in judgements were not driven by the visual appearance of the data, but by the interpretation of the data in relation to the range of values in the colour legend.

This finding is also connected to research on the interpretation of quantity in colourmap visualisations. Schiewe (2019) observed that assessment of values presented in choropleth maps are influenced by the coverage of different colours within a map (i.e., the relationship between colour and region size). We expand upon this work by identifying another factor which biases judgements of data in choropleth maps, yet does not change the appearance of the map itself. Like Correll et al. (2018), we demonstrate that manipulating a colour legend is sufficient to influence participants' responses. Schloss et al.'s (2019) results demonstrated that a colourmap's background colour is interpreted as corresponding to the smallest quantity when a scale appears to vary in opacity. That is, background colour provides a cue to the size of data points when taken to represent the minimum value. The present experiment

demonstrates that, like quantity judgements, magnitude judgements are also driven by visual cues to the minimum and maximum values.

A bias wherein the same values are judged differently depending on their surrounding context is often described as a framing effect (Tversky and Kahneman, 1981). This bias involves using inessential accompanying information to inform one's judgement, rather than discounting this information in order to generate a wholly disinterested assessment. Other research has also demonstrated that the interpretation of numerical values depends on their placement within a range. For example, the same salary is rated as more desirable when it appears near the top rather than the bottom of a range (Brown et al., 2008). The present experiment translates this effect to the visual domain. As Yang et al. (2021) suggest, biases in viewers' processing of information in data visualisations can be explained with reference to Grice's (1975) cooperative principle. Applied to the present experiment, this suggests that viewers would interpret the implication of certain magnitudes through the colour legend design as indicative of the designer's intention to communicate values' true magnitudes.

5.5.3 Limitations and Future Research Directions

Choropleth maps are typically designed to communicate differences between values, rather than values' absolute magnitudes. Discrimination between values is facilitated when the colour legend's bounds are equal to the minimum and maximum values in the dataset. Therefore, designers may have to make a trade-off between conveying absolute magnitude and conveying differences. Which aspect of the data a designer wishes to emphasise will depend on the purpose of their data visualisation. For example, a designer may wish to highlight the geographical differences in the construction of new houses, or may wish to highlight the fact that there is no region where targets are being met. The work reported here suggests that extending the range of the colour legend beyond the range of the observed data would promote the latter message.

It is important to recognize that a colour scale's bounds may not always be interpreted as a complete and accurate source of context for assessing magnitude. Pollution measurements are likely not among the most intuitive numbers to interpret, and in the present experiment, even viewers well-versed in pollution data were prohibited from applying their knowledge, since the fictitious data were presented using fictitious units. The influence of existing knowledge was eliminated to facilitate examination of the cognitive mechanism involved in magnitude judgements. Therefore, in this experiment, there were no *external* cues to magnitude. Consequently, our findings are most relevant for understanding interpretation of magnitude where units are unfamiliar or insignificant. Familiarity with a data visualisation's subject matter will typically provide an ability to independently assess magnitudes based on presented values only, which may reduce the influence of design choices. In addition, certain forms of number may carry cues to magnitude even in the absence of existing knowledge. For example, when assessing certain proportions, viewers are likely to be aware that 100% is the maximum possible value

and 0% the minimum. Future work should explore the degree to which these scenarios affect how colour legends inform magnitude judgements.

Future work should quantify the difference between different colour legend ranges in concrete units (e.g., a specific difference in financial investment, or a specific time-frame for resolving an issue). The visual analogue scale used in our investigation does not permit this. However, it was able to reveal that interpretations of magnitude differed between conditions, reflecting the type of inferences that are likely to precede decision-making. The within-participants design ensures that participants' different notions of urgency do not interfere with comparisons between experimental conditions. Future work should also examine a wider variety of topics beyond pollution data in order to examine generalizability. However, our investigation has nonetheless produced informative results, and the observed bias, a framing effect, occurs widely.

Numerical labels at the extremes of colour legends are sometimes open-ended. That is, a label at the lower bound may be '<30' rather than '30'. This interrupts the one-to-one mapping between colours and values. Instead, a specific position and colour on the colour legend may represent multiple corresponding numerical values. Consequently, *more extreme* values may exist in the data than those represented by the extremes of the legend. This introduces ambiguity regarding the relevant range of values to consider when assessing magnitude, making the colour legend a less informative reference. Future research should examine whether the present findings are replicated when a colour legend uses this type of numerical label at its extremes, or whether viewers treat colour legends with these labels as a weaker cue to plotted values' magnitudes. Experiments varying the range of values included in classified and multi-hue legends would also be beneficial.

5.5.4 Implications

The present experiment contributes to our understanding of cognitive mechanisms involved in assessing magnitudes in choropleth maps. We observed that assessments are informed by the range of the colour legend, demonstrating that colour legends can be exploited to influence viewers' judgements of data points' absolute magnitudes. Further work is required in order to identify various factors influencing the strength of this effect, but the essential implication entails designers considering how magnitude appears as a result of their chosen colour legend's range. Without deliberate consideration about the choice of value for a colour legend's upper bound, misleading visualisations may emerge. However, like Correll et al. (2020), we argue there can be no *a priori* system for identifying a range of values that guarantees an unbiased visualisation. Instead, the range of the colour legend should be appropriate for the data displayed, the intended message, and the task. There are also implications for data visualisation software developers in facilitating designers' ability to specify a custom colour legend range when required.

5.6 Conclusion

Understanding the consequences of design choices is crucial for understanding how to present data effectively. In choropleth maps, the upper bound of the accompanying colour legend influences how large or small plotted values appear to viewers. Data points' proximity to the upper bound increases impressions of their absolute magnitude. This finding provides insight into the processing of choropleth maps designed to convey overall magnitude, and promotes use of a suitable range of values on a colour legend.

5.7 References

References

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**:255–278. doi:10.1016/j.jml.2012.11.001

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* **67**. doi:10.18637/jss.v067.i01

Bertini E, Correll M, Franconeri S. 2020. Why Shouldn't All Charts Be Scatter Plots? Beyond Precision-Driven Visualizations. *arXiv:200811310 [cs]*.

Brown GDA, Gardner J, Oswald AJ, Qian J. 2008. Does Wage Rank Affect Employees' Well-being? *Industrial Relations* **47**:355–389. doi:10.1111/j.1468-232X.2008.00525.x

Brychtova A, Coltekin A. 2015. Discriminating classes of sequential and qualitative colour schemes. *International Journal of Cartography* **1**:62–78. doi:10.1080/23729333.2015.1055643

Correll M, Bertini E, Franconeri S. 2020. Truncating the Y-Axis: Threat or Menace?Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM. pp. 1–12. doi:10.1145/3313831.3376222

Correll M, Moritz D, Heer J. 2018. Value-Suppressing Uncertainty PalettesProceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal QC Canada: ACM. pp. 1–11. doi:10.1145/3173574.3174216

Cromley RG, Ye Y. 2006. Ogive-based Legends for Choropleth Mapping. *Cartography and Geographic Information Science* **33**:257–268. doi:10.1559/152304006779500650

Dasgupta A, Poco J, Rogowitz B, Han K, Bertini E, Silva CT. 2020. The Effect of Color Scales on Climate Scientists' Objective and Subjective Performance in Spatial Data Analysis Tasks. *IEEE Transactions on Visualization and Computer Graphics* **26**:1577–1591. doi:10.1109/TVCG.2018.2876539

Dent BD, Torguson J, Hodler TW. 2009. Cartography: Thematic Map Design, 6th ed. ed. New York: McGraw-Hill Higher Education.

Driessen JEP, Vos DAC, Smeets I, Albers CJ. 2022. Misleading graphs in context: Less misleading than expected. *PLOS ONE* **17**:e0265823. doi:10.1371/journal.pone.0265823

Dykes J, Wood J, Slingsby A. 2010. Rethinking Map Legends with Visualization. *IEEE Transactions on Visualization and Computer Graphics* **16**:890–899. doi:10.1109/TVCG.2010.191

Edler D, Keil J, Tuller M-C, Bestgen A-K, Dickmann F. 2020. Searching for the "Right" Legend: The Impact of Legend Position on Legend Decoding in a Cartographic Memory Task. *The Cartographic Journal* **57**:6–17. doi:10.1080/00087041.2018.1533293

Fischer J, Ali A. 2021. A Federal Ban on Abortion is Wildly Unpopular in All 50 States. *Data For Progress*.

Galesic M, Garcia-Retamero R. 2011. Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making* **31**:444–457. doi:10.1177/0272989X10373805

Garcia-Retamero R, Cokely ET, Ghazal S, Joeris A. 2016. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making* **36**:854–867. doi:10.1177/0272989X16655334

Garcia-Retamero R, Galesic M. 2010. Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine* **70**:1019–1025. doi:10.1016/j.socscimed.2009.11.031

Gołębiowska I. 2015. Legend Layouts for Thematic Maps: A Case Study Integrating Usability Metrics with the Thinking Aloud Method. *The Cartographic Journal* **52**:28–40. doi:10.1179/1743277413Y.0000000045

Grice P. 1975. Logic and Conversation In: Cole P, Morgan JL, editors. Syntax and Semantics Vol.3: Speech Acts. New York: Academic Press. pp. 41–58.

Harrower M, Brewer CA. 2003. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal* **40**:27–37. doi:10.1179/000870403235002042

Hegarty M. 2011. The Cognitive Science of Visual-Spatial Displays: Implications for Design. *Topics in Cognitive Science* **3**:446–474. doi:https://doi.org/10.1111/j.1756-8765.2011.01150.x

Hepburn J, Fairbairn D, James P, Ford A. 2021. Do we need legends? An eye tracking study. doi:10.5281/ZENODO.4665195

Hu T-Y, Jiang X-W, Xie X, Ma X-Q, Xu C. 2014. Foreground-background salience effect in traffic risk communication. *Judgment and Decision Making* **9**:8.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**:90–95. doi:10.1109/MCSE.2007.55

Jenks GF, Caspall FC. 1971. Error On Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers* **61**:217–244. doi:10.1111/j.1467-8306.1971.tb00779.x

Kraak M-J, Ormeling FJ. 2013. Cartography: Visualisation of Spatial Data, 3rd ed. Routledge. doi:10.4324/9781315847184

Kumar N. 2004. Frequency Histogram Legend in the Choropleth Map: A Substitute to Traditional Legends. *Cartography and Geographic Information Science* **31**:217–236. doi:10.1559/1523040042742411

Lenth RV. 2021. Emmeans: Estimated Marginal Means, aka Least-Squares Means.

Li Z, Qin Z. 2014. Spacing and alignment rules for effective legend design. *Cartography and Geographic Information Science* **41**:348–362. doi:10.1080/15230406.2014.933085

Lin S, Fortuna J, Kulkarni C, Stone M, Heer J. 2013. Selecting Semantically-Resonant Colors

for Data Visualization. *Computer Graphics Forum* **32**:401–410. doi:10.1111/cgf.12127

Okan Y, Stone ER, Parillo J, Bruine de Bruin W, Parker AM. 2020. Probability Size Matters: The Effect of Foreground-Only versus Foreground+Background Graphs on Risk Aversion Diminishes with Larger Probabilities. *Risk Analysis* **40**:771–788. doi:10.1111/risa.13431

Pandey AV, Rall K, Satterthwaite ML, Nov O, Bertini E. 2015. How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion TechniquesProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. Seoul, Republic of Korea: ACM Press. pp. 1469–1478. doi:10.1145/2702123.2702608

Paul BK. 1993. Choropleth Map Review: A Class Exercise. *Journal of Geography* **92**:227–230. doi:10.1080/00221349308979658

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* **51**:195–203. doi:10.3758/s13428-018-01193-y

R Core Team. 2022. R: A Language and Environment for Statistical Computing.

Retchless DP, Brewer CA. 2016. Guidance for representing uncertainty on global temperature change maps. *International Journal of Climatology* **36**:1143–1159. doi:10.1002/joc.4408

Sandman PM, Weinstein ND, Miller P. 1994. High Risk or Low: How Location on a "Risk Ladder" Affects Perceived Risk. *Risk Analysis* **14**:35–45. doi:10.1111/j.1539-6924.1994.tb00026.x

Schiewe J. 2019. Empirical Studies on the Visual Perception of Spatial Patterns in Choropleth Maps. *KN - Journal of Cartography and Geographic Information* **69**:217–228. doi:10.1007/s42489-019-00026-y

Schloss KB, Gramazio CC, Silverman AT, Parker ML, Wang AS. 2019. Mapping Color to Meaning in Colormap Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics* **25**:810–819. doi:10.1109/TVCG.2018.2865147

Stone ER, Sieck WR, Bull BE, Frank Yates J, Parks SC, Rush CJ. 2003. Foreground:background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes* **90**:19–36. doi:10.1016/S0749-5978(03)00003-7

Stone ER, Yates JF, Parker AM. 1997. Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied* **3**:243–256. doi:10.1037/1076-898X.3.4.243

Stone M, Szafir DA, Setlur V. 2014. An Engineering Model for Color Difference as a Function of Size. Boston, Massachusetts: Society for Imaging Science; Technology. p. 6.

Szafir DA, Stone M, Gleicher M. 2014. Adapting Color Difference for Design. Boston, Massachusetts: Society for Imaging Science; Technology. p. 6.

Taylor BG, Anderson LK. 1986. Misleading Graphs: Guidelines for the Accountant. *Journal of Accountancy* **162**:126–135.

Tobler WR. 2010. Choropleth Maps Without Class Intervals? *Geographical Analysis* **5**:262–265. doi:10.1111/j.1538-4632.1973.tb01012.x

Tversky A, Kahneman D. 1981. The Framing of Decisions and the Psychology of Choice. *Science* **211**:453–458. doi:10.1126/science.7455683

Voeten CC. 2022. Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects.

Wickham H. 2016. ggplot2. New York, NY: Springer Science+Business Media, LLC.
Witt JK. 2019. Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis. *Meta-Psychology* **2**:1–20. doi:10.15626/MP.2018.895
Yang BW, Vargas Restrepo C, Stanley ML, Marsh EJ. 2021. Truncating Bar Graphs Persistently Misleads Viewers. *Journal of Applied Research in Memory and Cognition* S2211368120300978. doi:10.1016/j.jarmac.2020.10.002

Postscript to Chapter 4

Overview

The second study in this thesis was designed to examine whether colour legends influence interpretations of absolute magnitude in choropleth maps. Colour legends convey the mapping between numerical values and colours, and can be extended beyond the range of the dataset to incorporate other relevant values. This study revealed that magnitude was interpreted as greater when plotted data points were numerically high within the range of values on the colour legend. Thus, a colour legend provides a frame of reference which informs subjective judgements of magnitude.

Rationale for Experimental Design

Unlike the previous study, generating stimuli for this study involved identifying topics which could plausibly be associated with geographic regions. The type of data presented was also a relevant consideration in order to broaden the scope of my investigation. The previous study on dot plots presented relative frequencies, and displaying absolute frequencies is not appropriate in choropleth maps (Dent, 2008). Therefore, to manage these constraints, reduce complexity, and expand inquiry to a different type of data, all choropleth maps in this study displayed pollution measurement data.

Using fictitious standardised units avoided several factors interfering with magnitude judgements: participants' prior knowledge, explanation of pollution measurements, and difficulty comprehending scientific units. The use of single topic and the reduced ecological validity ultimately afforded greater experimental control.

This study explored interpretations of magnitude using rating scales, rather than measuring behaviour. However, participants were asked to rate the *urgency* with which plotted pollution levels should be addressed. This specific judgement assists in capturing impressions of overall magnitude, and avoids asking participants the abstract question of how 'large' values plotted are. Furthermore, assessing urgency requires consideration of the degree of action required for addressing pollution levels, engaging participants' decision-making capacities.

Summary of Findings

This work demonstrated that colour legends' limits contribute to interpretations of absolute magnitude in choropleth maps. The corresponding location of plotted data points within the accompanying colour legend influenced magnitude judgements, despite consistency in the colours used to convey plotted values. Magnitude ratings were higher when data points were located closer to the upper limit of the colour legend. This suggests that the colour legend was used as a frame of reference when interpreting how high or low plotted values were. This investigation also revealed that magnitude judgements were not influenced by the numerical values of plotted data points. It is likely that the use of fictitious standardised units prevented participants from drawing on contextual knowledge when assessing magnitude.

Choropleth maps are usually designed in order to display spatial variability in a dataset: relative magnitude. Indeed, prior work has explored in depth how designs may promote optimal discrimination between plotted values (Jenks and Caspall, 1971; Brychtova and Coltekin, 2015; Dent et al., 2009). However, this typical application does not require that choropleth maps are only used for that purpose. This study illustrates that choropleth maps can also be legitimately employed to convey absolute magnitude. Thus, alternative approaches to data visualisation design which break with convention can still be valid.

This study also accords with the observation that interpretations of values in choropleth maps can depend on inferences about the minimum and maximum values in a colour legend (Schloss et al., 2019). Schloss et al. (2019) found that a map's background colour influenced judgements about whether plotted values were large or small (quantity), whereas this study found that legends' limits influenced judgements about *how* large or small plotted values were (subjective magnitude).

Relationship to the Study in Chapter 3

The first study in this thesis demonstrated that axis limits affect interpretation of magnitude in dot plots. This second study demonstrates that these findings generalise to a different type of data visualisation: choropleth maps. These results build upon the first study's observations by demonstrating that the numerical context accompanying data points influences interpretations even when the physical characteristics of the plotted values do not change. This also accords with the first study's finding that judgements were primarily informed by the relative numerical positions of plotted values, as opposed to their absolute physical positions. Other contributions include demonstrating that this bias occurs when manipulating only one axis limit and when using a horizontal (rather than vertical) colour legend.

Preface to Chapter 5

The empirical research conducted in this thesis thus far has demonstrated that interpretations of magnitude are influenced by the numerical context in which plotted values are presented. However, this thesis has not yet explored denominator values, which are another relevant aspect of context for magnitude judgements. Knowledge of a denominator is relevant to judging absolute frequencies. For example, when evaluating a business expense, awareness of the total available budget can provide perspective on the magnitude of the amount spent. Including denominator values in axes is another example of conveying relevant values that do not appear in plotted data. Therefore, the third and final investigation in this thesis explores how this type of contextual information informs interpretations of magnitude in bar charts.

This final study can also assist in understanding how the influence of visualisation design choices on a viewer's interpretation is affected by their awareness of a dataset's characteristics. The first study in this thesis observed a relatively small bias in the interpretation of probability values in familiar scenarios. The second study in this thesis observed a relatively large bias in the interpretation of ambiguous pollution measurements. The third and final study in this thesis contains a manipulation of participants' knowledge about plotted datasets. This provides insight into how knowledge of the numerical context behind plotted data can affect magnitude judgements, in addition to graphical cues.

6 Axis Limits and Denominator Information Influence Magnitude Ratings in Bar Charts

Consider a statistic corresponding to the number of public transport users in a particular town. Gauging whether this number is large or small requires awareness of the total population (the denominator). In data visualisations, an axis extended beyond plotted values can act as a graphical cue to a denominator value, but default axis upper limits (e.g., in ggplot2) are typically based on the highest plotted value. In two experiments (combined N = 350), we explore the influence of default and extended axes on interpretations of the magnitude in bar charts. We also investigate the influence of accompanying denominator information on participants' assessments. We observe that values plotted using default axes were rated as higher, compared to values plotted using extended axes. The absence of denominator information amplifies the effect of axis limits on judgements. This demonstrates that axes which incorporate denominator values influence interpretations of presented data. Whereas prior work has often focused on judgements of the *differences between values*, this work contributes to an understanding of how the magnitudes of *the values themselves* are interpreted by viewers. We also discuss implications for effective design, which involve considering both axis limits and accompanying contextual information.

6.1 Introduction

The question 'Is it a big number?' is often raised on the BBC radio programme *More or Less* when probing eye-catching statistics. A figure of several million pounds may initially seem large, but may represent a small proportion of total government spending. Awareness of a denominator value can influence judgement of a number's magnitude. In data visualisation, this contextual information can be displayed by extending an axis to accommodate the denominator value. However, this approach is infrequently used, since typical default axis settings are based on plotted data only. In this study, we investigate how these axis limits affect interpretations of how large or small plotted values are.

6.1.1 Overview

Across two experiments, we investigate the interpretation of magnitudes in bar charts. We plotted fictitious datasets which contained multiple observations, all with the same denomina-

tor values. In the first experiment, we displayed charts with default axes which terminated just above plotted values, or axes which extended to the denominator value specified in accompanying text. Participants rated values' magnitudes as higher when default axes were used, compared to extended axes. In the second experiment, we manipulated the axis limits, as before, and also the presence of the denominator information in accompanying text. Variation in responses to the two difference axis settings was starker when denominator information was not supplied. This indicates that this information influences the biasing effect of a chart's appearance.

6.1.2 Related Work

A wealth of research demonstrates biases in the interpretation of numbers. A survival rate elicits different judgements of a disease compared to its corresponding mortality rate (Tversky and Kahneman, 1981), the fat content of a meat product elicits different judgements of the product compared to its corresponding lean content (**levin_associative_1987?**). The units used to express the same values (e.g., months vs. years) affect comparisons (**burson_six_2009?**; **monga_years_2012?**) and also interpretations of precision and accuracy (**zhang_how_2012?**).

Various biases in magnitude judgements reveal the importance of accounting for numerical context. Base rate neglect describes difficulty acknowledging population-level characteristics when making judgements about a sample (**cosmides_are_1996?**). Format neglect describes a bias against incorporating set size information when judging percentage formats (top 20%) and numerical formats (top 10, (**sevilla_format_2018?**)). Denominator neglect occurs in judgements which overweight numerator information at the expense of denominator information (**reyna_numeracy_2008?**). The latter also leads (in part) to large percentages of a small number appearing greater than the *numerically equivalent* smaller percentages of a larger number (**li_big_2013?**). The general mechanism responsible for these biases is a failure to properly acknowledge numerical context.

Visualisations can help combat biases. Denominator information becomes visually available when icon arrays present both focal outcomes (e.g., number deceased) *and also* alternative outcomes (e.g., number survived). Research suggests that the combined array acts as a visual cue to the denominator (e.g., total number at risk), facilitating reasoning. For example, including alternative outcomes in an icon array, and therefore displaying denominator information, reduces denominator neglect, increasing comprehension of relative risk (Garcia-Retamero and Galesic, 2010). Icon arrays displaying both types of outcome are particularly helpful for understanding datasets with unequal denominators, and for individuals with high graph literacy (**okan_individual_2012?**). However, these effects are largest when depicting small probabilities (Okan et al., 2020).

Stacked bar charts function similarly to icon arrays: lower bars represent the focal outcome, upper bar represent the alternative outcome, and their combination represents the denomina-

tor. Like icon arrays, stacked bar charts lessen the influence of denominator neglect (Stone et al., 2003). However, denominator information can be displayed in bar charts without using additional stacked bars representing alternative outcomes. Extending an axis to incorporate the denominator value also communicates relevant numerical context. In this case, the blank space between the bars for focal outcomes and the upper axis limit corresponds to the alternative outcomes. Research has demonstrated that bar charts representing alternative outcomes using blank space increase perceptions of risk likelihood compared to those representing alternative outcomes using stacked bars (**stone_designing_2017?**). This research did not examine how presenting *denominator information* influences interpretation, since identical axis limits were employed across conditions.

Directly manipulating bar charts' upper axis limits provides insight into use of this source of denominator information. Bar charts with axes which extended to the denominator value produce more accurate estimates of changes in risk (Garcia-Retamero and Galesic, 2010). This design also elicits decreased ratings of risk perception (**okan_designing_2018?**), though numerical labels reduce this effect. In both studies, accompanying text included the denominator value. These studies demonstrate that extending axes to incorporate denominator values influences interpretation of risk. However, they do not provide specific evidence on how interpretations of *plotted values' magnitudes* are affected. Garcia-Retamero and Galesic (2010) measured risk understanding: how faithfully participants represented the exact numbers displayed. Only assessing comprehension fails to capture individuals' *impressions* of plotted information (**feldman-stewart_further_2007?**). Understanding the 'gist' obtained from a visualisation is crucial since this takes precedence over 'verbatim' information when making decisions (**reyna_theory_2008?**). (**reyna_theory_2008?**) argues that assessing gist requires consideration of how plotted values' magnitudes are interpreted, since plotted values' relative differences are only one aspect of a dataset conveyed by a visualisation. Indeed, (**okan_designing_2018?**) collected magnitude ratings, yet these cannot be examined in isolation, since they were assimilated into a combined measure of perceived risk, along with ratings of the degree of *difference between* plotted values. Outside the risk communication literature, a substantial body of research has demonstrated that judgements of the difference between values change as a function of axis range (Correll et al., 2020; Pandey et al., 2015; Witt, 2019; Yang et al., 2021). In spite of this, research has neglected the effects of axis range on judgements of the magnitude of *the values themselves*. This is the focus of experiments presented in this work.

Various accounts have sought to explain the consequences of including denominator information in visualisations. Depicting denominators may facilitate understanding of part-to-whole relationships, diminishing class-inclusion errors associated with denominator neglect (**reyna_theory_2008?**). This argument is consistent with Fuzzy Trace Theory, which also predicts the influence of physical attributes in gist representations, such that the appearance of short bars in charts with extended axes may contribute to smaller magnitude judgements (**reyna_theory_2008?**). Additionally, (**stone_salience_2018?**) suggest facilitation of proportional reasoning may be largely responsible for observed effects. Increasing the salience

of the denominator in text fails to affect judgements, yet a graphical representation effectively communicates the true scale of the denominator, helping put numerators into perspective.

Prior work on extending axes, discussed above, did not disclose methods for determining axis limits in charts without denominators. The values used as upper limits appear to be arbitrary. In the present study, we increase ecological validity by employing default axis limits from {ggplot2} (Wickham, 2016), a popular visualisation tool used in the R programming environment. Furthermore, previous studies' statistical power and generalisability have been limited by the use of one (**okan_designing_2018?**) or two (Garcia-Retamero and Galesic, 2010) trials per participant. Our experiments explore a range of scenarios (32 experimental trials per participant). The data presented are also unrelated to risk judgements, the domain of this prior work.

6.1.3 Open Research Statement

Data, analysis code, and pre-registrations are available at osf.io/854uc/.

6.2 Experiment 1

6.2.1 Introduction

This experiment investigates the influence of axis limits on interpretations of plotted values' magnitude. Participants viewed bar charts with default axes, or axes which extended to a denominator value well above the bars. Comparing participants' interpretations captures the influence of displaying the same data with and without numerical context.

6.2.2 Method

6.2.2.1 Materials

We developed 40 scenarios about fictitious studies. Each study evaluated a specific outcome across five categories (e.g., the number of items produced without defects, for five manufacturing methods). The denominator (e.g., total number of items produced) was identical for each category.

We generated bar charts in R (R Core Team, 2022) using {ggplot2} (version 4.1.2), {tidyverse} (version 1.3.1) and {ggh4x} (version 0.2.1). The two versions of each chart displayed the same five values, but employed different y-axis limits. Denominator values (400, 500, or 600) were used to generate datasets: data were sampled from normal distribution with a mean equal to either 20% or 40% of a given denominator value, and a standard deviation equal to 1% of the denominator value.

For charts with extended axes, the denominator value was used as the y-axis upper limit. The other charts used a y-axis upper limit which was dictated by ggplot2's default axis settings. These settings automatically identify a set of convenient breaks for each dataset, then slightly extend the plot area, adding an additional 5% of the axis range. In both conditions, a smaller expansion factor of 1% was applied to the lower axis limit, in order to eliminate visible space below the 0 baseline. Figure 6.1 shows example charts for both conditions.

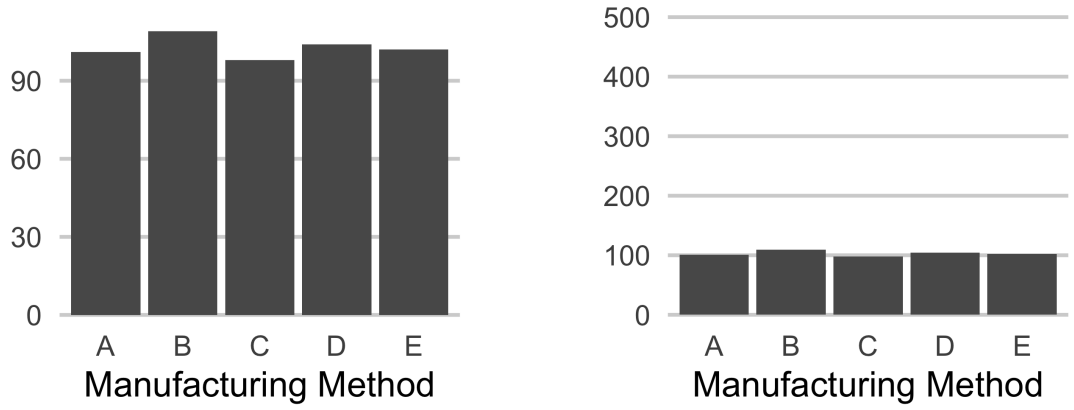


Figure 6.1: Example charts for Experiment 1. The same data appears in both charts. Accompanying text explained what the values represented: *'The graph shows, for each manufacturing method, how many of the items were free from defects'.* The chart with the default axes (left) employs an upper limit determined by ggplot2. The chart with the extended axes (right) employs an upper limit equal to the dataset's denominator value.

For the majority of datasets generated, the default settings produced charts where the highest gridline did not exceed the tallest bar. For consistency, when the opposite situation occurred, we used a different random seed to generate an alternative dataset for both conditions. 10% of datasets used were generated using this method.

In experimental trials (32 total), plotted values consisted of relatively small proportions of the dataset's denominator value (roughly 20% or 40%). To introduce variety and encourage attention, eight filler trials showed plotted values which were roughly 90% of the corresponding denominator value. Denominators for filler trials were selected so that numerical labels on the y-axis would approximately resemble either extended or default bar charts from experimental trials.

We included six attention check trials to assess participants' engagement with the task. These trials were similar to experimental and filler trials, consisting of text, a bar chart, a question and a visual analogue scale. However, participants were instructed to ignore the bar chart and

provide a specified response on the visual analogue scale.

6.2.2.2 Design

We employed a within-participants design: participants viewed 16 different charts in each of the two conditions (32 experimental trials total). The correspondence between scenarios and conditions was counterbalanced using two lists. However, all participants saw the same versions of the eight filler items and six attention check items. There were a total of 46 trials, which were presented in a random order.

6.2.2.3 Participants

Participants were recruited using Prolific.co. The experiment was advertised to fluent English speakers with normal-or-corrected to normal vision, who had previously participated in at least 100 studies on the site.

Data were returned by 157 participants. Per pre-registered exclusion criteria, seven participants' submissions were rejected because they answered more than one of six attention check questions incorrectly. Participants whose submissions were accepted received £3.50.

The final sample consisted of 150 participants (57.33% male, 39.33% female, 3.33% non-binary). Mean age was 33.18 years ($SD = 12.93$). The mean data visualisation literacy score was 21.35 ($SD = 4.73$), out of a maximum of 30.

This experiment was approved by the University of Manchester's Division of Neuroscience and Experimental Psychology Ethics Committee (ethics code: 2022-11115-24245).

6.2.2.4 Procedure

We programmed the experiment using PsychoPy (version 2022.1.4, (Peirce et al., 2019)). Participants were instructed to carry out the experiment using a laptop or desktop computer (not a mobile phone or tablet). After providing informed consent, participants completed a demographic questionnaire and Garcia-Retamero et al.'s (Garcia-Retamero et al., 2016) five-item subjective data visualisation literacy scale.

Participants were asked to imagine they were a researcher tasked with determining the outcome of experiments and surveys. They were instructed to make an overall assessment of all data presented in a graph after studying the text, graph, and question. All questions asked about plotted values' magnitudes (e.g., 'How successful were the manufacturing methods?'), with participants responding on visual analogue scales with anchors at the extremes (e.g., 'very unsuccessful', 'very successful'). Figure 6.2 shows an example trial.

Figure 6.2: An example trial from Experiment 1, showing a bar chart with a default axis limit.

Participants were permitted to move the response marker as many times as they liked before proceeding to the next trial, but could not return to previous trials. The response scale's granularity was altered for each attention check item, such that participants could only respond at the extremes or the middle of the scale. Finally, participants were informed that all data presented was fictitious and were given the option to provide comments on the experiment and describe any strategies used. Average completion time was 19.2 minutes.

6.2.3 Analysis

We conducted analysis using R (R Core Team, 2022) (version 4.2.1). Linear mixed models were built using {lme4} (Bates et al., 2015). Each model was based on a maximal model with by-participant and by-item random effects (Barr et al., 2013), and {buildmer} (Voeten, 2022) was used to identify the final random effects structure, ensuring convergence and removing terms not significantly contributing to explaining variance.

6.2.3.1 Magnitude Ratings

Figure 6.3 shows the distribution of ratings for charts with default axes and extended axes.

```
Picking joint bandwidth of 0.0261
```

Linear mixed-effects modelling revealed that participants awarded higher ratings to charts with default axes, compared to charts with extended axes: $F(1, 152.54) = 44.90$, $p < .001$, partial $\eta^2 = 0.23$.

This model employed a maximal random effects structure, capturing the baseline responses (intercepts) and differences between the two axis settings (slopes) separately for each individual participant and each individual item. The model formula was as follows: r print_formula(e1_mag).

6.2.3.2 Magnitude Ratings and Data Visualisation Literacy

Accounting for differences in data visualisation literacy did not change the significant effect of axis limits: $F(1, 152.57) = 44.95$, $p < .001$, partial $\eta^2 = 0.23$.

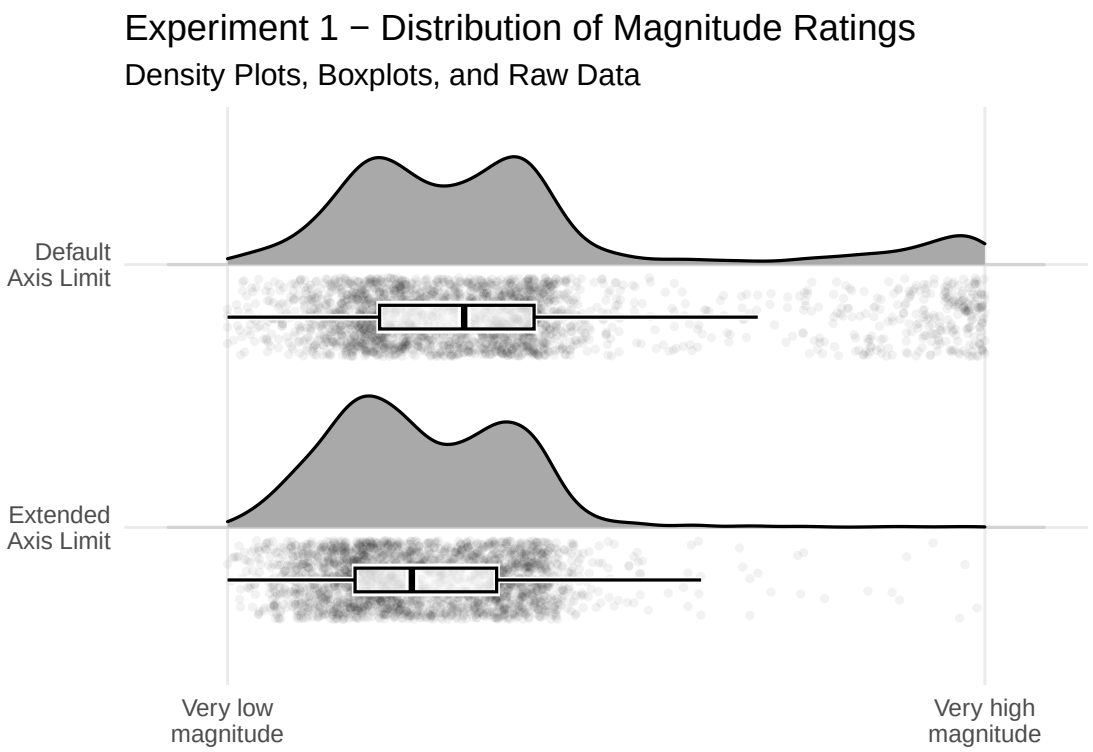


Figure 6.3: The distribution of visual analogue scale ratings in response to default and extended axis limits. Each circle represents a participant's response to an individual bar chart.

6.2.4 Discussion

This experiment explored the consequences of including a graphical cue to a denominator value using bar charts' axes. We observed that plotted values' magnitudes were interpreted as smaller when a default axis limit was used, compared to an axis limit equal to the dataset's denominator value. Therefore, assessments of data were biased by the presence or absence of numerical context in bar charts.

Denominator information informs magnitude judgements. In bar charts with extended axes, denominator information was available through the accompanying text and through axes. Comparison with bar charts employing default axes, where denominator information was *only* available through text, reveals the contribution of the graphical cue to the denominator value. Inconsistency in the differences between conditions illustrates variation in interpretation. The relative similarity of lower magnitude ratings across conditions indicates some attention to denominator information in the absence of a graphical cue. However, some extreme high magnitude ratings suggest that the appearance of tall bars carried the implication of large values. These ratings may indicate a failure to account for denominator information in the absence of a graphical cue. We investigate the role of denominator information further in Experiment 2.

6.3 Experiment 2

6.3.1 Introduction

Experiment 1 found differences in interpretations of data presented using different axes limits. Overall, plotted data were associated with lower magnitudes when presenting using axes which extended to a denominator value. Compared to bar charts with extended axes, charts with no graphical cue to a denominator value elicited a wider variety of responses. This variety appears to reflect differences in how the denominator information supplied in accompanying text is used in magnitude judgements. This raises questions about how text, including denominator information, might influence the interpretation of different chart designs.

By manipulating the presence of denominator information in accompanying text, in addition to manipulating axis limits, we investigate how these textual and graphical cues inform assessments of data. This allows us to understand how different chart designs are interpreted with and without additional numerical context. This 2x2 design also allows us to determine whether we can replicate the findings from Experiment 1 and also gives us the opportunity to explore whether ratings in the absence of denominator information correspond to the previously observed pattern of extreme ratings.

This second experiment requires minor adaptations to materials and procedure. First, there is a risk that highly ambiguous trials without denominator information supplied in text will elicit unreliable random ratings. Therefore, we collect additional confidence ratings to directly

index this aspect of participants' evaluations. This provides a more comprehensive view of participants' cognitive states and interpretations. Second, when denominators are not supplied in text, participants may use denominator values supplied in previous trials to inform their judgements. A limited range of denominators (as in Experiment 1) would artificially diminish uncertainty regarding possible values, inhibiting authentic, spontaneous judgements. Therefore, we expand the range of denominator values in Experiment 2. Third, increasing the number of fillers (which depict relatively high magnitudes) to match the number of experimental items (which depict relatively low magnitudes) will avoid priming effects by ensuring high and low magnitudes seem equally plausible.

6.3.2 Method

6.3.2.1 Materials

We generated bar charts in R using {ggplot2} (version 4.2.1), {tidyverse} (version 1.3.2) and {ggh4x} (version 0.2.3).

Bar charts were generated using the same method as in Experiment 1. We used the same scenarios from Experiment 1, and generated 24 new scenarios for use as additional filler items, thus employing 32 experimental items and 32 filler items. To increase variation across datasets, we employed a wider range of denominators (200, 400, 600, and 800) meaning the plotted values differed from Experiment 1.

We added the word 'surveyed' or 'assessed' to the accompanying text for seven items where the absence of a denominator may have implied that data were collected for the entire population under study. For example, where the study concerned data collected in five towns, the final sentence read 'The graph shows, for each town, how many people *surveyed* used public transport regularly', to avoid the implication that the denominator was equal to an entire town's population. This ensured that the inclusion of denominator values was equally informative across all scenarios.

16% of datasets used were re-generated to ensure that the highest gridline of a default axis did not exceed the highest plotted value.

6.3.2.2 Design

We employed a within-participants 2x2 Latin-squared design with two factors: axis limits (default vs. extended) and denominator information (present vs. absent). Participants viewed 8 different charts for each combination of conditions (32 experimental trials total). The correspondence between scenarios and conditions was counterbalanced using four lists. However, all participants saw the same versions of the 32 filler items and six attention check items.

6.3.2.3 Participants

Participants were recruited using Prolific.co, using the same inclusion criteria as Experiment 1. Additionally, the experiment was not advertised to individuals who completed Experiment 1.

Data were returned by 208 participants. Per pre-registered exclusion criteria, eight participants' submissions were rejected because they answered more than one of six attention check questions incorrectly. Participants whose submissions were accepted received £5.00.

The final sample consisted of 200 participants (60.00% male, 38.00% female, 1.00% non-binary, 0.50% other, 0.50% prefer not to say). Mean age was 33.17 years ($SD = 10.34$)[1]. The mean data visualisation literacy score was 21.72 ($SD = 4.84$), out of a maximum of 30.

This experiment was approved by the University of Manchester's Division of Neuroscience and Experimental Psychology Ethics Committee (ethics code: 2023-11115-28428).

6.3.2.4 Procedure

The procedure was identical to Experiment 1, except for the addition of a confidence rating, where participants were asked 'How confident are you in your response?'. The anchors on the response scale were 'Not very confident' and 'Very confident'. Figure 6.4 shows an example trial.

For attention check items, participants were asked to provide a specific response on the magnitude rating scale, and a random response on the confidence rating scale.

Average completion time was 29.3 minutes.

6.3.3 Analysis

6.3.3.1 Magnitude Ratings

Figure 6.5 shows the distribution of magnitude ratings for charts with default axes and extended axes, where denominators were absent from text, and where they were present.

```
Picking joint bandwidth of 0.038
```

```
Picking joint bandwidth of 0.0292
```

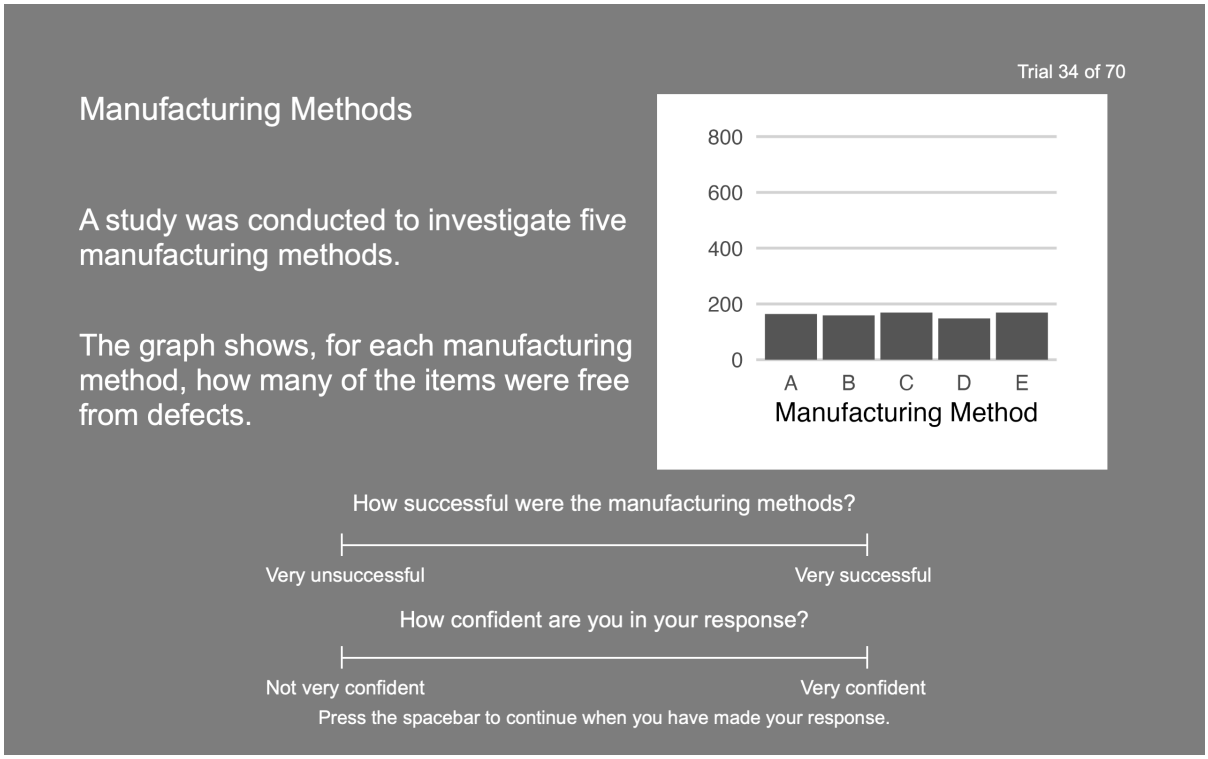


Figure 6.4: An example trial from Experiment 2, showing a bar chart with an extended axis limit. Note the presence of an additional confidence rating scale.

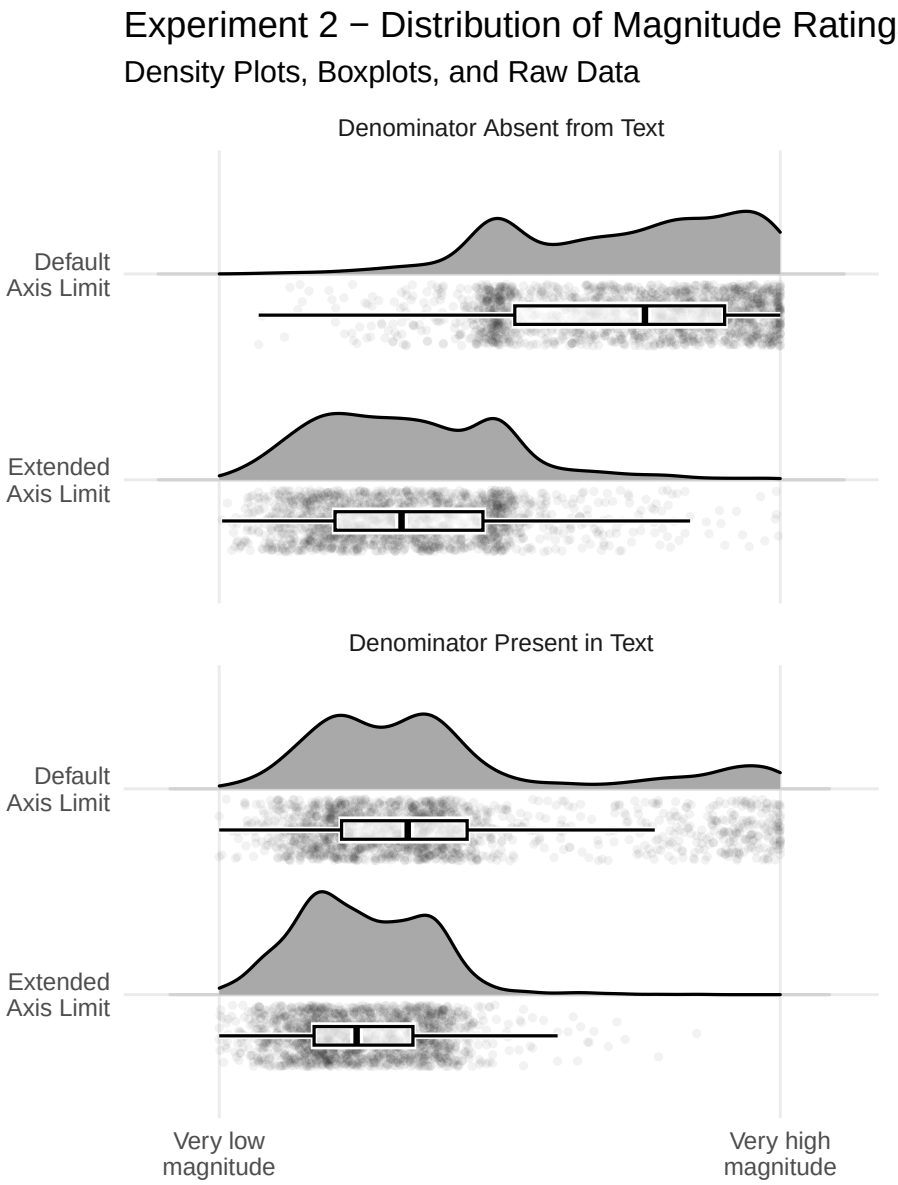


Figure 6.5: The distribution of visual analogue scale ratings in response to default and extended axis limits, shown separately for trials where denominator values were absent from accompanying text (top), and trials where denominator values were present in accompanying text (bottom). Each circle represents a participant's response to an individual bar chart.

A mixed effects model revealed that charts with default axes elicited higher ratings compared to charts with extended axes ($F(1, 198.22) = 311.45$, $p < .001$, partial $\eta^2 = 0.61$) and charts not accompanied by a denominator in text elicited higher ratings than those accompanied by a denominator ($F(1, 82.23) = 380.50$, $p < .001$, partial $\eta^2 = 0.82$).

Crucially, there was also a significant interaction between axis limits and denominator information: $F(1, 5{,}741.16) = 1{,}540.86$, $p < .001$, partial $\eta^2 = 0.21$. Figure 6.6 plots this interaction.

Pairwise comparisons produced using {emmeans} (Lenth, 2021) revealed that charts with extended and default axes were rated differently when the denominator was present, replicating the effect from Experiment 1 ($z = -9.19$, $p < .001$), and also when the denominator was absent ($z = -25.35$, $p < .001$). Therefore, the interaction indicates that the magnitude of influence exerted by a bar chart's axis varied according to whether the denominator was present or absent.

This model employed by-participant and by-item random effects. For each participant, there were random intercepts, plus random slopes for axis settings and denominator information. For each item, there were random intercepts, plus random slopes for denominator information. The model formula was as follows: r print_formula(e2_mag).

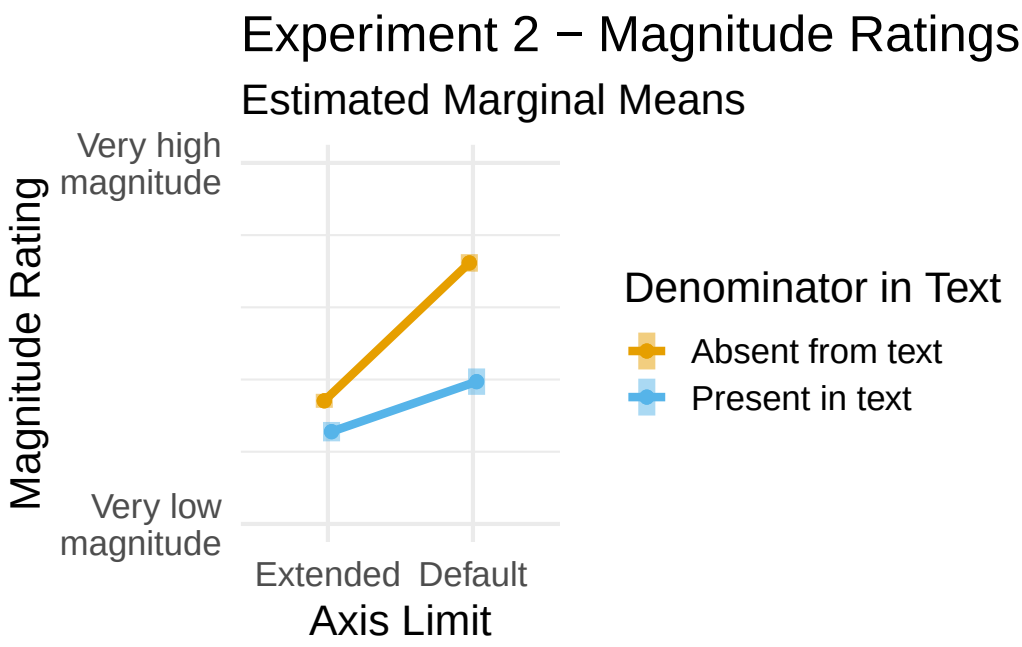


Figure 6.6: The interaction between axis limits and denominator information, for magnitude ratings. Estimated marginal means are generated by the linear mixed model used in analysis. Translucent bars show 95% confidence intervals.

[1]Age data was unavailable for 2 participants

6.3.3.2 Magnitude Ratings and Data Visualisation Literacy

Accounting for differences in data visualisation literacy did not change the significant interaction: F(1, 5,741.14) = 1,540.85, p < .001, partial η^2 = 0.21., the main effect of axis limits (F(1, 198.24) = 311.46, p < .001, partial η^2 = 0.61) or the main effect of denominator information (F(1, 82.29) = 380.60, p < .001, partial η^2 = 0.82).

6.3.3.3 Confidence Ratings

```
Picking joint bandwidth of 0.0711
```

```
Picking joint bandwidth of 0.0288
```

Figure 6.7 shows the distribution of confidence ratings for charts with default axes and extended axes, where denominators were absent from text, and where they were present.

A mixed effects model revealed a main effect associated with axis limits (F(1, 199.00) = 5.97, p = .015, partial η^2 = 0.03), a main effect associated with denominator information (F(1, 198.99) = 184.93, p < .001, partial η^2 = 0.48) and an interaction F(1, 5,799.00) = 27.74, p < .001, partial η^2 < .01. This interaction consisted of a difference between extended and default charts when the denominator was absent from text (z = -4.69, p < .001), but no difference between charts when the denominator was present (z = 0.42, p = .988). However, it is clear from Figure 6.8, as well as the partial η^2 values, that the effect sizes associated with axis limits and the interaction are trivial.

```
# A tibble: 1 x 2
  terms       slopes
  <chr>       <chr>
1 participant denominator + axis
```

This model employed by-participant and by-item random effects. For each participant, there were random intercepts, plus random slopes for axis settings and denominator information. For each item, there were random intercepts, plus random slopes for denominator information. The model formula was as follows: r print_formula(e2_con).

6.3.3.4 Confidence Ratings and Data Visualisation Literacy

Accounting for differences in data visualisation literacy did not change the pattern of results. There was a main effect associated with axis limits (F(1, 199.01) = 5.97, p = .015, partial η^2 = 0.03) and a main effect associated with denominator information (F(1, 198.99) = 184.93, p < .001, partial η^2 = 0.48) and an interaction F(1, 5,798.99) = 27.74, p < .001, partial η^2 < .01.

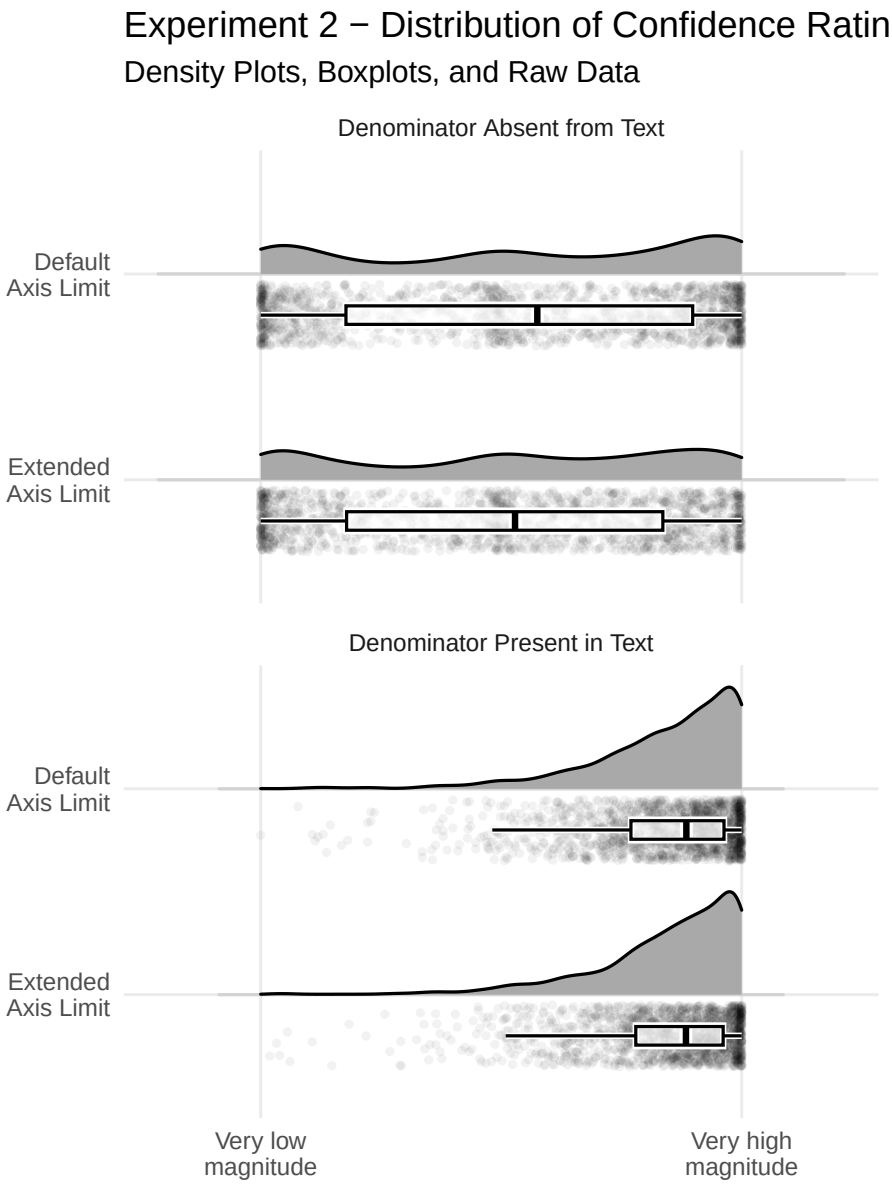


Figure 6.7: The distribution of confidence ratings in response to default and extended axis limits, shown separately for trials where denominator values were absent from accompanying text (top), and trials where denominator values were present in accompanying text (bottom). Each circle represents a participant's response to an individual bar chart.

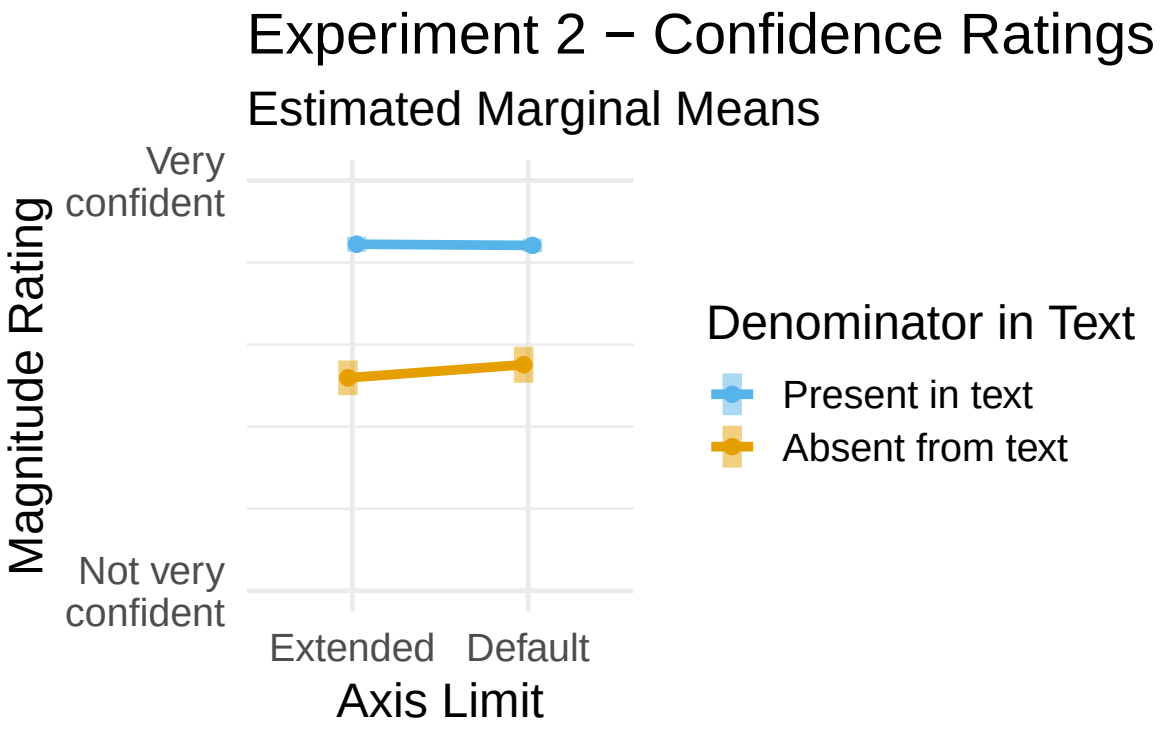


Figure 6.8: The interaction between axis limits and denominator information, for confidence ratings. Estimated marginal means are generated by the linear mixed model used in analysis. Translucent bars show 95% confidence intervals.

6.3.4 Discussion

This experiment manipulated bar charts' axis limits and the presence of denominator values in accompanying text. The results demonstrate that values presented in charts with default axis limits are associated with higher magnitude judgements than charts with extended axes, in both the presence and absence of denominator information. However, the absence of denominator information amplifies this bias. These results also suggest that extreme high magnitude ratings for default charts in the presence of a denominator value may be driven by a failure to incorporate that value into reasoning. Finally, confidence in judgements is reliably affected by the inclusion of denominator information in text.

6.4 General Discussion

Axis limits can be easily manipulated in common data visualisation software, in order to include a visual cue to denominator information. However, these defaults are based on plotted data only, so often omit denominator information. We demonstrate that plotted values' magnitudes were interpreted as smaller for bar charts with axes that extended to the denominator value, rather than those employing default axis settings. The influence of axis limits was particularly large when no denominator information was included in the text accompanying a chart. This provides insight into the cognitive process involved in magnitude judgements, indicating that denominator information is an important aspect in interpreting plotted values in bar charts.

In Experiment 1, we identified a framing effect, wherein charts with axes that accommodated a denominator value elicited smaller magnitude judgements compared to charts with default axes. In both conditions, the denominator was explicitly presented in the text. Additionally, some extreme responses in the default condition appeared to represent a disregard for denominator information. Given the apparent importance of this information, we conducted another experiment in order to examine the denominator's role in the cognitive processing of magnitude. We examined how interpretations were affected by the absence of denominator information, thus capturing how this information was incorporated differently across chart designs.

Experiment 2 makes several additional contributions. First, it replicated the main effect from Experiment 1. That is, we observed a propensity to interpret magnitudes as smaller when values were shown with an extended axis, rather than a default axis. Second, it illustrates the impact of including the denominator in accompanying text. This cue affects viewers' interpretations differently depending on whether a chart's axis also incorporates the same value. Without denominator information in text, the magnitude of values plotted using default axes can be ambiguous. Accordingly, drastically higher ratings in the absence of denominator information illustrate the denominator's role in reducing ambiguity. Interpretation of values plotted using extended axes was affected to a lesser extent by the denominator's absence. Thus, the impact of a bar chart's axis is greater when not accompanied by a denominator.

This suggests axis limits facilitate recognition of denominator information when interpreting magnitudes.

Third, Experiment 2 replicated the pattern of responses observed in Experiment 1 for charts with default axes and accompanying denominator information. This pattern consists of a small number of higher magnitude ratings, in contrast to the general tendency for lower magnitude ratings. Figure 6.5 reveals a close resemblance between the distribution of these higher ratings and the overall distribution of ratings for default charts *without* accompanying denominator information. This suggests that these extreme ratings may share a cause. Unusually high responses in the presence of denominator information likely result from failure to account for the denominator and a subsequent reliance on the chart's appearance. The analogous responses to charts without accompanying denominators (Experiment 2) can be considered an experimentally-induced instance of the same effect.

Fourth, additional ratings collected in Experiment 2 provide insight into participants' confidence. Although analysis of these ratings indicated a main effect of axis limits and an interaction between denominator information and axis limits, the minuscule effect sizes cast doubt over the practical significance of these effects. In spite of this, absence of a denominator clearly lowered confidence. This suggests that participants were hesitant to form magnitude judgements based solely on a bar chart's appearance. Inclusion of a denominator value in text was preferred regardless of graphical cues to context.

6.4.1 Relationship to Prior Work

Our focus on judgements of values' magnitudes is noteworthy because the vast majority of related work has explored participants' judgements of *differences between* values (Correll et al., 2020; Garcia-Retamero and Galesic, 2010; Okan et al., 2020; Stone et al., 2003; Witt, 2019; Yang et al., 2021; **okan_individual_2012?**; **okan_designing_2018?**; **stone_salience_2018?**). Responses to questions about values' magnitudes have often been obscured through inclusion in composite measures (e.g., **okan_designing_2018?**), or have been collected to assess comprehension, rather than interpretation (e.g., Garcia-Retamero and Galesic, 2010). As (**stone_effects_2015?**) discuss, failing to consider interpretations of values' magnitudes reflects two issues. First, neglecting values' magnitudes overlooks a relevant aspect of numerical information. Second, neglecting participants' *interpretations* limits insight into decision-making, which is not simply governed by accurate retrieval of information (see **reyna_theory_2008?**).

Whereas much prior research has been limited to interpretation of risk information (Garcia-Retamero and Galesic, 2010; Okan et al., 2020; Stone et al., 2003; **okan_individual_2012?**; **okan_designing_2018?**; **stone_salience_2018?**; **stone_designing_2017?**; **stone_effects_2015?**), we demonstrate that biases in interpretation extend to a wide range of non-risk scenarios. This provides confidence that these findings are widely applicable, and using *multiple* trials per participant enhances statistical power. When generating charts that do not

include denominator values, previous experiments (Garcia-Retamero and Galesic, 2010; **okan_designing_2018?**) appear to have employed arbitrary axis limits. By employing axis limits based on {ggplot2}'s default settings, our materials reflect a common practice, enhancing our experiment's ecological validity.

We contribute to a large body of evidence illustrating biases in the interpretation of numerical information, specifically *framing effects* (Tversky and Kahneman, 1981). Our results are consistent with research demonstrating that manipulating bar charts' axis limits influences interpretation of plotted values (Garcia-Retamero and Galesic, 2010; **okan_designing_2018?**). Furthermore, (**okan_designing_2018?**) found that participants' perceptions of risk were influenced more by bar charts' axis limits when labels containing numerators and denominators were excluded. Similarly, we observed that interpretations of magnitude were influenced more by bar charts' axis limits when denominator values were omitted from accompanying text.

A previous study exploring interpretation of magnitude in bar charts observed different responses according to whether stacked bars or blank space conveyed alternative outcomes (**stone_designing_2017?**). We demonstrate that manipulating the amount of blank space above bars can elicit different magnitude judgements, without plotting alternative outcomes explicitly. Earlier work investigating (in)consistency in the formats used to display numerators and denominators is also relevant. (**stone_effects_2015?**) found that displaying a value using icons, accompanied by a denominator in text, increased impressions of that value's magnitude, compared to when both the value and denominator were presented in text. We too found higher ratings when values displayed using bars were only accompanied by a denominator in text, compared to when a corresponding graphical cue to the denominator value was also present.

According to Fuzzy Trace Theory, different interpretations can arise due to different gist-level representations, despite accurate comprehension of presented values (**reyna_theory_2008?**). Therefore, access to denominator information in accompanying text did not prevent our chart designs influencing judgements. Encoding of gist is reported to be influenced by the appearance of graphical elements (**reyna_theory_2008?**). This suggests that the taller bars in our default axis conditions were responsible for impressions of greater magnitude, compared to shorter bars in our extended axis conditions. That charts with extended axes elicited lower magnitude ratings is also consistent with Stone et al.'s (**stone_salience_2018?**) proportional reasoning account, which suggests that part-to-whole displays facilitate processing of a larger numerical context.

We did not find evidence that data visualisation literacy affected our results. This is contrary to the finding that data visualisation literacy predicted the efficacy of using icon arrays to reduce denominator neglect (**okan_individual_2012?**). However, this is consistent with the finding that the impact of manipulations like ours (axis range, numerical labels) are independent of data visualisation literacy (**okan_designing_2018?**). This measure may capture whether people have sufficient ability to extract information from a visualisation, rather than predicting

the degree to which they will be influenced by subtler design choices (Yang et al., 2021). Numeracy is associated with decreased sensitivity to framing effects (**peters_numeracy_2006?**), so this may be a better candidate for understanding individual differences in response to visualisation design.

6.4.2 Implications

When conveying values' magnitudes, both axis limits and accompanying text warrant consideration from data visualisation designers. A bar chart produced using default settings is not equivalent to a bar chart with an axis that incorporates a denominator value. Extending an axis in this manner increases consistency in judgements and may provide insurance against individuals who fail to account for accompanying denominator information. Similarly, where constraints prevent inclusion of a denominator value in text, an extended axis should facilitate viewers' recognition of this numerical context. We observed that confidence ratings were consistently high in the absence of a denominator in text, despite use of an extended axis. Explicitly providing denominator values in text, regardless of graphical cues, would therefore promote viewers' confidence in their judgements.

It is also worth considering situations which may accentuate the observed bias. High cognitive load exacerbates the numerosity bias (**pelham_easy_1994?**), therefore may also interfere with magnitude judgements. Even when denominator information is supplied in text, high cognitive load could prevent this information from informing interpretations. This would likely increase reliance on bar charts' appearances, like in Experiment 2. Additionally, assuming that an audience has knowledge of a dataset's denominator may increase biases in individuals who are unfamiliar with the topic.

6.4.3 Limitations and Future Work

This work is concerned with visualisations intended to convey plotted values' magnitudes. However, design considerations will differ when conveying *differences* between values. In this case, axis ranges should be determined by the magnitude of the differences (Correll et al., 2020; Witt, 2019; Yang et al., 2021). Consequently, our recommendations are not relevant for all communicative scenarios. However, maintaining awareness of the implication of plotted values' magnitudes may help avoid misinterpretation of data, even if this type of judgement is not a primary concern.

Our experiments apply best to controlled scenarios, such as surveys and experiments where all plotted values share the same denominator. These findings may also extend to datasets with unequal denominators, if bars are used to depict proportions or percentages, permitting use of a single meaningful axis limit. However, this design will not be suitable for plotting other types of dataset. We also acknowledge that proportions are not the only factor influencing magnitude judgements: subject matter is also likely to inform assessments. For example, bars

clearly depicting one or two hours spent on administrative tasks within a 35-hour work week will still elicit some differences of opinion regarding whether these values are high or low.

All materials were produced using {ggplot2}. Therefore, our conclusions about default axis limits only pertain to bar charts created using this package's settings, though we expect other visualisation libraries' default settings to elicit similar responses, due to similarity in their behaviour. For uniformity in our materials, we only employed default charts where the highest gridline was positioned below the highest value, since this was the most common visual arrangement. We did not examine the minority of cases where the highest gridline exceeds the highest value. Whether this influences magnitude judgements could be explored in future experiments. In addition, future work should employ decision-making tasks to quantify the impact of axis limits on applied judgements.

6.4.4 Conclusion

In two experiments, we generated evidence on the effects of default and extended axis limits, illustrating the influential role of denominators in gauging magnitude. We provide insight into the cognitive processes involved in interpreting plotted values' magnitudes in bar charts and offer recommendations for facilitating judgements. Framing effects demonstrate the power of presentation choices on the interpretation of numbers.

6.5 References

References

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**:255–278. doi:10.1016/j.jml.2012.11.001

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* **67**. doi:10.18637/jss.v067.i01

Correll M, Bertini E, Franconeri S. 2020. Truncating the Y-Axis: Threat or Menace?Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM. pp. 1–12. doi:10.1145/3313831.3376222

Garcia-Retamero R, Cokely ET, Ghazal S, Joeris A. 2016. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making* **36**:854–867. doi:10.1177/0272989X16655334

Garcia-Retamero R, Galesic M. 2010. Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine* **70**:1019–1025. doi:10.1016/j.socscimed.2009.11.031

Lenth RV. 2021. Emmeans: Estimated Marginal Means, aka Least-Squares Means.

Okan Y, Stone ER, Parillo J, Bruine de Bruin W, Parker AM. 2020. Probability Size Matters: The Effect of Foreground-Only versus Foreground+Background Graphs on Risk Aversion Diminishes with Larger Probabilities. *Risk Analysis* **40**:771–788. doi:10.1111/risa.13431

Pandey AV, Rall K, Satterthwaite ML, Nov O, Bertini E. 2015. How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion TechniquesProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. Seoul, Republic of Korea: ACM Press. pp. 1469–1478. doi:10.1145/2702123.2702608

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* **51**:195–203. doi:10.3758/s13428-018-01193-y

R Core Team. 2022. R: A Language and Environment for Statistical Computing.

Stone ER, Sieck WR, Bull BE, Frank Yates J, Parks SC, Rush CJ. 2003. Foreground:background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes* **90**:19–36. doi:10.1016/S0749-5978(03)00003-7

Tversky A, Kahneman D. 1981. The Framing of Decisions and the Psychology of Choice. *Science* **211**:453–458. doi:10.1126/science.7455683

Voeten CC. 2022. Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects.

Wickham H. 2016. ggplot2. New York, NY: Springer Science+Business Media, LLC.

Witt JK. 2019. Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis. *Meta-Psychology* **2**:1–20. doi:10.15626/MP.2018.895

Yang BW, Vargas Restrepo C, Stanley ML, Marsh EJ. 2021. Truncating Bar Graphs Persistently Misleads Viewers. *Journal of Applied Research in Memory and Cognition* S2211368120300978. doi:10.1016/j.jarmac.2020.10.002

Postscript to Chapter 5

Overview

The final study in this thesis was designed to examine the role of denominator information in the interpretation of absolute magnitude in data visualisations. A data visualisation's y-axis may be expanded to incorporate not just plausible values, but a *denominator* from the underlying dataset. When this graphical cue is accompanied by a textual cue to denominator information, its influence on interpretations of magnitude is diminished, compared to when no textual cue is present. Thus, this study revealed that accompanying contextual information, not just visualisation design, informs subjective judgements of magnitude.

Rationale for Experimental Design

This study aimed to explore how axes which incorporate denominator values affect magnitude judgements. Manipulating upper axis limits in bar charts necessarily alters the length of the bars. Therefore, the differences between conditions in this study are less subtle than in the first study in this thesis, which employed dot plots. However, changes in the length of bars themselves should not necessarily be considered a confound. It is impossible to avoid these changes, therefore the resulting *overall* appearance of the visualisation should be considered part of the manipulation. A lack of precise experimental control over separate elements does not preclude valuable insight into the cognitive processing of magnitude in bar charts.

This study employed default axis limits from a popular visualisation tool in half of experimental trials. As default settings are frequently adopted by data visualisation designers, this reduces artificiality in stimuli, enhancing the experiment's ecological validity. In addition, adopting default settings means this study's conclusions are directly applicable to common visualisation designs, assisting the development of relevant considerations for designers.

Summary of Findings

In the first experiment in this study, participants rated magnitudes as lower when axes terminated well above plotted data, compared to just above plotted data. The former condition provided a graphical cue to the relevant denominator value, whereas the latter condition,

which used the default axis setting, did not. When there was no graphical cue to the denominator, a proportion of extreme high magnitude ratings reflected substantial differences in the interpretation of charts in the two conditions. The second experiment revealed that substantial differences in interpretation regularly occur when accompanying text omits denominator information, eliciting a greater bias than when denominator information is included. This interaction demonstrates that awareness of the numerical context of plotted values diminishes the influence of axis limits on interpretations of magnitude.

Risk perception research on axes and denominators in bar charts has observed that including labels with numerical values decreased the impact of graphical depictions (Okan et al., 2018). The final study in this thesis corroborates this claim by manipulating numerical values in accompanying text. Prior work has measured risk understanding (Garcia-Retamero and Galesic, 2010), or combined absolute magnitude ratings with other judgements to create a general risk perception measure (Okan et al., 2018). These studies have not examinated absolute magnitude judgements in isolation. Therefore, the final study in this thesis expands upon this work by examining absolute magnitude judgements. This study also explores responses to a different topic, which does not concern risk.

Relationship to the Studies in Chapter 3 and Chapter 5

This study concludes the empirical investigations in this thesis by revealing how axis limits and additional cues interact to influence magnitude judgements. This demonstrates that considering contextual information is important for understanding the effects of design choices on interpretations of magnitude. However, it also provides insight into the differences in the two preceding studies regarding the *extent* to which axis limits influenced magnitude judgements. The first study observed relatively small differences in participants' ratings between conditions, whereas the second study observed much larger differences. In light of this final study's findings, it seems likely that differences in contextual information may have contributed to these contrasting results. In the second study, use of fictitious standardised units may have increased reliance on axis limits as a source of information when assessing magnitude. Consequently, this resulted in increased bias in judgements.

This study used denominator values as upper axis limits. This contrasts with preceding studies, where the data (i.e., relative frequencies, measurements) do not have corresponding denominators. Therefore, the final study extends the line of enquiry to examine a different approach to setting axis limits, based on the underlying dataset. By incorporating a denominator, an axis presents the entire range of possible values, rather than a range of plausible values. Viewers' inferences about magnitude may differ in these two cases. Denominators may increase awareness of whether plotted values are *numerically* high or low, whereas plausible values may increase awareness of whether plotted values are *comparatively* high or low, in the context of other relevant magnitudes. Alternatively, viewers may not be sensitive to this difference. This

would be an interesting topic for future work, providing greater insight into the framing effects elicited by axis limits.

Both this study on bar charts and the second study on choropleth maps explore approaches to designing data visualisations beyond typical defaults. They suggest that not all data visualisation designs are suitable for communicating all aspects of data. Therefore, considering alternatives to the default options may be useful when conveying absolute magnitude. Atypical approaches to data visualisation design may provide clarity for viewers.

7 Conclusion

Data visualisations can facilitate comprehension of data by exploiting humans' perceptual and cognitive capabilities. Systematically manipulating data visualisation designs in experimental studies can help reveal cognitive mechanisms involved in interpretation. This, in turn, can inform recommendations for effective design, leading to successful communication of data.

Data visualisations can illuminate many aspects of a dataset, from correlations to distributions. Viewers may also form impressions of values' absolute magnitudes: how large or small values are. To understand this process, it is necessary to examine how graphical cues contribute to a viewer's mental representation. Techniques from experimental psychology provide a means of investigating this aspect of cognitive processing.

7.1 Research Objectives

The aim of this thesis was to investigate interpretations of absolute magnitudes in data visualisations. Specifically, this work sought to understand how *axis limits* may inform these judgements, with a focus on viewers' *gist*-level representations of data, rather than their apprehension of precise values. The experiments were designed to provide insight into the cognitive mechanisms involved in assessing magnitude, exploring how various aspects of presentation which affect the processing of this information (e.g., axis orientation, colour encoding, accompanying information). Exploring various visualisation formats and various types of data yields generalisable findings. In addition, examining conventions and default settings contextualises this work within current design practices.

7.2 Main Findings

The three sets of experiments in this thesis consider the same overarching question. However, each provides a unique contribution.

The study in Chapter 3 manipulated axis limits in dot plots. This manipulation altered the position of plotted values, so that they appeared at high or low positions. The first experiment in this study established that interpretations of values' magnitude were affected by axis limits. In the two subsequent experiments, I explored whether judgements were influenced by the relative positions of data points within axis limits, or the absolute physical positions

of data points. An inverted axis orientation made it possible to distinguish between these potential explanations. Despite conventions around displaying magnitude, participants rated magnitudes as greater when values appeared closer to the bottom of the visualisation, which was associated with higher numerical values. This illustrates that the *relative* positions of data points within axis limits inform magnitude judgements.

The study in Chapter 4 manipulated the limits of colour legends accompanying choropleth maps. Rather than altering the appearance of plotted values, this manipulation only modified the range of numerical values and corresponding colours in colour legends. Participants interpreted magnitude as greater when the corresponding positions of the plotted values were located at a colour legend's upper limit, rather than its halfway point. This illustrates that magnitude judgements are informed by the relationship between data points and axis values, not simply the appearance of plotted data points.

The study in Chapter 5 manipulated axis limits in bar charts. This manipulation compared upper axis limits based on ggplot2's defaults to upper axis limits corresponding to denominator values in datasets. The first experiment established that participants interpreted magnitude as greater in the latter condition. In the subsequent experiment, I explored how awareness of numerical context affects the use of graphical cues to magnitude. When text containing dataset's denominator value was removed, the degree to which axis limits informed magnitude judgements increased. This illustrates that knowledge about data points' numerical context limits the influence of axis limits on interpretations.

The stimuli used in these studies reveal different approaches to conveying magnitude using axis limits. *Absolute frequencies* plotted in bar charts (Chapter 3) were associated with denominators, so upper axis limits were intrinsically linked to datasets. Conversely, *measurements* plotted in choropleth maps (Chapter 2) upper axis limits were not determined by denominator values. The *relative frequencies* plotted as percentages in dot plots (Chapter 1) could not be less than 0 or more than 100, so axis limits were subject to logical constraints. Therefore, designers must also consider the type of data at hand when identifying suitable axis limits for data visualisations. What constitutes *context* is not uniform.

Overall, this work provides consistent evidence of the influence of *framing effects* on magnitude judgements in data visualisations. Axes served as a frame of reference biasing the interpretation of numbers, thus mental representation of information was influenced by its presentation (Tversky and Kahneman, 1981; Kahneman, 1992). Furthermore, according to models of graph comprehension (Pinker, Carpenter and Shah), visual patterns are encoded prior to comprehension of numerical values and contextualisation of information. Therefore, first impressions elicited by a data visualisation may prejudice overall interpretations (Pandey et al. 2015; Yang et al. 2021).

7.2.1 Data Visualisation Literacy

In each study, I also investigated the role of participants' data visualisation literacy levels in their interpretations of absolute magnitude. Data visualisation literacy levels did not account for variation in responses to axis manipulations. This accords with prior research on axis truncation, which found that the same subjective data visualisation literacy measure was not associated with the degree of bias in judgements of relative differences (Yang et al., 2021). A similar finding has also been observed using an objective data visualisation literacy measure (Okan et al., 2018). Another experiment on the same issue, using yet another literacy measure, did observe that data visualisation literacy predicted bias (Driessen et al., 2022). However, a lack of diversity in the sample's literacy levels diminished confidence in the utility of this measure.

In this thesis, variability in participants' responses was evident. This suggests that the data visualisation literacy measure employed was not best placed to capture these differences. This measure may register basic skills for comprehending visualisations, rather than susceptibility to visualisation design (Yang et al., 2021). Ge et al. (2023) suggest that critical thinking skills predict the degree of bias elicited by deceptive designs, and produced a data visualisation literacy measure designed to capture this capacity. However, its considerable length makes it unsuitable for use as a covariate in experimental studies. A shorter literacy measure of critical thinking in comprehension of data visualisations would be a valuable tool for future research.

7.3 Reproducibility

Historically, research on data visualisations has rarely been aligned with the principles of open and reproducible science (Haroz, 2018). Sharing data, code, and materials provides transparency and allows for independent authentication of published findings (Klein et al., 2018). The empirical work in this thesis was conducted with a focus on ensuring reproducibility. In addition to publicly sharing all data and code, I used containerisation software to capture the computational environment used for data processing, analysis and visualisation. This allows the manuscript for each chapter to be reproduced from the raw data, enhancing the transparency and sustainability of this work. The adoption of these open and reproducible working practices in future work would increase the credibility of data visualisation research.

7.4 Contributions and Implications

This thesis addresses an important yet neglected topic in data visualisation research. Empirical results demonstrating viewers' sensitivity to *absolute magnitude* present new considerations for data visualisation design. Experiments were designed to generate robust evidence, extending

prior work on this topic (Sandman et al., 1994) by increasing statistical power and investigating a range of visualisation formats. Furthermore, a focus on identifying the underlying cognitive mechanisms provides insight into how graphical cues to absolute magnitude are processed and how additional factors influence interpretations. This thesis expands and strengthens the evidence base on the interpretation of absolute magnitude in data visualisations.

This thesis challenges the notion of 'objectively correct' data visualisation design and presents a more nuanced perspective. Certainly, misrepresenting numerical values in data visualisations is an invalid practice. However, the axis manipulations explored in this thesis do not alter the numerical values presented, but the *subjective* judgements of these values. As Correll et al., (2020) argue, because these judgements depend on context, there is no *inherently truthful* axis setting which avoids influencing viewers. This thesis demonstrates that designers should be aware of what their visualisation implies about plotted data, and make design choices which faithfully portray the relevant characteristics of the plotted dataset. Similarly, Sandman et al. (1994) also assert that *intentional* design necessarily involves considering suitable interpretations. In response to the concern that this is a manipulative practice, they argue *"[I]t is no solution to renounce the intention and keep the bias"* (Sandman et al., 1994; pg. 44).

A further contribution of this work is the argument that visualisation design conventions and default software settings can suppress graphical representations of absolute magnitude. Typical choropleth map design depicts spatial variability (i.e., *relative* magnitude) by mapping the maximum and minimum values to the colour legend's upper and lower bounds, respectively. This may inappropriately imply that values near the extremes have particularly high or low *absolute* magnitudes. Rejecting this convention can assist in conveying that *all* values should be considered high or low. Similarly, default upper axis limits can inflate absolute magnitude judgements in bar charts, relative to axis limits based on denominator values, particularly in the absence of other contextual cues. Despite this, this default setting is necessary because axis limits cannot automatically account for denominators, which are not necessarily available within a dataset. Regarding visualisation of absolute magnitude, issues may arise from automatic deference to conventions and defaults, not the conventions and defaults by themselves. Considering alternative approaches to visualisation design contributes to an appropriate presentation of data.

The present work discusses *bias*, but use of this term does not necessarily warrant a negative perspective. Unquestioning acceptance of the message conveyed by a visualisation is obviously detrimental. However, *failing* to account for visual cues to the magnitude of values could also be considered a limitation in cognitive processing. Sensitivity to the relative positions of values within an axis range and the ability to weight this information in accordance with other knowledge constitutes a powerful capability. Appropriate use of graphical cues to magnitude provides designers with an opportunity to leverage this bias for effective communication.

Sensitivity to norms in communication is another relevant factor for interpreting the findings of this thesis. Gricean pragmatic principles (Grice, 1975) suggest that contributions in conversation are expected to be truthful, relevant, clear, and sufficiently informative. Under the assumption that visualisation design follows these principles, a chart's implied message will

guide viewers' judgements. Values' relative positions within axes would be considered relevant and reliable information, thus informing inferences about magnitude.

7.5 Limitations and Future Directions

This work does not suggest that conveying absolute magnitude will be appropriate in all scenarios. Other aspects of data may be considered more pertinent. Conflicting graphical cues may prevent designers conveying magnitude *in addition to* other aspects. For example, a truncated axis may clearly display relative differences between data points but omit context illustrating the absolute magnitudes of values. Alternatively, an extended axis may clearly display absolute magnitudes but obscure the relative differences between values. A possible compromise may involve generating two visualisations: one to illustrate absolute magnitudes and another to promote discrimination between values. A similar approach is discussed in work on axis truncation (Correll et al. 2020), and has been reported to benefit viewers (Ritchie et al., 2019). Future work could investigate the utility of this approach for conveying both absolute magnitudes and relative differences.

Eliciting judgements using rating scales is an established method for measuring interpretations of visualisations (see Correll et al. 2020, Witt, 2019, Yang et al. 2021, Pandey et al. 2015, Stone et al. 2017, 2018, Okan et al. 2018, Okan et al. 2020, Sandman et al. 1994). The measures used in this thesis sought to ensure that participants' responses reflected message-level interpretations. Therefore, all scales captured ratings pertaining to the information presented, not simply the graphical elements. However, future work employing decision-making measures would complement this work. First, this would provide insight into how the use of axis limits in conveying magnitude influences *behaviour*. Second, this would more effectively illustrate the strength of effects observed. Measuring responses in terms of well-defined units (e.g., monetary value) can capture the magnitude of bias in a more concrete manner. Other experiments in the field offer inspiration for effective and comprehensible decision-making tasks (e.g., Kale et al. 2020, Bancilhon et al., 2019).

Further investigation of the different cues to context in interpretations of absolute magnitude would provide a more detailed account of the cognitive mechanisms. There are two subtly different ways in which an axis could influence judgements, which have not been fully differentiated in this thesis. The first cue is the numerical location of a data point on an axis, meaning a data point may appear *numerically* large or small. The second cue is the physical location of a data point on an axis, meaning a data point may appear *comparatively* large or small. Further experiments could measure the relative contributions of these two cues by manipulating the numerical range of an axis whilst maintaining the approximate physical locations of data points. Alternatively, experiments could compare axes which contain only *plausible* values to axes which display the entire range of *possible* values. These investigations would reveal the degree to which an axis provides context by implying a relevant range of numerical values, versus providing a purely visual cue. The second experiment in this thesis investigated the role

of numerical labels, and found that they were not used as a cue to magnitude. However, use of unfamilar units prevented participants from considering numerical values in a typical manner. Future work comparing numerical and visual cues using *familiar* data would be a valuable contribution to this area. More generally, this thesis focuses on how axis limits inform magnitude judgements, but many other factors may also influence interpretations, from numerical units to prior beliefs. A comprehensive account of the cognitive processing of magnitude in data visualisations will also need to consider additional biases and cognitive mechanisms.

An algebraic approach to visualisation design suggests that material changes in data should be reflected accordingly in a visualisation (Kindlmann and Scheidegger, 2014). However, this capacity requires the ability to convey material changes in the *specific aspect* of a dataset of interest. For example, a design supporting comprehension of relative differences may be poorly suited to conveying other messages. This thesis demonstrates that conventional data visualisation formats present an opportunity for communicating an often-overlooked aspect of a dataset: absolute magnitude. Just as adjusting axis limits influences interpretations about absolute magnitude, other design choices may help convey other messages. Graphical cues to other under-explored features of data, such as noise or numerosity, would benefit from future investigation. Developing an understanding of cognitive processes involved in interpreting these features would inform effective design.

7.6 Closing Remarks

Context makes numbers meaningful. In data visualisations, context can be established not only by plotted data, but also extrinsic details. By determining the physical characteristics of a visualisation, designers may also influence the message conveyed about the data. Viewers' sensitivity to the framing of data illustrates the power of design choices to mislead or enlighten.

7.7 New

Go through intro and select relevant prior work - one sentence on each study Kindlmann and Scheidegger - Affordances in assessing magnitude if there is some numerical context.

References