Assessing Individual Differences in Graphical Perception

by

Russell Davis

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

degree of Master of Science

in

Data Science.

April 2023

APPROVED:

Professor Lane T. Harrison, Thesis Advisor

Professor Randy Paffenroth, Thesis Reader

Professor Elke Rundensteiner, Head of Department

Abstract

Data is often presented in the form of graphical visualizations rather than as raw data, with encodings frequently chosen to optimize for accuracy of interpretation by the audience. Visualization guidelines have been drafted to help designers select visualizations that optimize the reader's ability to understand it. However, most visualization guidelines are derived from studies that focus on population-level rankings of accuracy, disregarding possible individual differences in peoples ability to interpret visualizations. This thesis considers variations in individual performance by replicating and extending Cleveland & McGill's widely-studied visualization experiment. By implementing Bayesian multilevel regression, we generate models that facilitate exploration of differences between individual participants and between each visualization type. We confirm that a substantial percentage of individuals show accuracy judgments that deviate from the canonical rankings. We discuss between-individual differences as a relevant factor for design effectiveness, with respect to its capacity to highlight individual variation from population-level aggregates, and with respect to its ability to differentiate factors to between-individual variation; implications for research focused on providing guidance to visualization designers; and proposed further modifications to research in the mode of Cleveland & McGill.

Acknowledgements

I would like to express my deep gratitude to my advisor, Professor Lane Harrison, who was willing to talk to a 1st Lieutenant from out in Arizona about visualization and has given me many hours of mentoring and stimulating discussion.

I would like to thank my thesis reader, Randy Paffenroth, who gave me valuable feedback and encouragement.

I would like to thank Dr. Mi Feng, a member of the WPI VIEW Lab, for her help building the experimental framework and hours of collaboration on Bayesian statistics and data analysis.

Finally, I would like to thank my wife, whose support during this process has been truly invaluable.

This material is based upon work supported by the National Science Foundation under Grants No. 1815587 and 1815790. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

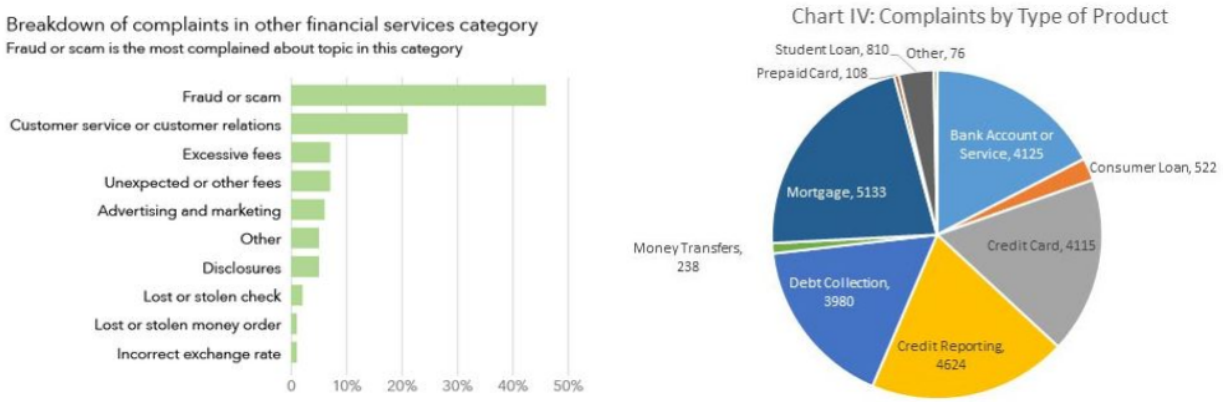
In the modern age of computers, smartphones, and omnipresent data, people are asked to parse information at a higher rate than ever before. Instead of subjecting audiences to bulky and potentially inscrutable tables of data, visualizations are regularly employed to engage readers and communicate more effectively. These graphics enable flexible and elegant presentations, but their flexibility increases the risks of misrepresenting data to an audience. Audiences for modern visualizations can be highly diverse, representing a population with varied skill sets and experiences; these variations can lead to differing interpretations of the information encoded in graphics and a potential for miscommunication between the designer of the visualization and its readers.

1.1 Relying on Visualizations To Communicate With A Diverse Audience

Visualization designers balance the need to communicate effectively with the need to present an appropriately engaging aesthetic. For organizations with a story-telling bent and a knowledgeable audience, the aesthetics that come with a sophisticated visualization may justify the effort while bringing minimal risk of misleading their readers. In contrast, organizations that focus on ease of public access are more likely to choose simplistic visualizations, even (or perhaps especially) when working with complex data. One

such organization is the Consumer Financial Protection Bureau (CFPB), a government agency created to work towards education and protection from financial predation and misinformation. This organization creates visualizations that are intended to be easily understood and accessible to a broad audience.

A pair of examples, obtained from the Consumer Financial Protection Bureau, are shown in Figure 1.1 [27, 26]. Designed for quick reference, these charts are intended to provide information about the relative severity of financial complaints. The visualizations shown in these cases, a bar graph and a pie graph, are simple designs that are sometimes taught as early as primary education [41]. They allow viewers to compare relative and absolute magnitudes or proportions of a whole, respectively, which enables a high degree of versatility in presenting various sorts of data.



(a) CFPB Bar Chart (b) CFPB Pie Chart

Figure 1.1: A pair of simple charts created by the Consumer Financial Protection Bureau [27, 26].

Ideally, these graphical representations accurately capture the structure of the data, as well as facilitating an interpretation of the data that is shared between viewers. Because the CFPB has engaged in tens of millions of customer interactions since its inception and is responsible for communicating with Americans across all walks of life, consistent and coherent communication is essential [21]. The CFPB is provided here as an example; newspapers, magazines, and numerous other sorts of media rely on visualizations to convey information concisely.

1.2 Design Guidelines For The Average Person (But Not Every Person)

The chart types chosen by the CFPB, as seen above, are familiar to most Americans. Their popularity has prompted many studies of similarly common and approachable visualizations in order to investigate their nuances and inform design guidelines. The outcome of these studies is used to inform visualization creators and aid them in generating readable graphics for as wide an audience as possible [38, 39, 73, 98].

A series of commonly used graphs are summarized in Table 1.1 [98]. These graphs were analyzed by Cleveland & McGill in a seminal paper in 1984, in an effort to characterize and rank the effectiveness of elementary graphical encodings [29]. Two broad categories of charts can be seen:

1. Bar and bubble graphs, which facilitate data lookup and comparison.
2. Pie graphs and tree maps, which represent part-to-whole relationships
3. Stacked bar graphs, which combine these types; as each vertical grouping is analogous to a pie chart.

1.2.1 Missing The Trees For The Forest

Although they are simple, relevant aspects of the way audiences perceive these graphical representations are not fully understood, leading to potentially misleading conclusions. For example, Cleveland & McGill's work concluded that bar charts were better than pie charts for examining proportional size of graphical elements. Their work and subsequent research was focused on assessing accuracy in judgments made during graphical perception tasks based on the type of graphic used. These studies assume that a population-level analysis is sufficient without delving into individual variation. In other words, the analysis focuses on what might be thought of as the 'average' person and is achieved by aggregating

Descriptive Characteristics	Encoded Data	Primary Purpose	Probable Strategy For Assessment
Bar graph	One quantitative attribute (bar height); One categorical attribute (separates bars)	Look up and compare values	Examine bar height
Bubble graph	One quantitative attribute (bubble area); One categorical attribute (separates bubbles)	Look up and compare values	Examine area; Examine diameter or radius
Stacked bar graph	One quantitative attribute (bar height); two categorical attributes (separates bars; separates interior of bars)	Part-to-whole relationships; comparison of bars	Examine area of interior segments; Examine overall bar height
Pie graph	One quantitative attribute (proportions); One categorical attribute (separates slices)	Represent part-to-whole relationship	Examine interior angle; examine external arc; examine area
Tree map	One quantitative attribute (proportions); One categorical attribute (separates nodes)	Represent part-to-whole relationship	Examine area; Examine length and width

Table 1.1: Characterizing simple graphs to demonstrate categorization

judgments made by multiple people during an experimental study. These results are as shown in Figure 1.2, in a visualization from a paper by Heer & Bostock reproducing Cleveland & McGill's work [54]. The conclusion of this research is that bar charts reliably outperform other elementary data encodings; further commentary in their paper helped contribute to the canonical rankings of *Bar* > *Pie* > *Stacked Bar* > *Bubble* [98].

The analytic approach used by Cleveland & McGill assumed that an individual's error when reading a visualization could be adequately modeled by solely considering the type of graphic used. This assumption has been maintained in work replicating and building off of Cleveland & McGill's research. Heer & Bostock, who drew from a different pool of experimental subjects but obtained the same results in a frequently-cited paper, supported their arguments with a theoretical foundation derived from psychophysics, a field which studies the mechanics by which individuals perceive stimuli. Their research also focused on recommendations for an 'average' person without consideration for individual variation or sub-populations which may meaningfully deviate from the aggregate

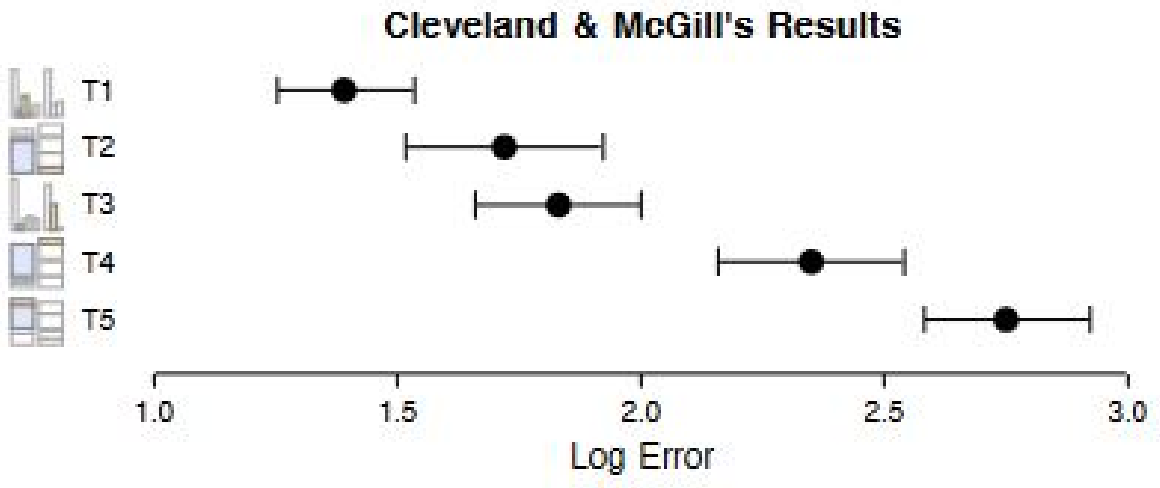


Figure 1.2: Selected results from Cleveland & McGill's work, comparing versions of bar and stacked-bar charts.

results [121, 88, 54]. Hullman *et al.* also extended Cleveland & McGill's work to consider the effect of social psychology, but again preferred an approach that did not consider individual bias [56].

This repeated assumption leaves a gap for consideration of individuals who may not show the same bias, whether due to variations in comfort with graphs in general, familiarity with the specific data under discussion, differences in strategy used to parse the visualization, or other factors. The *Probable Strategy for Assessment* identified for each chart type in Table 1.1 highlights the possibility of variations in strategy; a *probable* strategy is not universal, and the employment of less-common strategies represent a possible driver for individual variation [74]. Because organizations like the CFPB have a mandate that extends to hundreds of millions of Americans, each of whom is an individual, understanding the validity of applying aggregate-level metrics to the population as a whole is crucial.

1.3 Finding the Individual With Hierarchical Modeling

By focusing only on the 'average' viewer, visualizations made by relying on these guidelines may fail to adequately inform members of the populations who are not represented by the average participant. A visualization developed with experienced practitioners in mind may be too simplistic, with the creator leaving out information that would be beyond the grasp of a layman but easily understood by a expert. Conversely, an inexperienced viewer presented with a complex visualization may fail to derive any useful information, or could even draw false conclusions after misinterpreting the data. In cases such as the graphs in Figure 1, individual perception of the ratios between complaint sources can meaningfully affect decision-making. This may be especially relevant for small business owners and entrepreneurs, who are at risk for a variety of financial problems and lack the resources to thoroughly address all possible avenues of financial complaints and issues [81].

As a result, the ability of a business to accurately parse information from the CFPB can represent significant value. For example, an individual who believes that scams are vastly more prevalent than customer service complaints may focus their efforts accordingly, perhaps by investing in fraud protection instead of employee training; meanwhile, one who sees a smaller gap is likely to allocate resources in a more evenly distributed manner. If their decision is based on inaccurate reading of a graph, their business could be negatively impacted.

Cleveland & McGill opted to avoid analysis at the individual level, suggesting that because the error judgments made by each experimental subject's participation were correlated with each other, and so modelling that internal correlation represented a "substantial chore". Models that allow for analysis at different layers (in this case, both the visualization level and the participant level) are known as hierarchical models, multilevel models or mixed-effects regression models, among other names; until relatively recently, they have not been used to facilitate study of intraindividual variability [53].

Substantial advances in computing technology and techniques have made fitting mul-

tilevel models far more accessible, particularly in cases of complicated multilevel modeling where such calculations were previously complex to the point of intractability [24, 46]. Bayesian methods may be particularly well-suited to this modeling approach [46], and we find that other reasons exist to justify a Bayesian approach.

1.3.1 Opting For Bayesian Statistics

The population-level guidelines generated by Cleveland, McGill, and their peers rely on statistical analysis centered around null-hypothesis significance testing (NHST), as is standard in many fields. This approach is characterized by binary questions regarding the existence of an effect caused by some experimental condition. Concerns about this approach have grown recently due in large part to the "replication crisis" in psychology, where numerous published results were found to not be reproducible [110]. Its prominent use has also been called into question by practitioners within the visualization and Human-Computer Interaction fields.

These writers argue against continued use of NHST, which they refer to as 'frequentism', as they suggest it leads to "unjustified or arbitrary inferences" and frequent misapplication of results [94]. A number of them propose Bayesian statistics as an appropriate solution, suggesting that their use would be better suited for researchers [94, 102, 78].

In addition to similar statistical concerns, researchers in HCI and visualization have begun to argue that the Bayesian approach is a better fit for the nature of the field. Kosara discusses various threats to study validity, presenting several that could be addressed through appropriate application of Bayesian statistics [75]. Kay argues that Bayesian statistics allow for more nuanced questions such as "How strong is the effect [of the variable being studied]?" rather than settling for a binary approach [69]. A number of recent well-received papers have included results from both frequentist and Bayesian statistics [66, 35, 37].

Cleveland & McGill's work remains one of the most frequently-cited papers in the visualization field, and serves as a foundation for research performed as recently as the past few years, showing that its relevance has not faded [56, 113]. By employing a Bayesian

approach to hierarchical modeling, we seek to add nuance to their results and investigate perception at the individual level. We also propose to demonstrate that these techniques can be used to answer the same questions addressed by Cleveland & McGill using NHST, while providing further detail. By investigating on an individual level, we determine whether sub-populations exist that are not well-served by existing design guidelines. The hypotheses we will investigate are as follows:

- H1: Existing population-level results concerning graphical perception rankings are not universally true at the individual level.
- H2: Experiments can be used not only to rank graphical representations but also to rank individuals based on their skill at graphical perception.

1.4 Contributions

Here, I describe both the contributions of the Risks of Ranking paper, published in IEEE in 2022 as part of a team collaboration [31], and my personal contributions, both to the published paper and the unique portions of this thesis.

1.4.1 The Risks of Ranking paper

The methodology, results, and analysis are substantially in line with work done in a collaboration with Prof. Matthew Kay of Northwestern University, Dr. Xiaoying Pu of UC Merced, Brian Hall of the University of Michigan, and Karen Bonilla, Yiren Ding, Dr. Mi Feng, and Prof. Lane Harrison of WPI [31].

This work includes the following contributions:

- **Demonstrating that considerable variation exists between participants**. The average difference in performance across pie, stacked-bar, and bubble charts ranges from 1-1.5 percentage points (pp), which is less than the expected difference between individuals (1.5-3pp). Variation in average performance of over 20pp could be seen between members of the population for some charts. While the variations were generally not this pronounced, they were consistent and present across the

population. Additionally, in analysing the correlation that Cleveland & McGill passed by, we show that individuals can be expected to display consistent success (or lack thereof) in comparative judgments, further highlighting the distinction between participants.

- **Findings that the canonical rankings of visualization performance do not hold for the population as a whole**. Generally, we found that between-person variance outweighs the selection of visualization in cases that don't include bar charts. Even in the case of bar charts, around 20-25% of the population showed higher error for bar charts in their estimation judgments than on some other chart.

We discuss further insights that can be gained in the visualization field using our hierarchical Bayesian approach, such as assessing variations in strategy and their impact. The consistency of individuals (positive and negative) even across chart types that seem to favor different strategies suggests that some other individually-driven factor is at play, or that individuals are likely to develop effective or ineffective strategies across multiple visualizations. Furthermore, the significance of between-person variation relative to between-chart visualization suggests that designers might preferentially focus on aspects of visualization design (*e.g.* aesthetics or designer preference) rather than making decisions on design guidance pertaining to chart type.

1.4.2 Personal contributions

In the collaborative work, I was a significant contributor to the experimental process, especially with regard to the generation of the experimental stimuli and the data points needed to display them appropriately. I also helped develop the web design and training data for soliciting user input, alongside general collaboration for the work as a whole.

As an addition to the collaborative work, I present the following personal contributions:

- **A more detailed explanation of the modelling approach used in the paper submission**. The abbreviated nature of paper submissions, along with the expected technical expertise of the audience, leads to a more rapid development of

the analytical approach. I proceed more slowly through the modelling approach in order to elaborate on the technical composition of the model, as well as outlining in more detail our implementation in the modelling packages used.

- **A stratified ranking of individuals by graphical perception performance**. As an expansion of the positive correlation between individuals' performance and the establishment of between-person variance in performance, I demonstrate a ranking of participants from this experiment. Participants can be grouped into three basic categories: consistently high-performant, consistently low-performant, and middling-performant, which shows greater variation in participant ranking between the chart types. The preliminary identification of these strata provides opportunities for more focused investigation of those factors that characterize and differentiate the individuals in each.
- **Commentary on the relevance of true proportion to graphical judgments.** As an additional step in the modelling process, I explore incorporating additional portions of the experimental data into our model. I discuss the implications of this for our model, review an initial set of results based on preliminary modeling, and discuss the relevance of this approach for future research.

Chapter 2

Background and Related Work

A significant portion of information visualization research focuses on the ability of a viewer to accurately judge data present in graphical form, *i.e.* graphical perception [79]. These studies provide the foundation for research that investigates effective ways to communicate with audiences, whether comprised of the general public or a subset of particular interest, as well as novel ways to present data and orchestrate tools for professional environments. The reliability of these studies is supported by statistical analysis, primarily null-hypothesis significance testing (NHST), as is the case in most contemporary fields. This thesis is built on existing graphical perception studies and other work in the information visualization field, as well as recent publications on effective statistical analysis.

2.1 Graphical Perception

Graphical perception studies, in general, focus on how people retrieve information encoded in a visualization. This includes both the investigation of relatively well-known encodings of information (*e.g.* geospatial data, tree maps, scatter plots, dot plots) [9, 71, 67, 20, 52], the nuances of human perception within those encodings [116, 121, 112, 123], and the investigation of novel encodings [62, 72, 67]. Visualization research also contains research focused on scenarios for data analysis during practical application of visualization tools, including understanding the professional environments where those tools are employed, their efficacy in that specific domain, and whether or not the visualization supports

decision-making and understanding in both casual and focused use [79].

Research regarding better-known encodings provides useful context and helps to remove wrongful assumptions and ambiguities from the existing literature and guidelines [72, 20], and provides a foundation for the development of more complex tools [79]. However, Lam *et al.*'s analysis of papers submitted to leading visualization conferences showed that papers such as these make up a far smaller percentage papers that focus instead on creating and exploring the space of novel visualizations. As a result, visualization creators who want to employ well-understood graphics to connect to their audience have relatively sparse recent research to draw from.

This challenge, which is especially relevant for inexperienced visualization creators, becomes more relevant when considering the increased use of the internet for public-facing visualizations for federal agencies, the financial sector, journalism, sports writing, and more [120, 122, 28, 115, 1, 27, 26]. These visualizations are generally not created by specialists in visualization theory, but rather by professionals in other fields. Furthermore, they are accessible to anyone with internet access, and intended for use across a wide spectrum of demographics and backgrounds. Therefore, efforts to make them equally useful to all potential users requires detailed insight regarding people's ability to understand graphical encodings.

The study by Cleveland & McGill provides a useful foundation for investigating personal variation due to its widespread acceptance, as well as cases of successful replication [72, 54]. Cleveland & McGill asked participants to compare a smaller graphical element (*e.g.* a bar, or segment of a pie chart) to a larger one and estimate its proportional size, as shown in Figure 2.1. This study concluded that graphical representations that separate data using position (*e.g.* bar charts) enabled users to make more accurate assessments than those that separate based on angle or area (*e.g.* pie charts and bubble charts).

The work by Cleveland & McGill makes well-known conclusions about the effectiveness of visualizations for accurately conveying proportions. The results from Heer & Bostock's replication are shown in Figure 2 [54]. These rankings, which are frequently

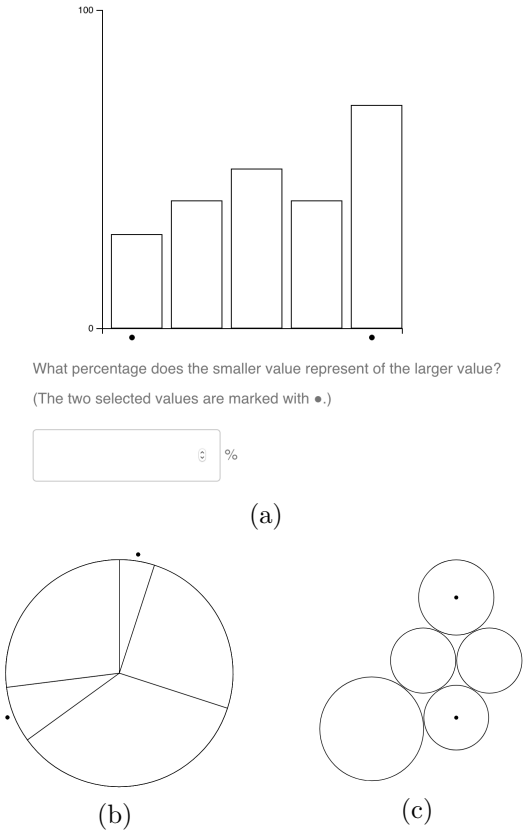


Figure 2.1: Example tasks from Cleveland & McGill's work. Participants were asked to compare the ratio of the smaller element (B) to the larger element (A). The experimental prompt shown with the bar chart appears under every graphic.

referenced, are generated using aggregated data from participants and make no comment on results at the individual level. As a result, while they are frequently viewed as being universally applicable, there is no analysis to suggest whether this is an accurate interpretation [72, 98].

Because of their widespread usage, a variety of fields are invested in visual design and have publications on the topic. Many of them reference Cleveland & McGill's work in isolation when crafting advice for other designers; these include textbooks on information visualization, pedagogical research, forays into specific uses of visualizations, and more casual articles written by professional visualization designers [91, 109, 84, 38]. Their work is also cited during development of automated design tools and formal mathematical treatments of graphical design [96, 88]. Other guidelines, such as those given by the Harvard Business Review, offer more consideration to other visual research but recognize Cleveland & McGill's work as seminal and still relevant for consideration [98, 12, 33, 42].

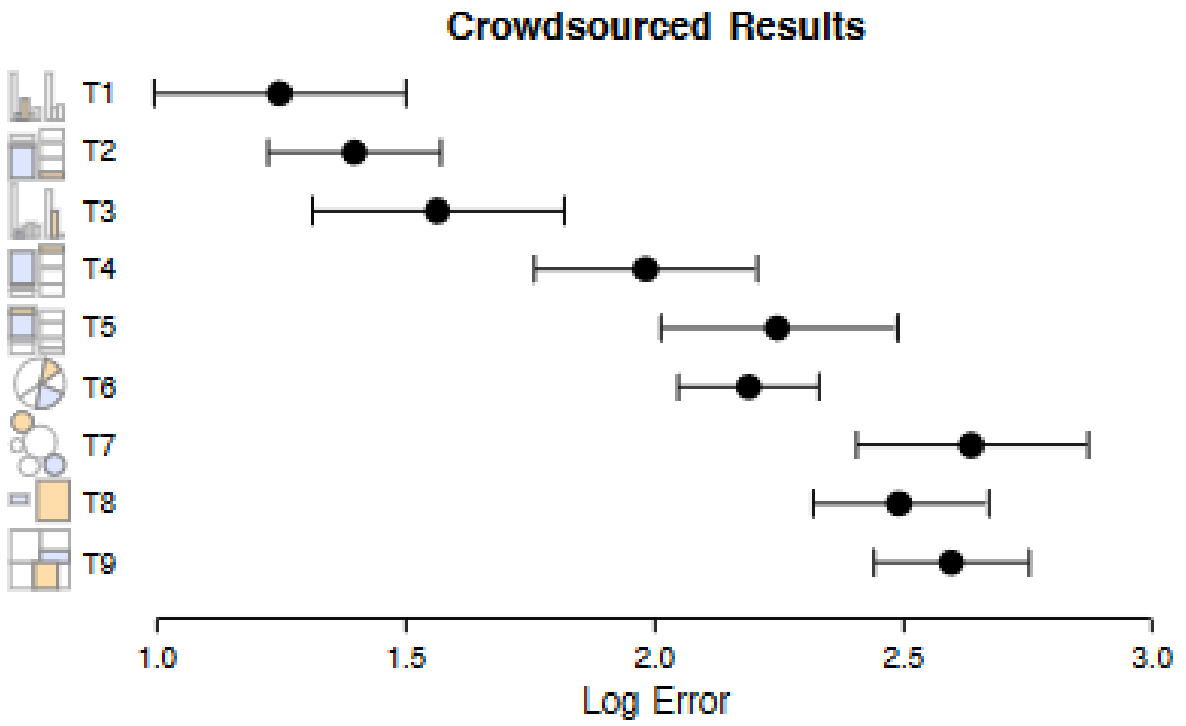


Figure 2.2: Results from Heer & Bostock's replication of Cleveland & McGill's work. Bar graphs show noticeably less error than pie charts and other encodings that are likely to require more complex assessment strategies.

2.1.1 Studies of the Individual In Visualization

While the above instances of research do not explicitly incorporate information about the individual in their approach, visualization studies have been done to identify factors that might cause variation between individuals. Existing research in visualization concerning the individual can be generally separated into two approaches: The individual whose personality informs their approach to a visualization, and the individual whose experience and level of comfort with visualizations informs their understanding of a visualization. These differences are sometimes understood at the group level (*e.g.* binning experimental subjects by personality factors), and sometimes with the individual as the focal point (*e.g.* developing testing paradigms to glean pertinent information about participants).

Assessing Intrinsic Sources of Individual Variation

Liu *et al.* suggest in a recent State-of-the-Art report for the Computer Graphics Forum that these differences can be broken down into three main dimensions: *Cognitive traits*,

cognitive states, and *experience* [86]. They differentiate between the cognitive dimensions as being intrinsic to the individual (traits) or an ephemeral state predicated on physical circumstances and recent events (states).

A series of "aptitudes", such as perceptual speed and spatial orientation, with associated tests for measurement were developed by the Educational Testing Service [32, 36]. These aptitudes resemble Liu's *cognitive traits* in that they are intrinsic and not readily mutable, if at all. Of these aptitudes, spatial orientation (or spatial ability) is the closest match to our focus on graphical perception; other examples of these aptitudes include inductive reasoning, perceptual speed, and number facility, none of which present themselves as being particularly relevant to our experimental task. Spatial ability is supported as a differentiating factor between individuals by recent research [101, 119].

Personality traits have been identified by multiple researchers as a relevant factor in visualization contexts: as a proposed proxy for their preconceptions in how to approach and interpret data [125], as a direct factor on the manner in which individuals engage with tasks and graphic interfaces [51], and generally as a necessary consideration for designers [126].

The cultural upbringing of participants has also been found to have bearing on their interactions with graphical interfaces [22], their approach to parsing information with a spatial component [11], and their level of comfort with a visualization [105].

Visualization Literacy As A Source Of Individual Variation

Visualization literacy as a term has a variety of definitions, but most generally capture the spirit of "a group of skills which enable an individual to understand and use visuals for intentionally communicating with others" [6]. Presentations of data can vary significantly, from the elementary tasks used by Cleveland & McGill [29] to more complex representations of uncertainty [61], or even interactive graphics built on modern technology [40]; in light of this, remaining sufficiently literate to understand the array of possible data representations can be difficult.

Numerous studies have concluded that visualization literacy is important to society

as a whole, and is lacking. Galesic & Garcia-Retamero find that the issue is cross-cultural and discuss the need for research to address "barriers of low numeracy and graph literacy" [43]. Börner *et al.* found interest but limited knowledge regarding visualizations, particularly more esoteric ones such as network visualizations [15]. Maltese *et al.* found significant differences between experts and novices, with middling groups showing less differentiation [89]. Peck *et al.* argue that data visualization studies are often performed by individuals with a relatively high degree of familiarity and comfort with visualizations, which is not characteristic of the population as a whole and implies a need for more focus on the "data poor" [103].

Visualization literacy stands out as a component of individual variation because it is the easiest to impact pedagogically, compared to more innate concepts such as perceptual speed. To address these concerns, educators are working to establish frameworks for assessing and improving visualization literacy in online coursework [14] and as early as elementary school [5].

2.2 Use of Statistics in Data Visualization

As with a number of other fields, NHST have served as the primary form of statistical analysis in visualization and studies focusing on human-computer interaction [34, 67]. NHST focuses on the 'null hypothesis', which is stated in opposition to an alternative hypothesis that represents its inverse. The purpose of experimentation is to collect evidence that will either suggest that the hypothesis is false or fail to do so. To make this assessment, we assume that the given experiment is a sample from an infinite series of identical experiments, with the random data representing a random sampling from a fixed population distribution [80].

2.2.1 Arguments Against NHST and For A Bayesian Approach

Concerns regarding this approach have existed for decades, as even professional statisticians find that it can be confusing and difficult to properly apply [108, 10, 17, 50, 63, 94,

102, 30, 3, 25]. The intent of NHST is to assess the probability of observing the available data given the null hypothesis, but accurately translating that intent to practice has been shown to have difficulties by students, practitioners, and mathematicians; oftentimes, individuals from the same field cannot agree on even simple statements about the theory behind confidence intervals. [94, 93, 95]

In addition to its complexities, NHST is limited by its binary nature. While questions can be posed to provide useful information (*e.g.* assessing a reasonable estimate of the mean of a value, as in Cleveland & McGill), it is difficult to ask more complicated questions (*e.g.* "How confident can we be regarding this conclusion about the mean?"). NHST also suffers from an inability to *confirm* the null hypothesis, which limits the amount of useful information that can be produced from a study [77].

Bayesian statistics also address concerns regarding replication in visualization, which have been written about extensively in other fields [102]. Kosara & Haroz wrote on a number of concerns in experimental design and methodology, several of which are ameliorated by a shift away from NHST. They suggest that the nature of a 95% confidence interval inherently allows for false positives, which is not true in a Bayesian context. Additionally, they create a motivation for "p-hacking" and other questionable research practices undertaken with the intention of generating publishable results at the expense of statistical validity [90], which is not possible when employing Bayesian statistics. Bayesian statistics instead often present what is often called a "credible interval", which show the values a parameter is likely to take on, along with the probability of each value.

By presenting results in the form of a probability distribution, analysts are able to obtain a more accurate portrayal of their solution space. This is particularly important in the case of discrete events, which are inherently unrepeatable and therefore difficult to fully address using confidence intervals, since p-values and confidence intervals require that probability is founded on a long-run frequency [64]. Showing results as probability distributions, however, has the potential to confuse readers, particularly in contexts where the audience is not comfortable with statistics [4, 2, 85]. The underlying issues with respect to the public's familiarity with statistics are broadly pedagogical in nature, but

application of novel encodings of probability can help with reasoning about uncertainty and probability distributions [37, 85, 57].

In visualization and HCI, Kay *et al.* support the use of Bayesian statistics due to sample size and difficulties in conducting meta-analyses. Kay & Heer argue that the use of prior and posterior distributions simplifies and formalizes the process of using past work as a foundation for future work, as well as providing more detailed information about parameters of interest [66]. Both papers make use of multilevel models incorporating random effects at the individual level and find that individual differences are relevant. Other authors have taken a similar approach [37, 70, 97].

2.2.2 Strengths and Weaknesses of a Bayesian Approach

Bayesian statistics have seen increasing use in recent years [92]. While the potential contributions of Bayesian statistics have been acknowledged in theory for decades, it is only in recent years that computational power has grown to the point that it can be used as a part of standard practice [69]. Bayesian statistics are oriented around the notion that the parameters of interest in a system are inherently uncertain but can be described using the available data, which is treated as a fixed and invariant set of information [80]. This approach relies on Bayes' rule:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{2.1}$$

where $p(\theta|data)$ denotes the probability of a hypothesis θ given the available data.

In practice, Bayes' rule has three parts:

1. The likelihood, $p(data|\theta)$, which represents the probability of generating the observed data across a range of possible values for θ.

2. The prior, $p(\theta)$, which is used to describe our belief about the parameter prior to observing the data.

3. The posterior probability distribution, $p(\theta|data)$, which represents plausible parameter values, conditional on the model, after considering both the prior and the data.

The denominator, $p(data)$, is a normalizing constant and does not affect the posterior probability distribution, which is usually the focus of Bayesian models. Consequently, this variable is often excluded [99, 87].

The most well-known challenge with Bayesian inference is the specification of prior distributions for all parameters of interest. The modeling process and and prior both require designation of an appropriate probability distribution, which ideally the researcher arrives at by considering the domain space based on his own awareness of the field [107]. Prior specification also requires estimation of the value of the parameters, which again is ideally determined based on the domain knowledge of the researcher.

The concept of prior selection is seen with skepticism by many, as some find it too subjective for modern scientific standards [92]. As such, strategies and techniques for setting priors are discussed at length in most books on Bayesian analysis [76]. A proper approach to setting priors in a given domain requires primarily two things:

1. Sufficient grounding in statistics to understand what various functions can be used to capture prior knowledge
2. Sufficient domain expertise to understand the functions that experts in that domain use for modelling

A researcher with insufficient statistical grounding or domain expertise could easily end up selecting an inappropriate model or incorrect prior estimates for parameter values, so while the Bayesian framework can contribute greatly to statistical rigor, the skepticism is not without merit.

Even without a strong background in statistics, a Bayesian approach is not unachievable. Proponents of a Bayesian framework sometimes make a conservative choice in model selection, using what are called "uninformative" priors; these models are ones where the model is chosen to be generic and therefore contribute little to the posterior probability distribution [117]. An uninformed prior would be one such as a uniform distribution, where any possible value is equally likely to occur.

However, because researchers generally have some understanding of the science surrounding their domain, it is often possible for researchers to choose more informative

priors that reflect either previous experiments or experience in the field [92]. In some circumstances, this can take the form of a differently chosen prior, based on other research focused on modelling in a specific field, *e.g.* using a linear log-odds model when assessing human perception of probability [124]. Other researchers may conduct pilot experiments or choose to use some portion of their data to inform an exploratory analysis, using those results to choose a more appropriate model, *e.g.* beta regression models used by Kay *et al.*, where a less informed approach would have seen a linear regression model used [68]. In order to take full advantage of the Bayesian framework, analysis can use previous research to inform both the model selection and the parameters of the model [66].

2.2.3 Types of Priors

The assignment of prior distributions is relevant primarily insofar as it affects the final analysis, as that is the focal point of the research being conducted [48]. Gelman *et al.* discuss a series of options for establishing priors, including:

- A fully defined, fully subjective prior (*i.e.* composed by the statistician based on their domain expertise, and made *a priori* without the likelihood in mind) that is intended to capture all aspects of the information available prior to engaging in new data collection and modelling. This type of prior assumes that the statistician's intent is to leverage existing information to build a deeper prediction about the relevant context, and require confidence on their part in their ability to incorporate information from outside the scope of their current research.
- A *minimalist* prior, which is another way of discussing the "uninformative" prior discussed above. The stimulus behind choosing such a prior is generally to minimize the impact of the prior distribution on the likelihood, and is often seen captured with the optimistic use of a diffuse prior such as a uniform distribution.
- A *reference* prior, which is a formal approach to the notion behind minimalist priors. The intent behind reference priors is again to minimize information added by the prior to the sample information; however, rather than relying on ad hoc selection of hopefully diffuse priors, this approach applies information theory to calculate the

prior distribution that would be the *least informative* and therefore make the sample information the dominant contributor to the posterior distribution [13].

- A *structural* prior, so called because it encodes "structure about the model rather than particular numerical values" [44]. Because a structural prior encodes structural information about the distribution at hand and therefore provides pressure on the posterior distribution towards that structure, it is not properly minimalist. Gelman *et al.* reference maximum entropy models as a context where structural priors can be applied. Jaynes characterized the appropriate circumstances for these as being where there exists "a well-defined hypothesis space and noiseless but incomplete data" [60]. These priors put on the statistician the responsibility, discussed above, of applying knowledge from previous experiments or experience in the field to assess the structure accurately.
- A *regularizing* prior, which, in keeping with the concept of **regularization**, biases the posterior distribution toward "smoother, more stable inferences" [48]. This category of priors again pressure the posterior distribution towards certain characteristics; it differs from structural priors in that they are less concerned with the generative structure of the data, and more concerned with the statistical properties of the model output.
- A *weakly informed* priors, which relies on understanding the structural domain of the problem at hand (as with structural priors) and also bias the resulting inference towards a more regularized output (as with regularizing priors). This type of prior generally draws lightly from domain-specific information, in order to capture structural information and some quantitative aspect of prior knowledge, but does not seek to aggressively inform the posterior distribution with information from prior research. A weakly informative prior should accurately reflect the shape of the data, *e.g.* a Cauchy distribution for studying logistic regression, where the prior is normalized such that the center of the distribution represents no effect [47]. This choice helps to predispose the posterior probability distribution towards a conclusion of no effect, unless the impact of the experimental data is strong enough to overcome

the conservative no-effect model, which limits potential objections from Bayesian skeptics.

2.2.4 Impetus for a Bayesian Approach To Hierarchical Modeling

Aside from the differences in the output of a Bayesian and frequentist model, it is worth noting that the fundamental mathematics behind each approach differ in ways that render other advantages to a Bayesian approach. In the research performed by Cleveland & McGill, as well as subsequent reproductions, the stimuli for each individual's responses are the same graphics; consequently, as noted by Cleveland & McGill:

> Because each subject judged all of the experimental units in an experiment, the judgments of one unit are correlated with those of another, and modeling this correlation would have been a substantial chore.

A Bayesian approach is better suited, conceptually, to answer this question, but until recently was limited by the computation complexity required. Mixed-effect models that account for dependencies on multiple levels, as seen here in the hierarchy of individuals to the population, are not new, but have not historically provided information beyond the mean of the response distribution [23]. Advances such as modern Markov chain Monte Carlo methods [24] have enabled Bayesian models that assess individual-level parameters, *e.g.* mean and variance for an individual, in tandem with the population-level parameters, eliminating much of the difficulty behind Cleveland & McGill's "chore". Assessing individual-scale results would not otherwise be practicable.

Chapter 3

Methodology

This thesis uses Cleveland & McGill's work as a foundation to examine individual variation when performing graphical perception tasks. The quantity of works referencing or built on their research makes it a valuable starting point for research such as this, which examines studies following that paradigm. Cleveland & McGill's methodology was not designed for analysis of individuals; consequently, this section will review the foundational experimental protocol as well as necessary adjustments to enable measurement and comparison of performance between participants.

The methodology outlined here was developed as part of a collaborative effort performed under Prof. Lane Harrison of WPI, alongside Prof. Matthew Kay of Northwestern University, Dr. Xiaoying Pu of UC Merced, Brian Hall of the University of Michigan, and Karen Bonilla, Yiren Ding, and Dr. Mi Feng, all affiliated with WPI. This work was published in IEEE in 2022 [31].

3.1 Experiment Stimuli and Data Generation

Much of the experimental methodology is in keeping with Cleveland & McGill's research, as well as the crowdsourced replication from Heer & Bostock [54]. The visual stimuli were consistent with that of Cleveland and McGill; namely, black-and-white visualizations with five data points, with two elements identified for comparison, and a prompt for a numerical answer to their ratio.

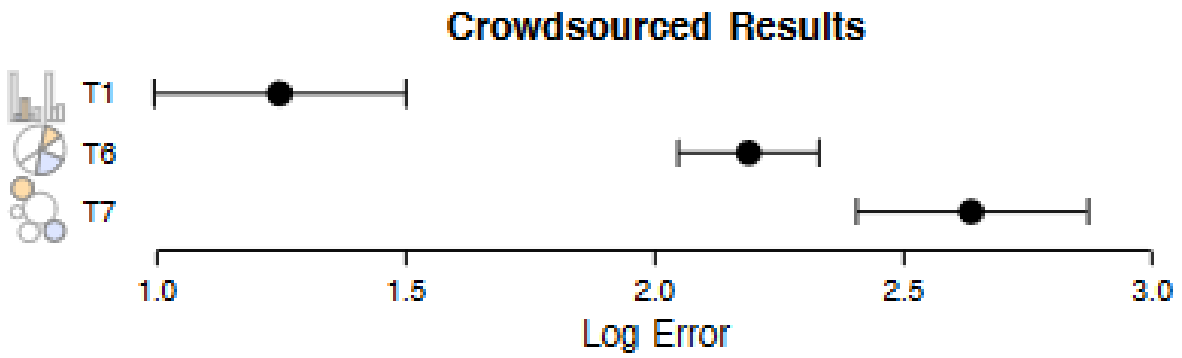


Figure 3.1: The confidence intervals resulting from Heer & Bostock's replication experiment for the selected chart types. Clear separation can be seen between the intervals for each.

The visualization types were modified to enable comparisons between participants while keeping the experiment at an appropriate scale. Rather than using all of the charts present in Cleveland & McGill's initial work (5 chart types) or in Heer & Bostock's replication (9 types), this experiment focused on bar charts, pie charts, stacked-bar charts, and bubble plots. This allows us to compare both within and across chart categories as described in the introduction, *i.e.* assessing charts that hinge on part-to-whole relationships (pie chart) alongside those that rely on adjacent shapes of different size (bar and bubble chart) and a combination of each (stacked-bar charts). As a result, we anticipate that individuals may apply different strategies for estimation, which is more likely to yield differentiation in accuracy. Additionally, the confidence intervals from the previous studies which describe estimation error show clear separation between several of these chart types, particularly bar charts to the rest, as seen in Figure 3.1. As a consequence, any individual inversion of the expected population-level will be more convincing.

The data generation process was broadly in keeping with the above studies, with adjustments made to enable consistent stimuli between participants, as well as repeated trials to enable variance analysis. Each trial utilized five values: one smaller and one larger value for comparison purposes, along with 3 distractor values. Cleveland & McGill employed random number generation to obtain true proportions ranging from 10% to 99.7%; in this experiment, the spectrum of true proportions of the smaller value to the larger value ranged from 5% to 95% in increments of 5%, except in two cases, as well as a propor-

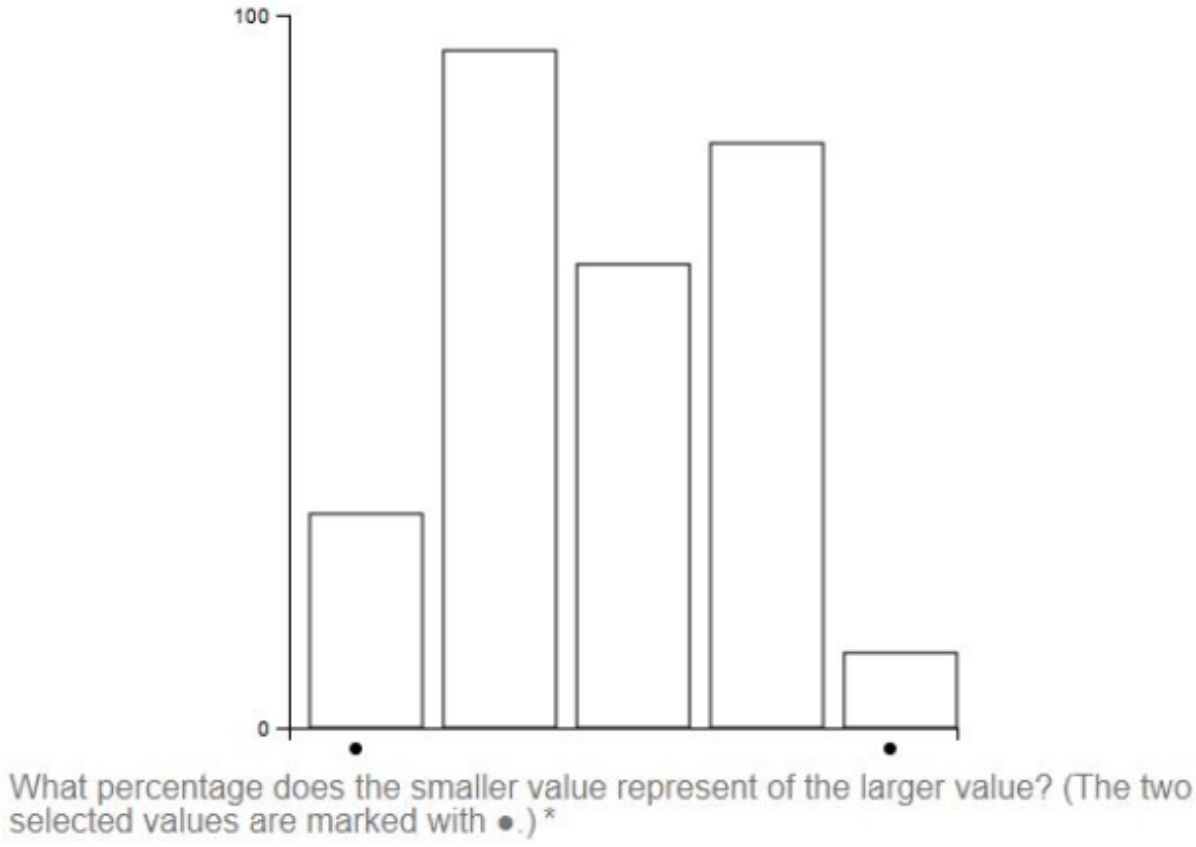


Figure 3.2: An example of a bar graph shown to experimental participants. The prompt for each trial appears underneath the graph; this particular graph shows a ratio of approximately 40%. This is similar to the prompt of Cleveland & McGill, but with a more refined graphical presentation and a slightly adjusted prompt.

tion of 99%, for a total of 20 distinct proportions. The exceptions were a substitution of 0.33 (representing a judgment of one third) for 0.35, and of 0.66 (representing a judgment of two thirds) for 0.65. This spectrum of true proportions was chosen to mitigate the effect on error of participants' likely use of rounding, as shown by Talbot *et al.* [113]. For the *Bar*, *Stacked Bar* and *Bubble* charts, the distractor values were randomly generated within a normalized 0 to 1 range. For the *Pie* chart, the distractor values were generated randomly, with the additional constraint that they must sum with the comparison values to 1. For each *visualization-dataset* pair (4x20), there were 15 *repetitions*, yielding a total of 1200 trials per participant (4 *visualizations* x 20 *datasets* x 15 *repetitions*. These 1200 trials were shuffled after generation to ensure random presentation to participants.

3.2 Experiment Procedure

The experiment was conducted in three stages: *Training*, *Trials*, and *Demographics*.

Training:

Before conducting the experiment, each participant was shown a series of training trials to ensure that they understood the task. Two training trials were shown for each chart type, for a total of 6 training trials. Each trial resembled those shown in the experiment, with added feedback after each trial to inform the participant of the correct answer. The training trials included both rounded and non-rounded answers and ranged from a true proportion of 13% to 100%, to avoid systemically biasing participants towards providing rounded answers or towards providing answers within a certain range of numbers.

Trials:

As with Heer & Bostock's experiment, this phase prompted participants to make a "quick visual judgment", discouraging them from physically measuring the size of the comparison graphics. Breaks were given after every 60 trials to mitigate the effect of fatigue and to limit participant dropout rate.

Demographics:

After all trials were completed, a series of demographic information was collected, including age, gender, country of origin and level of education. Participants were also asked to assess their experience with visualization and statistics on a 1-7 scale.

3.3 Experimental Participants

Participants were recruited from Amazon Mechanical Turk, which was validated by previous graphical perception research validated as a reliable source for graphical perception experiment participants [54]. The compensation for participating in the experiment was $22. Of the 200 participants who began the experiment, 118 completed all trials. With an average completion time of 2 hours 40 minutes, this yielded an hourly rate of $8.3. Of the 82 participants who did not complete all trials, 57.3% (*i.e.* 47 participants) failed

to complete 10 trials, and 80.5% (*i.e.* 66 participants) failed to complete 100 trials. All participants received the same set of trials to enable a within-subjects experimental design.

3.3.1 Exclusion Criteria

To ensure that participants were making good-faith assessments, we used values at each end of the true proportion range (5% and 100%) as attention checks. These attention checks served to identify participants that were failing to focus on the prompts, not trying to be accurate, or not understanding the instructions. Participants were excluded only when consistently (*i.e.* more than 25% of the time) making errors of greater than 50% on such questions. This resulted in the exclusion of 9 participants, leaving 109 participants in all.

Chapter 4

Analytical Approaches

Our primary purpose is to place Cleveland & McGill's analysis inside a structure where within- and between-participant variance can be studied. Because our experiment's methodology followed theirs, we have the opportunity to replicate their analysis using our data, confirming that our data solicitation process yielded comparable results, and then construct an alternate model using Bayesian techniques. The latter approach will employ a model expansion, as demonstrated by Gelman *et al.* in their discussion on Bayesian workflows, to facilitate an exploration of within- and between-participant variance [49].

4.1 Reviewing Cleveland and McGill's approach

We will directly replicate the process followed by Cleveland & McGill, permitting us to verify that our data is in keeping with prior work. Heer & Bostock's methodology is similarly comparable to Cleveland & McGill's, so the findings from their research will be included as well in in our comparative analysis.

The error metric captured by these researchers was the absolute value of the difference between the judged percentage, *i.e.* percentage assessed by the participant, and the true percentage. The formula they used is: $\log_2(|\text{judged percent} - \text{true percent}| + 1/8)$ [29] Cleveland & McGill stated that they used log base 2 as it "seemed appropriate to measure relative error", and added a slight positive adjustment to the terms to mitigate the impact of the log scale where the error is close to 0.

Because Cleveland & McGill saw frequent outliers in their experimental data, they relied on midmeans to mitigate their impact rather than incorporating them into their measurements. A **midmean**, or **interquartile mean**, is the mean of the middle two quartiles in a set of data. By taking the mean of a midmean rather than the mean of the entire set of observations, a more robust measure of the central value of a dataset is achieved, albeit at the expense of potentially relevant data in the first and fourth quartiles.

Cleveland & McGill employed bootstrapping, *i.e.* random sampling with replacement, to generate a sampling distribution of these midmeans. In this approach, 1000 bootstrapped samples were drawn, each with the same observation count as the experimental data, yielding 1000 means of midmeans.

Unlike our experiment, Cleveland & McGill's experimental structure did not have repeated trials for a given pair of true proportion and visualization type. In our replication of their work, we minimize the impact of these repeated trials by calculating the mean response from each of our repeated trials. This sees our bootstrapping draw from a single unique combination of true proportion, visualization, and participant, as in their work.

Each bootstrapped sample provides 20 midmeans (based on the participants' estimates for each true proportion) for each visualization type. The mean of these 20 midmeans provides a relatively robust assessment of the experimental participants' relative error. Generating 1000 bootstrapped samples provides 1000 means of these 20 midmeans, enabling us to create 95% confidence intervals for each visualization type.

4.2 Expanding Cleveland & McGill's work in a Bayesian context

Cleveland & McGill felt that modelling the relationships between an individual's judgments represented a "substantial chore" and did not pursue it in their work. Because we seek to capture the variation in variation across individuals' judgments, as well as the variation between chart types for a given individual, we develop a model that achieves these goals without losing the ability to describe errors relating to visualization type.

Following a model expansion approach will allow us to begin by replicating the outcomes from Cleveland & McGill and Heer & Bostock, and then proceed to further develop the model to achieve our goals.

4.2.1 Explaining model expansion

A Bayesian model must start with some likelihood, where each input has designated priors that capture our understanding relevant to that likelihood. Because mean absolute error is the experimental focus, we will start with a model that matches Cleveland & McGill's approach, outlined as follows:

1. By showing experimental participants a pair of visualized data points, a judgment of their relative size can be made. This judgment will have some amount of error compared to the true proportion. A precisely accurate judgment will have an error of zero.
2. Aggregating participant error in judgments across varying relative sizes, for a given visualization, can give an estimate of mean absolute error for that visualization type.
3. All true proportions, which are part of the calculation of mean absolute error, are comparable and therefore can be aggregated with no loss of information.
4. All participants are comparable and therefore can be aggregated with no loss of information.
5. Assumptions regarding use of confidence intervals are met (*i.e.* mean absolute error can be treated as a normal distribution).

This yields the following equation:

$$\begin{aligned} \text{Mean absolute error} &\sim \text{Normal}(\mu[i], \sigma) && \textit{normally distributed likelihood} \\ \mu[i] &= \beta[\text{vis}[i]] && \textit{mean submodel} \end{aligned}$$

This is to say, mean absolute error is normally distributed with a mean $\mu[i]$ that varies based on some aspect of the observation, and a constant standard deviation σ. Specifically, the point estimate of a given observation depends on the visualization type

used to solicit input for that observation. A separate mean is different for each visualization type, with the overall mean absolute error associated with that particular observation represented by $\beta[\text{vis}[i]]$.

We know that a few shortcomings from this model exist:

- The errors are constrained to describe the ratio of a **smaller** to a **larger** value, and on a percentage scale, so they cannot exceed 100; this constraint is not captured by a normal likelihood.
- Some individuals may be consistently better or worse on some, or all, visualization types relative to other participants, meaning that information is potentially lost when aggregating.
- Cleveland & McGill note that "When the true percentages are in the range of 25-50, subjects tend to underestimate values for [bar and pie charts]", and noted a similar negative bias for stacked-bar charts when the true percentages were between 30 and 70 of "some dependence of log error on the true percent". Accordingly, we lose information that explains some of the variation between chart types by aggregating across all true proportions.

We hypothesize that participants are not interchangeable with respect to their judgments. Consequently, accuracy across visualization types, or overall ability to make accurate proportional judgments, will vary across participants. We can therefore improve the precision of our model by incorporating individual-level differences. We will use model expansion to refine and augment the model to account for these individual variations as well as the concerns with error bounds.

As in the methodology, the discussion and analysis that follows are mostly in line with the collaborative work recently published in IEEE [31]. Distinct from this collaboration, I explore possible gains in precision from incorporating true percentage into the model. This approach will be outlined separately at the end of the following chapter, in order to not distract from the main theme of the thesis.

4.2.2 Choosing an appropriate distribution

Before taking steps to expand the model to incorporate participants as an explicitly understood source of variation, we first select a distribution that fits the needed specifications. The normal distribution is generally seen as a default distribution for linear regression, and was employed without commentary by Cleveland & McGill. It is not, however, an exact fit for the circumstances of this study, as suggested above. Use of a normal distribution requires that the bounds of mean absolute error be positive and negative infinity. The experiment limited the permissible input to a range of 0% to 100%, where 0 is possible (no error), but 100% is not possible (a proportional judgment cannot be 100% wrong), so the distribution should be chosen to reflect that.

Our need, then, is a distribution that is bounded by 0 inclusive and 1 exclusive, which has the flexibility to represent the shape of possible parameter values for mean absolute error. A standard Beta distribution spans from 0 to 1, both exclusive, making it commonly used for bounded data; additionally, its shape parameters give it substantial flexibility within these bounds [111, 23]. This model is usually denoted by its shape parameters as Beta(α, β) but can also be represented using its mean μ and the sum of its shape parameters $\phi = \alpha + \beta$, with ϕ now representing **precision**, a value that increases when variance decreases [76]. This parameterization is as follows:

$$\alpha = \mu\phi$$

$$\beta = (1 - \mu)\phi$$

$$\text{mean\_abs\_error} \sim \text{Beta}(\mu\phi, (1 - \mu)\phi) \qquad \textit{Beta-distributed likelihood}$$

In order to incorporate the inclusive 0, we can model the probability of a zero being present separately using the zero-inflated Beta distribution using the Bernoulli distribution, which discretely captures the probability π of some binary outcome, for that special case. The zero-inflated Beta distribution is as follows:

$$y \sim \text{ZeroInflatedBeta}(\mu, \phi, \pi)$$

$$\implies y = \begin{cases} 0 & \text{if } z = 1 \\ y^* & \text{if } z = 0 \end{cases}$$

$$y^* \sim \text{Beta}\left(\mu\phi, (1-\mu)\phi\right)$$

$$z \sim \text{Bernoulli}(\pi)$$

This leaves us with a model for mean absolute error that exchanges a normal distribution for a zero-inflated Beta distribution:

$$\begin{aligned} \text{mean\_abs\_error}_V &\sim \text{ZeroInflatedBeta}(\mu, \phi, \pi) && \textit{Beta-distributed likelihood} \\ \mu[i] &= \beta[\text{vis}[i]] && \textit{mean submodel} \end{aligned}$$

4.2.3 Modelling individual variation, *i.e.* random effects

At this point, our mean submodel for a given observation incorporates information about the average mean absolute error for the visualization type that the observation came from, but nothing further. Therefore, we cannot yet fit a model assessing individual variation without running afoul of what Hurlbert called *pseudoreplication*, which results from asserting impact from an effect that is not statistically independent [58]. In this case, we have not separated out the effect of individual variation, so we cannot reasonably model its effect.

We can incorporate this using *random effects* in the model, employing a new term we will call $RE[\text{participant}[i]]$ that acknowledges that the participant making the observation can deviate from the average. This also means that we are *no longer modelling mean absolute error*; the **mean** was referring to aggregation across individuals, which is no

longer needed as we are separating out individual contributions in the mean submodel.

Because the chart types vary meaningfully in characteristics (as indicated by the varying chart categories discussed previously) and because the psychophysical properties of chart types may prompt varying strategies by individuals [59, 61, 100, 74], we also incorporate the visualization type into this random effects term: $RE[\text{participant}[i], \text{vis}[i]]$.

The effect of a participant and visualization in this term may be correlated with each other. What if, for example, an individual happens to be particularly capable at assessing areas of circles, and consequently performs better than expected on both bubble and pie charts, and the absolute error from their judgments varies accordingly? To account for this, a covariance matrix is needed, which captures the extent to which each pair of variables is likely to over- or under-perform the mean together. This is a difficult endeavor given our current model; the mean parameter μ reflects an aggregate of possible values for a Beta distribution, and is therefore bounded by 0 and 1 exclusive and inappropriate to model as Gaussian. A *link function* can be used to map the (0,1) space to $(-\infty, +\infty)$. This will allow us to treat the random effects as having a multivariate normal distribution. The logit function serves as a suitable link function, and gives us the following model characteristics:

$$
\begin{aligned}
V = 4 &: \text{number of visualization types} \\
P = 109 &: \text{number of participants} \\
D = 20 &: \text{number of datasets per visualization type} \\
R = 15 &: \text{number of repetitions per dataset} \\
i \in \{1 \dots VPDR\} &: \text{index of observations (trial-level errors)} \\
\text{vis}[i] \in \{1 \dots V\} &: \text{visualization associated with observation } i \\
\text{participant}[i] \in \{1 \dots P\} &: \text{participant associated with observation } i
\end{aligned}
$$

$$\begin{aligned} \text{Absolute error}[i] &\sim \text{ZeroInflatedBeta}(\mu[i], \phi, \pi) && \textit{likelihood} \\ \text{logit}(\mu[i]) &= \beta[\text{vis}[i]] + RE[\text{vis}[i], \text{participant}[i]] && \textit{mean submodel} \end{aligned}$$

And, for $v \in \{1 \dots V\}$ and participant $p \in \{1 \dots P\}$:

$$\begin{bmatrix} RE[1,p] \\ \vdots \\ RE[V,p] \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \Sigma \right) \qquad \forall p \in \{1 \dots P\} \qquad \begin{matrix} \textit{correlated} \\ \textit{random offsets} \end{matrix}$$

Our earliest model using the Normal distribution incorporated a mean submodel, and so we have updated that accordingly. However, we have not yet accounted for the precision (ϕ) or the probability of a zero (π). The model currently assumes a fixed value for each, meaning that it assumes that every participant has identical consistency between themselves and across visualizations, datasets, and repetitions (in the case of ϕ) and identical chances of having a perfectly accurate estimate for a given trial (in the case of π).

We incorporate submodels for ϕ and π mimicking that of the mean submodel. The probability space for the Bernoulli distribution is (0, 1), making it appropriate to use logit space for it as well. For the precision parameter ϕ, it is bounded by $(0, +\infty)$, so we will instead use a log scale. These adjustments give us:

$$\begin{aligned} \text{Absolute error}[i] &\sim \text{ZeroInflatedBeta}(\mu[i], \phi[i], \pi[i]) && \textit{likelihood} \\ \text{logit}(\mu[i]) &= \beta_\mu[\text{vis}[i]] + RE_\mu[\text{vis}[i], \text{participant}[i]] && \textit{mean submodel} \\ \log(\phi[i]) &= \beta_\phi[\text{vis}[i]] + RE_\phi[\text{vis}[i], \text{participant}[i]] && \textit{precision submodel} \\ \text{logit}(\pi[i]) &= \beta_\pi[\text{vis}[i]] + RE_\pi[\text{vis}[i], \text{participant}[i]] && \textit{zeros-probability submodel} \end{aligned}$$

Having developed these submodels, we will also establish our approach for modeling

for their associated random effects. Earlier, a multivariate normal distribution with a covariance matrix was used to capture possible correlations in the mean submodel. These submodels could be correlated with each other, *e.g.* a participant who shows a negative individual effect on mean error (*i.e.* has lower-than-average mean error) could have a corresponding positive effect on the zero-error parameter (*i.e.* higher-than-average chance of a zero-error estimate), and so we will have all random effects share a covariance matrix:

$$\begin{bmatrix} RE_\mu[1,p] \\ \vdots \\ RE_\mu[V,p]] \\ RE_\phi[1,p] \\ \vdots \\ RE_\phi[V,p]] \\ RE_\pi[1,p] \\ \vdots \\ RE_\pi[V,p]] \end{bmatrix} \sim \text{Normal} \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \Sigma \right) \qquad \forall p \in \{1 \dots P\} \qquad \begin{matrix} \textit{correlated} \\ \textit{random offsets} \end{matrix}$$

4.2.4 Implementation in BRMS

To summarize, we've identified the following elements that need to be brought into our models:

- Average effect from each visualization type, consistent across each group
- An effect (*i.e.* variation) for each participant, which changes based on the visualization type.
- A shared covariance matrix across submodels

This modelling can be performed in *brms* [23], a library inside R [104] that uses Stan [24] to fit models. Stan is a probabilistic programming language and Markov chain Monte Carlo sampler, and will be discussed more later.

The model, specified using the **brm** function, takes its framework from the more

widely-used **lme4** syntax [8]. This syntax follows the form

$$response \sim pterms + (gterms \,|\, group)$$

Here, *pterms* refers to population-level effects; these are consistent across observations [23]. *gterms* are group-level effects (here, we've referred to them as random effects), which vary across the entities of interest.

So, in keeping with our summary above, we render our model in *brms* as:

brm (

brmsformula (

$abs\_error \sim vis + 0 + (vis + 0\,|pt|\,participant)$,

$phi \sim vis + 0 + (vis + 0\,|pt|\,participant)$,

$zi \sim vis + 0 + (vis + 0\,|pt|\,participant)$

),

family = zero_inflated_beta,

...

)

In this syntax, we declare the **family** to be zero_inflated_beta, and provide a submodel for each required parameter, as discussed previously. The concept of a family in *brms* will be discussed later. The computation requires the visualization types to be encoded for modelling population-level effects, and so *vis* is present in each to declare that it should be considered as a variable term. Because visualization is a constant effect for each observation made from that visualization type, this should be understood as an intercept variable. The addition of +0 informs *brms* that **there should be no base population intercept, *i.e.* mean** from which the categorical levels have *offsets*; instead, *brms* will produce uniquely understood intercepts, which can be understood as an offset from zero. This is called *one-hot encoding* and is an important consideration for assigning priors, as using an offset from a non-zero value would require confusing modifications.

The parenthetical notation in this syntax covers the random, *i.e.* group-level, effects. Here, $vis + 0$ denotes that participants will not affect the constant effect of each

visualization, but will individually deviate from that constant effect. should be understood independently from each other, when considering their impact on each participant. The notation $|pt|$ is specific to *brms*, and was written to address the difficulty in having group-level effects be correlated across formulas (or submodels, for our terminology), provided that they share a grouping factor [23]. The *pt* identifier is arbitrary, and here is shorthand for the shared grouping factor, 'participant'.

4.2.5 Establishing Priors

A Bayesian model requires specifications of priors, which, as previously discussed, capture the knowledge we possess of the context when initiating our research. Because our research is intended to recreate and then provide further nuance to the non-Bayesian approach implemented by Cleveland & McGill, and subsequently by Heer & Bostock, we do not intend to put forth a fully informative prior that will have a strong effect on the probability distribution resulting from the analysis. Doing so would predispose our posterior distributions to resemble their research, which would eliminate the confirming power of replication. Instead, we use *weakly informative* priors, as described by Gelman *et al.* and discussed earlier in this thesis [48]. Because these priors do have some pressure on the resultant distribution, they should be chosen to reflect the general form of the data as understood from previous research in the domain *without* a bias towards the results from previous literature; in practice, this means that we will not closely investigate prior results, but will use them where possible to ensure that our priors are reasonably scoped.

Before establishing priors, we note that the covariance matrix will be decomposed for modelling into standard deviations for each random effect and a correlation matrix, and a prior will be set for each component part. We set our priors as follows:

- $\beta_\mu[v] \sim \text{Normal}(-2, 1)$: Here, we assign the same prior for mean error to each visualization type, on a **log-odds** scale. Because mean error was the focus of prior research, we can use that context to identify a range of values that should feasibly contain the mean absolute error. With our declared use of *weakly informed* priors, we choose our prior conservatively, to capture a range of values exceeding what

was identified for all visualization types investigated by the prior research we are considering. This prior is in logit-space, and covers from approximately [-4, 0] inside its 95% central interval. In the [0, 1) percentage space that we are considering, this covers roughly [1.7, 50]. We decide on this prior with awareness of the ranges found by Cleveland & McGill and Heer & Bostock, which saw results ranging from about [1, 3] in their adjusted $\log_2$-absolute-error scale. Converting back, using $2^error - 1/8$, yields a range of (1.875, 7.875). An average error of 50 is large enough to be reliably outside the bounds of expected results for the proportion judgment task we have posed, especially in comparison to the log errors seen from previous research, and our lower bound also exceeds the range that prior research identified.

- $\beta_\phi[v] \sim \text{Student}(5, 0, 10)$: Again, we set here the same prior for each visualization on a **log** scale: this time, for the precision parameter. We lack prior knowledge regarding how precise the estimates made by participants were. In this case, we'd have been specifically interested in research that grouped estimates by visualization type and observed their variance. Because we cannot use prior research to appropriately bound this prior, we choose a Student's t-distribution rather than a normal distribution. The prior is set in log-space, but covers roughly (1e-11, 1e11) when that transformation is inverted; this range is, again, intentionally conservative.

- $\beta_\pi[v] \sim \text{Normal}(-2.5, 1.25)$: The prior for the probability of a participant getting a 0-error (exactly correct) response for each visualization, on a **log-odds** scale. The goal here is again to make highly conservative estimates in light of limited prior data. This interval covers roughly [-5, 0] in log-odds space and roughly [0.7, 50] when inverting the transformation. Because there are 100 possible responses to the experimental prompt, we expect that even with random guessing, not less than 1% of responses would see zero error (we set our lower bound to 0.7 for the sake of having an integer in log-odds space), and that more than half of estimates being correct would be implausible.

- $\sigma_{RE_p} \sim \text{half-Normal}(0, 0.5)$: The prior for the standard deviation of the random

effect from each participant. This covers from [0, 1] in a log scale, making it a fairly wide prior.

- $\rho_{RE} \sim$ LKJ(4): For the correlation matrix, we use the Cholesky factor of a Lewandowski-Kurowicka-Joe prior [83], which is recommended as a standard probability distribution for correlation matrices [45]. A Cholesky factor of 1 sets all correlation matrices to be equally likely, with higher Cholesky factor limiting the likelihood of extreme correlations. The *lkj_corr* function from *ggdist* [65] was used to adjust this prior to ensure that the fitting process does not suffer from exploring too broad a range of possible correlation values; ultimately, a Cholesky factor of 4 is used to constrain the results somewhat. In practice, this constraint has a minimal effect on the posterior distribution obtained during the modelling process, but yields meaningful improvements to the efficiency of the MCMC algorithm.

4.3 Fitting a model using brms

The complexity inherent in our particular problem prompted Cleveland & McGill to call the thought of modelling it (especially the covariance matrix) a "substantial chore" where "mathematical deviations of sampling distributions [are] intractable". The *brms* package is based on an implementation of Markov Chain Monte Carlo (MCMC), which is an algorithm that uses sampling techniques to learn about "unknown and usually complex target distributions" and allows us to surmount Cleveland & McGill's chore [92]. MCMC relies on random movement from a current state, where the realm of movement is the range of possible parameter values, and the motion traverses the posterior probability of each. Ideally, a Markov chain shows signs of *stationarity*, which mean that the path traversed by the chain remains centered around a particular mean value, suggesting that it has converged on a likely parameter value. It also shows *mixing*, which means that each step of a Markov chain makes meaningful deviations within the posterior distribution (while remaining centered around the mean value). Because Markov chains act based on a current state and the movement is dictated by the probability of each parameter

value, it is possible for steps in different Markov chains to converge in different regions of the parameter space (say, in the case of a multimodal target distribution), or for high autocorrelation to limit the chain's ability to have steps substantially from their predecessors [118].

In implementation, *brms* is built on 'Stan'. Stan uses a particular method of MCMC called Hamiltonian Monte Carlo (HMC), which is more computationally expensive but more efficient. The default HMC algorithm used by Stan is called NUTS (No-U-Turn-Sampler) because it is built to minimize random-walk behavior that leads to MCMC algorithms collecting redundant information, and requires specifications of several parameters alongside the model itself [55].

4.3.1 Model specifications

We are most directly interacting with *brms*, which provides a layer of abstraction between us and Stan (which provides a layer between us and NUTS). This interface provides a number of "family functions", which are designed to facilitate use of commonly seen models [19]. Zero-inflated beta is one of these models, and it defaults to the link functions we arrived at in our model: logit for the mean and zero-likelihood, and log for the precision parameter.

Using the **brm** function, which initiates model fitting, requires specification of the model and formula as well as several other Stan-specific arguments:

- **Number of chains**: The number of Markov chains employed during the computation of the model. Increasing the number of chains and seeing each converge to the same target distribution provides confidence that the sampling approach is functioning properly. Richard McElreath suggests that "3 or 4 chains is conventional"; we use 20, to make use of the 20 cores in the computing system used
- **Number of iterations per chain**: This number defaults in Stan to 2,000, but the appropriate number is highly contingent on the model and parameters of interest. Because we are modelling all of mean, variance, and covariance, a higher number is merited; we double it to 4,000.

- **Number of warmup iterations**: When sampling using Stan's HMC implementation, the initial samples have not begun to converged to a parameter space and lack information about the parameters that the algorithm uses to perform effective iterations [114]. The NUTS algorithm uses warmup iterations to establish these parameters. We use the default value of 1,000.

- **adapt_delta**: This parameter helps to control divergent transition in HMC, which are cases where Stan's sampling saw a mismatch between its projections for the path it would follow and its true path. A higher value for adapt_delta prompts the sampler to take smaller steps on its path, limiting the chances of a problematic deviation but increasing the necessary computational time [114]. We use Stan's default value of 0.95 to balance the reliability of the sampler with the practical consideration of runtime.

In order to assess mixing, Stan uses the $\hat{R}$ metric, which compares the standard deviations between and within chains. Vehtari *et al.* propose $\hat{R} < 1.01$ as a standard for using a sample, a threshold that our model meets [118].

Stan also provides information on the number of effective samples, so called because a series of samples from a chain that are highly autocorellated will not provide as much information about the posterior distribution as samples that are less correlated, and so, effectively, are worth less to the model.

In practice, because the fit model represents a sizeable amount of data (over 800MB), data thinning can be employed to facilitate sharing or storage. Thinning has little practical import aside from reducing the amount of data needed, and is also facilitated by Stan [114]. For replication, our results use a thinned model, which has 6,000 post-warmup samples and an effective sample size ranging from 4,200-6,300 for participant-level variables and from 3,100-5,400 for population-level variables.

Chapter 5

Results

Before making full use of our posterior distribution, I will provide context to this thesis and establishing the validity of our data collection and modelling approach. Afterwards, I explore the new information that our model provides, and determine the answer to the two hypotheses: *Are conventional rankings of chart types universal?*, and, *Can we arrive at a ranking of individuals based on their perceptual accuracy?*

5.1 Building From A Collaborative Research Project

This thesis began life as part of an effort with Professors Lane Harrison and Matthew Kay *et al.* to apply Bayesian analysis to the problem addressed by Cleveland & McGill, funded in part by a grant from the US National Science Foundation (#1815587, #1815790). The drafting of the thesis was delayed, but the group continued to build off of our discussions and wrote a paper titled *The Risks of Ranking: Revisiting Graphical Perception to Model Individual Differences in Visualization Performance*, which was accepted to the 2022 IEEE conference. The paper was written to implement "Bayesian multilevel regression...to explore individual differences in visualization skill", which I will proceed to do here as well [31]. I recreate here the results and figures, as in the analytical approach, with a more thorough description of the model output and thought processes than is possible in an academic paper. This serves the dual purpose of making the statistical process in this thesis somewhat more accessible for a reader, and also serving as reference documentation

for others seeking to approach their research area in a similar manner. In addition to the published work, I add a discussion on identifying a hierarchy of individuals in terms of skill, as well as an exploration of modelling true proportion as a relevant parameter.

5.2 Comparison to Cleveland & McGill and Heer & Bostock

We recreate Cleveland & McGill using our data and their mechanics in order to validate our collected data and to place our charts in the context of their data, as the specific implementation of chart types varied slightly from study to study. This process uses bootstrapping, or, sampling with replacement, and calculation of the means of midmeans to generate a sampling of sample errors to calculate mean and standard deviation. Because Heer & Bostock's study bears close resemblance to Cleveland & McGill as a crowd-sourced replication, we include data from their research as well. The upper panel of Figure 5.1 is a rendering of findings from Cleveland & McGill's paper, and the middle portion from that of Heer & Bostock's data; in both cases, the presentation is borrowed from Figure 4 of Heer & Bostock's paper; our replication is present in the bottom panel [29, 54].

Each study contains a comparable implementation of bar charts, making comparison easy; all have comparable results. Heer and Bostock made pie and bubble plots, which serve as reference points for ours. Although Cleveland & McGill also studied pie charts, they used a distinct task format, making it unsuitable for direct comparison. Variations of stacked bar charts appear in each study as well, although the presentation varies slightly in each. We see that *Bar* is consistently lowest error across each study The error seen in our *Pie* chart is lower error than that of our *Stacked Bar*, while Heer & Bostock's shows stronger overlap between the two. In either case, the mean of midmeans covers a similar error range. *Bubble* shows highest error margin in both studies where it is present. For each of Pie, Stacked Bar, and Bubble, we see similar error estimates, mostly showing overlap, and almost entirely within 0.5 of each other in the adjusted log scale being used.

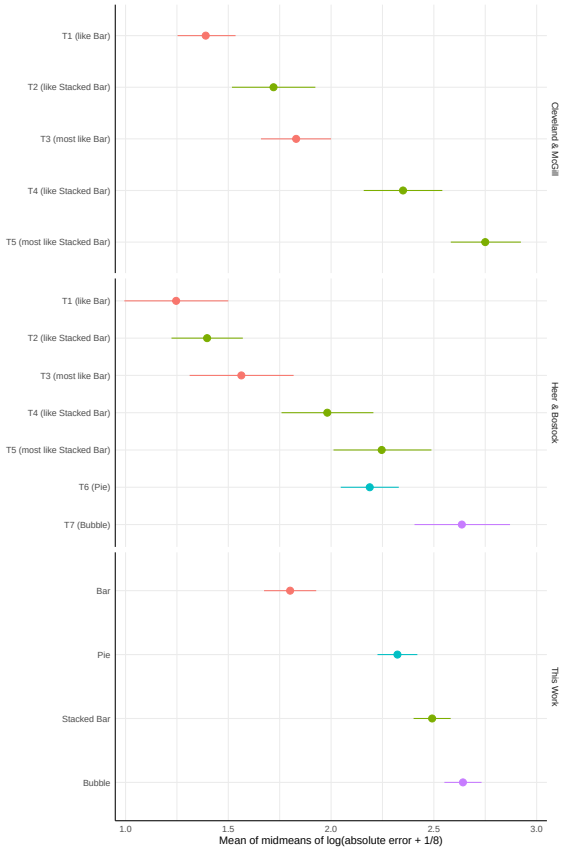


Figure 5.1: Comparing our results to those obtained by Cleveland & McGill and Heer & Bostock

This makes it difficult to assess a strong advantage from selecting one visualization rather than another.

As discussed previously, there are some limitations to this methodological approach. It fails to consider the full range of participant error due to the use of midmeans; we have lost information from the lowest and highest quartile. The outliers removed by doing this do contain information about graphical perception, albeit information that can contain outliers that might render the use of aggregate statistics problematic..

From a designer perspective, however, information about high perception error could be particularly impactful; while low error represents particularly effective communication, we already see that the adjusted log error is fairly low, so this does not represent a significant change in understanding. High perception error is not similarly bounded by 0 on our graphs, and leaves open the possibility for individuals to dramatically misunderstand the information presented to them. As in our example of CFPB activity earlier

in this paper, circumstances where individuals take no meaning, or worse, a misleading conclusion, could easily be seen to affect behavior. We note here as well the difficulty even from an academic perspective in translating the adjusted log scale to the actual metric of interest, that being percentage points of error. Lastly, as per the motivation of this paper, this methodology does not enable individual analysis or between-participant exploration, making it unfeasible to answer our key question: *Do relative rankings differ across people?*

5.3 Exploring the output of our model

Here, we demonstrate the possibilities of our Bayesian approach, both in achieving the same research goal as previous research and in providing new insights. We entered into the modelling process asserting that the visualization types are meaningfully different from each other, at least qualitatively, and therefore treated each as having a distinct impact on the judgments made by participants. The output of the model contains the posterior parameter distributions for group-level effects, which capture the impact of variation from participant to participant, alongside population-level effects. This latter resembles the previous implementation in that they similarly present a high-level assessment of how the visualization types differ without the context of random effects.

5.3.1 Replicating the visualization comparison

As part of the validation process for our model, we can compare our outcomes to those from previous research. Because our model's results are not on the same adjusted log scale, they cannot be directly compared to the prior studies we have considered or to our recreation of them. However, we do expect that the ranking of chart types is comparable. Because we modelled mean error and probability of zero error as normal distributions, we can present credible intervals comparing the likely error for each visualization in a similar graphic, as seen in Figure 5.2.

The posterior output for MAE by vis type is almost identical, qualitatively, to

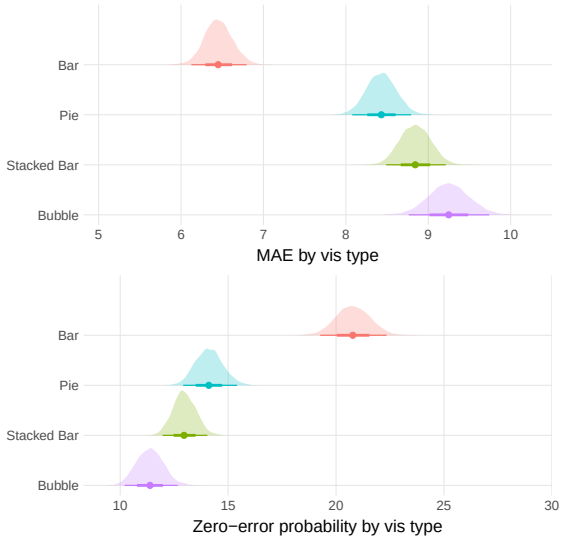


Figure 5.2: Posterior distribution of MAE, and zero-error probability, for each visualization type

that obtained using the previous approach. *Bar* remains the most high-performant chart type; the others go *Pie* > *Stacked Bar* > *Bubble*, but in a tight grouping that does not meaningfully differentiate them. A similar effect is seen for zero-error probability, where *Bar* outperforms the other chart types, which are clustered together.

The third population-level parameter is ϕ, which captures variance. In the context of our Beta distribution, ϕ is the sum of two shape parameters; while necessary for our modelling, it does not lend itself as readily to comparing visualizations.

5.3.2 Preparing To Move Beyond Cleveland & McGill

We've now validated the outcome of the Bayesian approach, and should consider the other avenues that are opened by our posterior distribution. This distribution represents a reasonable range of values for each parameter of interest; in addition to the population-level effects discussed above, they represent the variation that is expected by participants across visualization types, between participants, and between a participant's own responses to repeated stimuli (even when the conditions are exactly repeated). Because we have access to a *range* of values, the model can be used to simulated expected outcomes for the data used to fit the model with what are called *fitted* draws (randomly generated samples), assessing the expected behavior of the average participant within our sample. The model

can also be used to produce *predicted* draws to simulate the behavior of "new" participants with their own (simulated) idiosyncrasies. The latter has higher variance due to its incorporation of residual error (*i.e.* error that our model has identified as being unexplained by the variables we have accounted for), while the former's only source of variance is the uncertainty inherent in randomly drawing from a range of values. Our model simulates the expected outcome for posterior predictive checks, which takes the form of numerous fitted draws from the posterior distribution. This is the approach that was used to create Figure 5.2.

5.4 Complementing Confidence Intervals With Error Distributions

Synthetic data created using fitted draws allows us to examine the range of plausible mean errors. By representing this as a cumulative distribution function (CDF), which represents the probability that a parameter value is less than or equal to a particular value along a scale, we see the percentage of responses that were less than or equal to a given absolute error. CDFs make it easier to answer questions about the population as a whole, *e.g. What proportion of the population can be expected to make errors of less than 10pp?*, which we could not answer previously.

After randomly generating population-level parameters for μ, ϕ and π for each visualization from each of our 6,000 postwarmup samples, we use the GAMLSS (Generalized Additive Models for Location Scale and Shape) package to calculate CDFs using these parameters [106], as shown in Figure 5.3. Our creation of figures makes heavy use of ggplot2, whose *stat_lineribbon* function helps to generate shaded bands to show 66% and 95% credible intervals for the error density is estimated by the model. For comparison purposes, we overlay the model-generated CDFs onto a CDF of participant error from observed data from our experiment, represented by an opaque line. In doing this, we compare the output of our model at the population-level (so, considering each visualization separately and excluding participant-specific random effects) to the aggregate performance of the

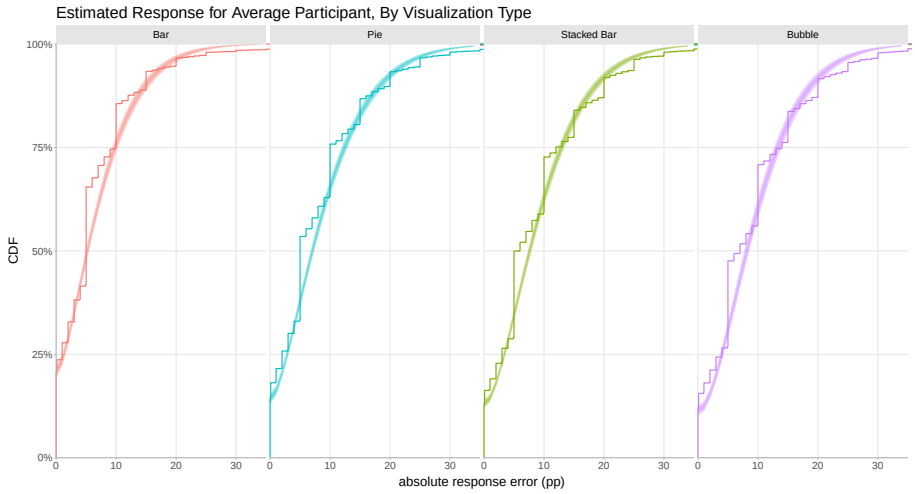


Figure 5.3: Cumulative distribution function of MAE for each visualization type

participants of our experiment. Showing the observed data alongside our model output shows how effectively our model (predictions) fits reality (the empirical data we collected) We found that most errors (about 98.5%) are less than 35pp, so that was used as an upper bound for the error scale.

When overlaid, the most obvious difference between the model-generated data and the observed data is the continuous nature of the former and the discrete nature of the latter. Discrete values are to be expected, given the experimental prompt that required entries of integer values. In some cases in the CDF, we see particularly sharp jumps; this is a consequence of participants rounding their answers to multiples of 5 [7, 113]. As an example, if the true proportion is 30pp, we expect to see a significant amount of answers that are either 25pp or 35pp, more so than any of the intervening integers.

This effect is made more prominent because we elected to present in the experiment true proportions that were multiples of 5 in 90% of cases; to be precise, 18 of the 20 true proportions. Had we not done so, we would have been more likely to observe errors that were not multiples of 5. For example, if the true proportion was 28pp, we are likely to have received estimates of 25pp (with an error of 2) or 30pp (with an error of 3), diminishing the sharpness of the jumps currently seen. The model does not account for these jumps, leading to some discrepancy between the two sources of data around the area of the sharpest jumps (5pp and 10pp). Since the jumps are seen here more as a function of our

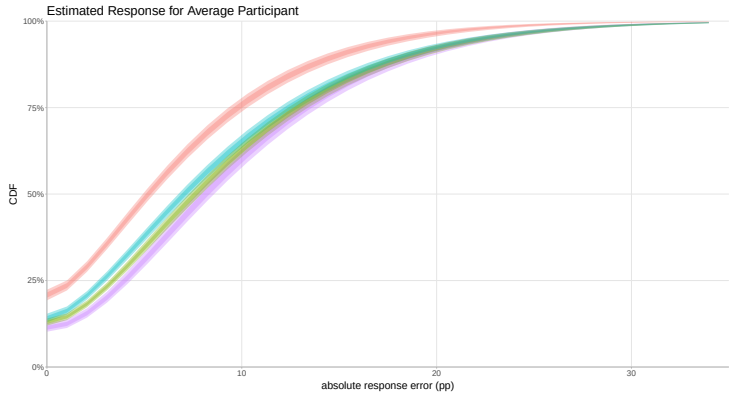


Figure 5.4: CDFs for each visualization, based on modelling the average participant

data solicitation methodology than a characteristic of estimating proportions, modelling this behavior in greater fidelity is left as an opportunity for investigation in future work.

Our observed data, and so our model, both show relatively high rates of 0-error responses (about 15% for bubble plot at the lowest, and around 20% for bar charts at the highest). If a higher number of the true proportions were not divisible by 5, we would have seen this rounding behavior work against zero-error proportions. The model CDFs and observed data both suggest that median error is roughly 5pp, and that the errors skew small; more than 75% of the density is not greater than 15pp, even in the worst performing chart.

5.5 Comparing Error Distributions

Direct comparison between the CDFs will also provide useful context for comparing the error obtained from each visualization. Our interest in comparing CDFs is to see the difference in their density at different levels of mean absolute error. Overlaying the model estimates for each visualization makes this an easier task, seen in Figure 5.4.

It was previously clear that *Bar* outperformed the other chart types, but now we see more clearly that the biggest advantage of bar charts is seen when error is in the range of 5pp to 10pp, where the gap between it and the next chart type is largest. *Pie* is demonstrably better than *Stacked Bar* and *Bubble* charts, particularly at the same 5pp range; however, the difference remains small throughout, rarely exceeding 5% of the cumulative density. *Bubble* and *Stacked Bar* track each other closely and do not show

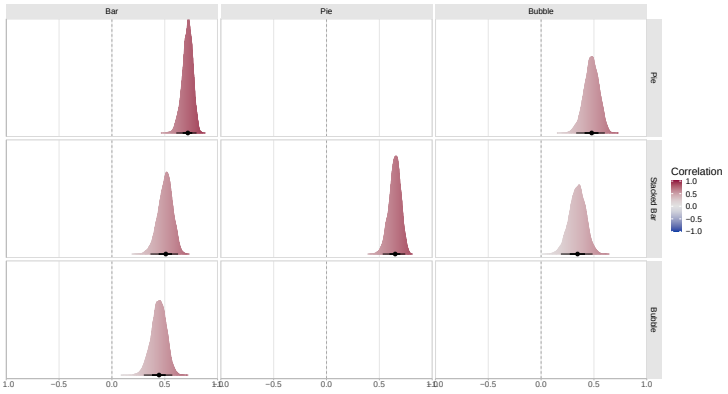


Figure 5.5: Correlation between visualization types, for individuals

any practical difference, perceptually. These results are broadly in keeping with canonical rankings, with the added benefit of showing the scale and scope of their differences.

5.6 Do Individuals Show Consistent Strength In Performance?

We will now take advantage of our model's ability to examine performance at the individual level. First, we consider the various visualization types in the context of single individuals, before proceeding to between-individuals performance. In each case, we simulate participants from our model. Because we hypothesized that above-average performance in one chart type could be correlated to above-average performance in another chart type, our model included a covariance matrix to capture correlation between all individual-level parameters. This is one of the parameters that the posterior distribution can provide a range for. Aggregating individual-level correlations across our fitted draws gives the correlations and their uncertainty.

Figure 5.5 was created in ggplot with non-overlapping ridgeline plots, using the *stat_pointinterval* function from the *ggdist* package to display the point estimate for correlation and the interval representing the uncertainty of the correlation parameter [65]. We see in every visualization type positive correlation, meaning that performance varies consistently across visualization types, either positive or negative relative to the

mean. This effect is strongest for *Pie* to *Bar* and *Pie* to *Stacked Bar* (approximately 0.65 and 0.62, respectively), and is lower but still positive for other chart types.

5.7 How significantly do individuals vary from each other?

We are now confident that a consistent effect can be seen within an individual's performance across chart types, meaning that some participants will be seen to reliably outperform the mean and others will reliably under-perform the mean. This suggests that meaningful variation will be observed when transitioning between individuals, and that individual variation is a demonstrable cause of meaningful changes in mean absolute error for a given visualization.

The between-person variance in mean errors for a given population can be derived from mean error data for that population; assuming a normal distribution, calculating the standard deviation is straightforward. However, this only represents a point estimate of standard deviation. In order to clarify the interval of values where the standard deviation might appear, we need information for multiple populations, each with their own set of participants causing their own between-person variance. Once we have this collection of information, we can establish a distribution describing our knowledge about the standard deviation. Our posterior distribution allows us to simulate multiple populations by creating batches of "new" participants and treating each batch as a population. In the IEEE submission, 6000 simulated participants were generated from each of the 6000 post-warmup draws.

For this thesis, I simulate this with 100 participants in the interest of an independent recreation; this data is shown in Figure 5.6.

The between-person variance for each chart for each chart type is strikingly similar for *Bar*, *Pie*, and *Stacked Bar*. *Bubble* shows the highest standard deviation, about 0.5pp greater than the other chart types. The results for the mean match those in the IEEE

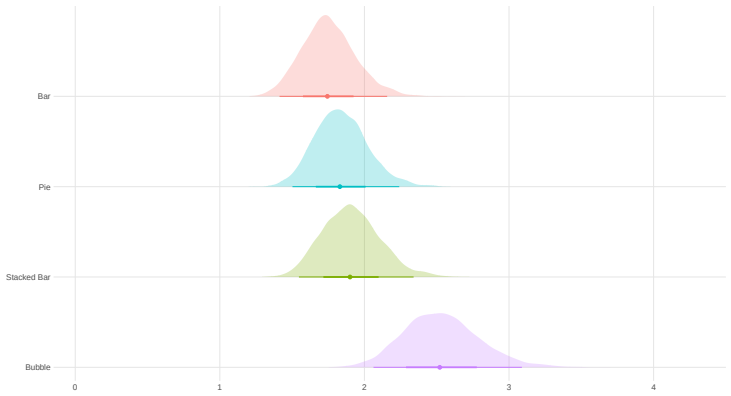


Figure 5.6: Standard deviation between simulated participants in MAE (pp)

submission. The uncertainty is significantly broader, as expected due to the vastly reduced amount of data being used, but shows the same sense of overlap between chart types.

5.8 Establishing A Skill Hierarchy In Individuals

We have observed that there is significant between-person variance, and that performance relative to the mean is consistent. Using this information, we can then reasonably conclude that some high-performant individuals will show better performance across all vis than low-performant people, in terms of mean error. This allows us to state unequivocally that some participants have the ability to more accurately perceive information in these fundamental graphical forms, in answer to the second of our two hypotheses of interest. Displaying it graphically will give us more confidence that there are specific individuals who set themselves apart in terms of accuracy, either positively or negatively.

We are interested in investigating both the accuracy of a participant's performance, *i.e.* mean absolute error, as well as their precision, *i.e.* how consistently a participant performed in their estimates. An individual with a low mean absolute error could have consistently low error, or a mixture of extreme accuracy and higher errors; the precision is less relevant for ranking, but more for characterizing the performance of each individual.

By completing multiple fitted draws for each participant, the random effects model generates information to determine a distribution of plausible mean absolute error values as a function of some central tendency in tandem with the participants' precision and zero-probability submodels. For example, an individual with very high output from the

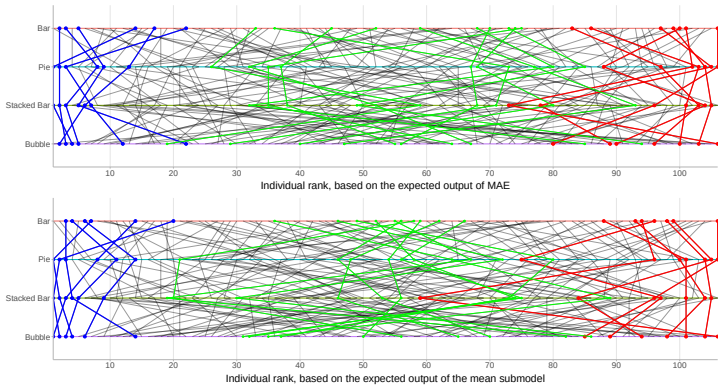


Figure 5.7: This chart shows the consistency of high-performant participants (blue), middling-performant participants (green), and low-performant participants (red)

mean submodel but correspondingly high probability of 0-error will, ultimately, show a lower mean absolute error. In other words, we have information about their average tendency, and the output of the model incorporates their consistency and likelihood of zero error as well. So, by developing a ranking based on an individual's mean error and the mean absolute error that they show, we can draw conclusions about their accuracy and consistency.

The top and bottom 5 participants in terms of expected MAE and μ were colored blue and red, respectively; the middle 5 from each are colored green. The same individuals are represented on each plot, in each color. In both cases, we see that high-performant individuals cluster together tightly. Low-performant individuals show similar consistency except for a pair of individuals who show unexpectedly high ranks in *Pie* for one and *Stacked Bar* for the other. The middling-performant individuals show less consistency in their outcomes.

5.9 What matters more? A change in visualization type, or a change in participant?

Because we observed earlier that *Pie*, *Stacked Bar*, and *Bubble* were difficult to differentiate from each other, and because we now know that standard deviation between

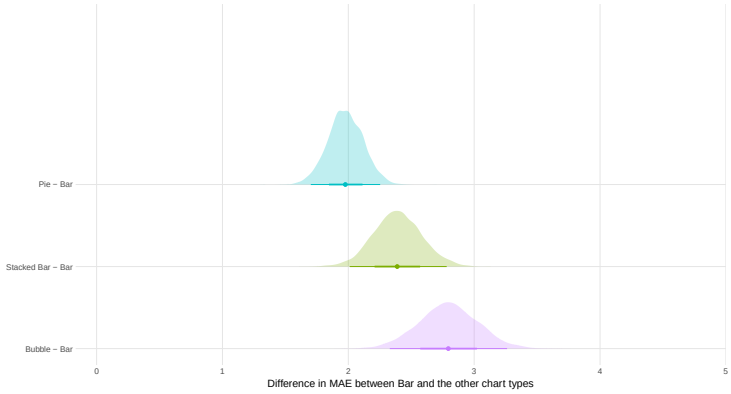


Figure 5.8: The difference between Bar charts and the other charts is readily observable. A quick observation of the chart shows that the error between the non-Bar chart types is, on average, less than 1pp.

participants is not insignificant, it is worth considering if changes in participants might have more power to explain mean error than the visualization type.

We have already modelled the mean absolute error for each vis type in Figure 5.2. We can effectively compare the difference between each chart by using one visualization as a common reference point. We choose *Bar* for this purpose, because our early assessment showed that it was clearly differentiated from the other chart types. Figure 5.8 captures this variation. Taken in tandem with Figure 5.6, we see that the effect of transitioning from *Bar* to another chart type slightly outweighs the standard deviation between participants, *i.e.* changing participants is likely to be less impactful to expected mean error than changing from bar charts to a different visualization. The change resulting from shifting participants could be positive (lower mean absolute error) or negative (higher mean absolute error), but in either case the magnitude would be lower. However, this is not true for transitions between other chart types.

5.10 How Universal Are The Canonical Rankings?

This question is the first of our two research hypotheses, and is the core question in our investigation of Cleveland & McGill's results. Our first task is to observe the relative performance of each individual (their mean absolute error, as estimated by the model)

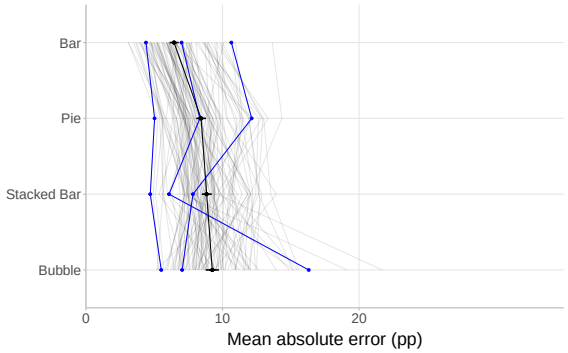


Figure 5.9: Each line represents the mean and 95% uncertainty interval for each participant; the solid black line represents the average participant

for each chart type, compared to other individuals.

Parallel coordinates plots are well-suited to this task, as seen in Figure 5.9. As a reference point, we show the average participant as a solid black line. In Figure 5.9, the chart types are ordered top-to-bottom by the canonical rankings. If all participants held to these, we would not see the crossing patterns in the plot that, here, indicate a decrease in mean absolute error. Because bar charts consistently show to have the best average performance while the other charts have a tighter grouping of error, we are particularly interested in the potential existence of crossing patterns going from *Bar* to *Pie*, which we do see. We also note the multitude of crossing patterns in the other chart types, suggesting that, as suggested above, variations in individual estimates outweigh whatever overarching effect the visualization type has. Figure 5.9 also highlights three individuals who show behavior that highlights our goal: the left-most individual whose performance shows little variation, the central individual whose performance in *Stacked Bar* exceeds their performance in *Bar*, and the right-most individual who outperforms *Bar* charts with both *Stacked Bar* and *Bubble*.

To capture more directly the population of individuals who do not share the canonical rankings, we generated fitted draws from each of our 6000 post-warmup samples to model participant behavior for each visualization type, and then calculate the internal ranking of visualization type error (1 through 4) for each simulated individual. An alluvial chart was generated using the *ggalluvial* package from those rankings [18]. The alluvial diagram in Figure 5.10 shows how many individuals ranking of individual's mean

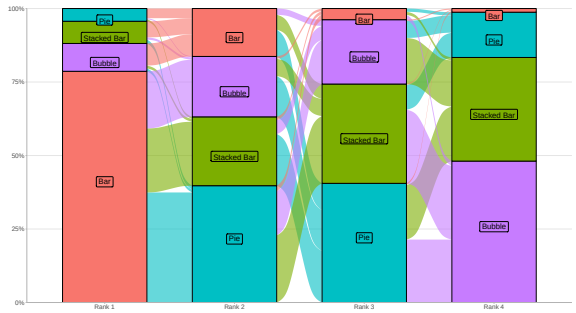


Figure 5.10: Rankings of visualization type for simulated experimental participants

error follows the common guidance. More than 75% of simulated individuals show *Bar* as their highest-performing chart type, which represents a significant proportion of the population. However, even if it is only a quarter of the population that deviates from the established rankings, this still represents a considerable amount of individuals.

Additionally, the heavy inter-mixing of the other chart types is expected based on our findings in Figures 5.8 and 5.6. Because between-person variance seems to outweigh between-charts variance, we would expect to see this behavior here given our focus on participant-level effects.

5.11 Exploring the relevance of true proportion as an effect

Having addressed our hypotheses of interest, I now turn to incorporating the true proportion into the model. As noted previously, this section is exploratory work not present in the IEEE submission, and does not provide insight into the primary hypotheses of this paper; accordingly, the analysis and discussion is contained within this section to keep it from being a distraction.

This exploration is stimulated by Cleveland & McGill's comments about negative bias for several chart types. This is, to recount, that "When the true percentages are in the range of 25-50, subjects tend to underestimate values for [bar and pie charts]", and noted a similar negative bias for stacked-bar charts when the true percentages were

between 30 and 70 of "some dependence of log error on the true percent" [29]. Cleveland & McGill's use of aggregate statistics meant that they could not readily investigate this notion of true proportion-driven bias, but exploration using the Bayesian model is relatively straightforward.

Before we take steps to incorporating true proportion into the model, we consider how it fits into the model that has already been implemented. When we established our initial framing of the model, we determined a posterior distribution for the fixed effect from each visualization type, We additionally modelled performance distribution for each participant, as well as the change in error to be expected when changing participants, with participants identified as a source for variation between estimates.

A priori, we conjecture that true proportion will have some population-level effect on absolute error, similar to visualizations, because of the visual difference the true proportion will have on the appearance of the data points of interest. Additionally, if Cleveland & McGill's observations about negative bias hold generally true, there is some effect that is seen across multiple chart types across broadly overlapping ranges, suggesting that this bias is not unique to any particular visualization or context. It is not unreasonable to think that individuals will vary in how they respond to different true proportions, given our findings as to the significant between-participant variation; however, I risk overextending myself by investigating too many levels at once, and prefer to leave its inclusion as a random effect to future work. Accordingly, true proportion will not be incorporated as a factor for the random effect of the participant.

Building on our earlier mathematical representation of error, our approach is depicted as:

$$\begin{aligned}
\text{logit}(\mu[i]) &= \beta_\mu[\text{vis}[i], \text{true_p}[i]] + RE_\mu[\text{vis}[i], \text{participant}[i]] && \textit{mean submodel} \\
\log(\phi[i]) &= \beta_\phi[\text{vis}[i], \text{true_p}[i]] + RE_\phi[\text{vis}[i], \text{participant}[i]] && \textit{precision submodel} \\
\text{logit}(\pi[i]) &= \beta_\pi[\text{vis}[i], \text{true_p}[i]] + RE_\pi[\text{vis}[i], \text{participant}[i]] && \textit{zeros-probability submodel}
\end{aligned}$$

$\beta_\mu = \beta_{vis} + \beta_{true\_p}$ *components of average mean absolute error for an observation*

$\beta_\phi = \beta_{vis} + \beta_{true\_p}$ *components of average precision for an observation*

$\beta_\pi = \beta_{vis} + \beta_{true\_p}$ *components of average zero-error probability for an observation*

Accordingly, we model each level of true proportion as having a distinct effect, further explaining variation between estimate judgments that is not explained by changes in visualizations. This will allow us to answer questions like "What characteristics in error are seen at different true proportions?" and "Does true proportion have any effect on estimation error for when varying visualizations?"

We represent this in **brms** as follows:

brm (

brmsformula (

$$abs\_error \sim (vis + 0) + (true\_p + 0) + (vis + 0\,|pt|\,participant),$$
$$phi \sim (vis + 0) + (true\_p + 0) + (vis + 0\,|pt|\,participant),$$
$$pi \sim (vis + 0) + (true\_p + 0) + (vis + 0\,|pt|\,participant)$$

),

family = zero_inflated_beta,

...

)

As before, we use a one-hot encoding of the population-level effect to facilitate setting of priors. The prior for each population-level effect was previously selected to be weakly informative, covering a broad range of effect, and we can reuse them here.

5.11.1 Discussing Limitations In Creating The Model

As a brief aside on the mechanics of the modeling process, I note that the computation of the posterior distribution from the previous modeling process was performed with the aid of significant computational power as part of research done by a larger team. This exploration was be performed with a personal computer. As such, the model was based on

a limited number of participants for ease of computation, as well as using a limited number of chains, iterations, and warm-up samples. These considerations limit the ability of the model to converge on a particular parameter distribution, resulting in the multimodal distributions seen subsequently. The data subset used was comprised of 26 randomly selected participants, representing just under a quarter of the 109 participants considered during the primary modelling process.

To confirm the validity of this approach, we create a model of MAE by visualization type using the initial model specifications in *brms* as given above with the subset of participant data, and compare it to the same model as created by the specifications that incorporate true proportion, in order to demonstrate that the limited model is comparable to the model derived from the full set of data. This is shown in Figure 5.11. The model obtained using the initial model formula substantially resembles the findings earlier in the thesis, with *Bar* outperforming the other three chart types, which are grouped together and broadly overlapping; this suggests that the data subset is representative of the full set of participants.

The long tails in the specifications incorporating true proportion reflect the model's uncertainty about parameter values, likely due to the paucity of data as discussed above. Multimodal behavior is also evident, which is suggestive of the chain's inability to converge on an appropriate range of parameter values. However, because the bulk of the posterior distribution space shows similar characteristics, albeit in a distended way, we will continue with this modelling process expecting that the results will be broadly indicative of what would be obtained when revisiting this model with more resources.

5.11.2 Error CDFs From True Proportions

To validate our *a priori* motivation to attempt this model, we show CDFs for a selection of true proportions was from our experimental data in Figure 5.12, similar to what was shown in Figure 5.3. These CDFs were selected to show the spectrum of deviations from the canonical rankings. In the majority of true proportions, which are not all shown here,

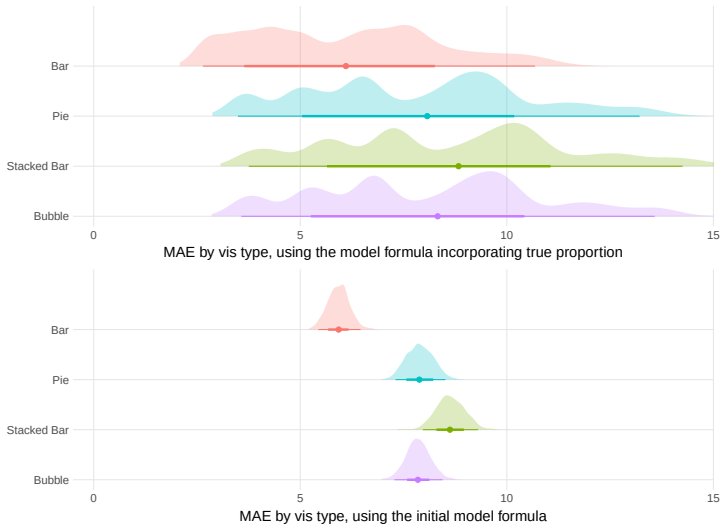


Figure 5.11: MAE by visualization type using both sets of specifications

Bar was the most performant, which is expected based on its aggregate results. However, the true proportions selected for display are not the only ones to show deviations in their behavior between each other and compared to the canonical rankings, simply ones that best allowed for highlighting points of interest.

First, we note that *Bar* did not show noteworthy success compared to the other charts in 3 of the 4 true proportions, and in fact noticeable underperformed in several cases. We also note the significant differences in zero-error estimates across the true proportions; when the true proportion was 1, zero-error estimates occurred in over 50% for every chart type. The substantial amount of zero-error probability at a true proportion of 1, taken in context with the substantial underestimate at 0.55 and 0.75, suggests a strong bias towards particular values even beyond the tendency to round to numbers divisible by 5. The low rate of zero-error outcomes for a true proportion of 0.75 is due primarily to a tendency of participants to estimate a value with an absolute error of 5pp, namely 80pp and 70pp; of those answers, nearly 75% were 80pp, which is in keeping with Talbot *et al.*'s findings of 0.8 being a strongly favored answer [113].

Figure 5.13 shows CDFs obtained by generated fitted draws, as done previously in this paper. The modelled data shows significantly larger 68% and 95% error bands than seen previously in this thesis, primarily due to the uncertainty caused by the limited dataset. The model does not reflect the inversion of performance of *Bar* compared to the

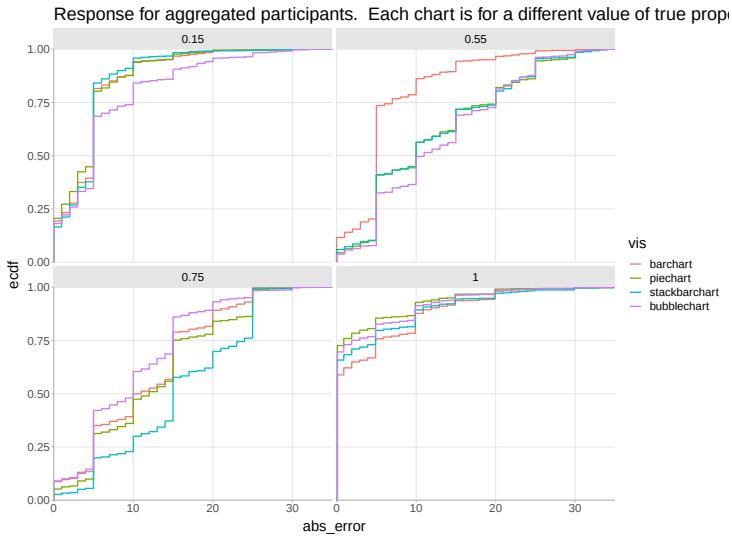


Figure 5.12: CDFs for observed participant data for a selection of true proportions

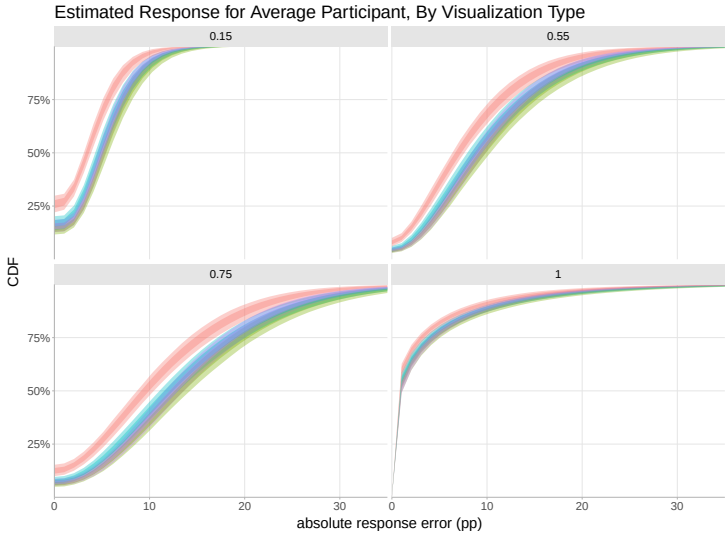


Figure 5.13: CDFs for modelled participant data for a selection of true proportions

other chart types, but otherwise effectively captures the shape of the data. The model reflects the dramatically different characteristics of each true proportion, as well as the significant variation in zero-error probability.

5.11.3 Mean absolute error by true proportion

Figure 5.14 shows the expected mean absolute error for each of the selected true proportions. Significant variation can be seen here: the true proportions of 0.15 and 1 show

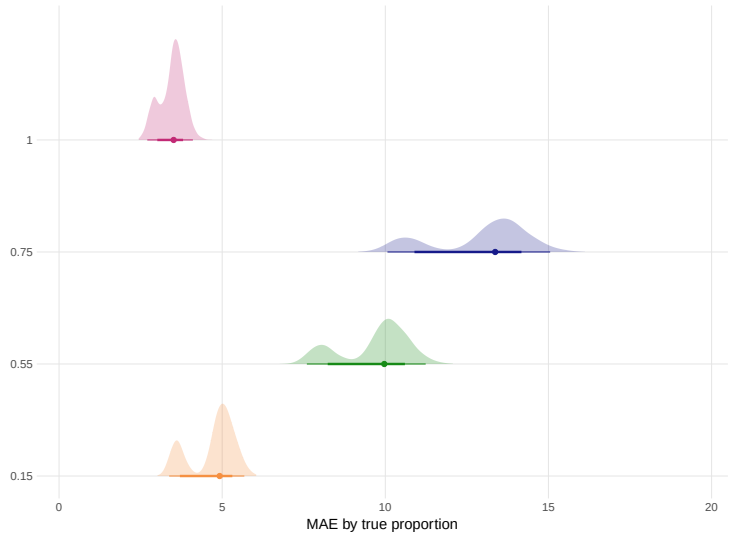


Figure 5.14: Mean absolute error for a selection of true proportions

similar behavior, but seem likely to differ by roughly 8pp from the true proportion of 0.75; this variation exceeds what was seen between participants and between chart types earlier in this thesis. As with the other plots, multimodality is visible and inhibits more exact comparisons between the true proportions, which would likely be mitigated by modelling with more data.

Incorporating the visualization type, as shown in Figure 5.15, shows relatively similar average MAE within each true proportion between the chart types alongside the significant variation across true proportions. As with Figure 5.13, the model differs from the observed data in showing *Bar* as being the best-performing chart.

5.11.4 Discussion of results from incorporating true proportion

Due to the less reliable nature of the modelling process, I discuss possible implications of the results here rather than in the following chapter.

The deviation in MAE between the selected true proportions is significant. Depending on the true proportion, the difference may approach 10pp, which significantly outweighs the variation shown between participants and visualization types earlier in this chapter. However, while the observed data showed significant variation between chart types for a given true proportion, the modelled data did not reflect this. Because of this discrepancy, it is difficult to make strong statements about their implications for design-

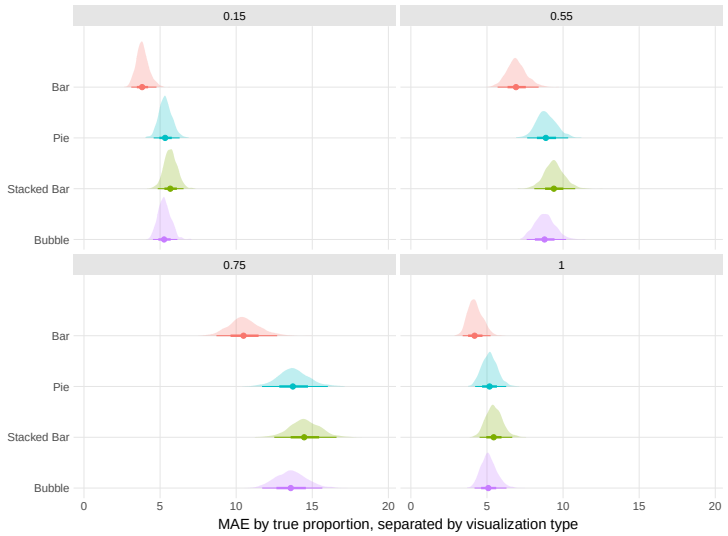


Figure 5.15: Mean absolute error for a selection of true proportions, separated by visualization type

ers, or about possible explanations based on the effect of true proportion on the encoding of data. If a revised model captured the behavior shown in Figure 5.12, with case-by-case outperformance of some charts versus another, the inclusion of true proportion could help provide further weight into the shortcomings of relying on high-level aggregation to drive design recommendations.

The model does, though, reflect the strong variation in zero-error probability across the true proportions. This variation is likely to have implications for designers who are aware of the true proportion(s) they are likely to present. A designer seeking to show that two data points are identical, for example, will be far more successful in doing so than a designer seeking to capture accurately a true proportion of 75%.

This also brings to light the importance of models that incorporate error, rather than absolute error. I noted previously that while 75% showed a common mean absolute error of 5, it would be inaccurate to assume that estimates of 70% and 80% were equally probable. By converting to absolute error, we lose information about individual estimation behavior that could have design relevance, *e.g.* recognizing circumstances where overestimation is more likely than underestimation. This may be difficult to achieve, as the presence of negative error will take the data outside the range where it can easily be modelled with a Beta distribution.

Chapter 6

Discussion

Visualization recommendations built off of foundational studies such as Cleveland & McGill are somewhat sparse, and their advice to prioritize or deprioritize particular visualizations often focuses on particular aspects of psychophysical perception. The results of this research effort, with its investigation of the participants as individuals, may suggest that a different sense of focus may be more practical.

6.1 Importance of Between-Person Variation

Research such as Cleveland & McGill's, which focuses on increasing "the chances of a correct perception of patterns and behavior [by choosing the graphical form that leads to more accurate judgments]" shows a clear focus on optimizing the choice of graphic [29]. Our work suggests that individual variation plays at least as great a role in individual perception as the variation from changing chart types. This research was undertaken in large part to address the hypothesis that population-level results concerning graphical perception rankings are not universally true at the individual level; by quantifying the impact of individual variation and comparing it to the impact of chart variation, we show this to be true. We believe that this highlights the need for continued research into the individual approach for graphical perception. This also provides motivation to research similar aggregate-level conclusions under our research's paradigm. For example, Cleveland & McGill's ordering of elementary tasks from most to least accurate (position

along a common scale; positions along nonaligned scales; length, direction, angle; area; volume, curvature; shading, color saturation) are similarly aggregate-level metrics [29].

Cleveland & McGill intended to leave strategies employed during their experimental task wholly up to their participants, stating that bubble plots composed of circles, for example, "might well [be judged in terms of] diameters or circumferences [or areas]" and wanting to avoid biasing participants toward one particular approach. Since then, study of the strategic approaches individuals might implement for some of these chart types has been undertaken, as with Kosara *et al.* in their study of pie charts [74]. Similar research into such strategies, as undertaken in recent years, assess the use of *perceptual proxies* in lieu of direct examination. For example, if the objects to be compared are two bars separated by a third, a participant may compare each bar to the middle bar to create a judgment. Qualitative research into individuals' approach to graphical perception for more complicated graphics has been undertaken as well; Kale *et al.*, for example, investigated uncertainty visualizations and found that individual strategy played a pronounced role in informing the participants' completion of a particular task [61].

The secondary hypothesis of this research, namely that experiments can be used to rank individuals based on their skill at graphical perception, was achieved by demonstrating that subsets of the population can be seen that display consistently high or consistently low performance. This recognition provides an opportunity for contextualizing findings on user strategy. The consistency in behavior seen by the high- and low-performant participants in Figure 5.7 suggests either that individuals may perceive consistent psychophysical effects across every chart type even when the various chart types are thought to focus individuals on differing psychophysical properties, or perhaps that some individuals have developed successful strategies to be employed across visualizations while others have failed to do so for any. Our findings of limited grouping inside the middling-performant participants in Figure 5.7 is consistent with those found by Maltese *et al.* while studying visualization literacy [89], and suggests that we can adjust our model that might provide explanations for this outcome. Some individual characteristics like spatial ability [101] and personality traits [51] are informative but less useful from a pedadogical perspec-

tive. However, identifying specifically the strategies that are used by highly performant individuals may provide an organic source of effective strategies; likewise, identifying the strategies that are used by low-performant individuals may provide information about ineffective strategies, and provide the opportunity for focused research into the reasons for their varying effectiveness.

We believe that continued use of our modelling approach will provide value for further understanding the impact of intrinsic factors [86], for example by performing a similar task to this study with the added context of cultural heritage to understand its impact on perceptual estimation. This approach also will enhance qualitative understanding of participants' visualization literacy, which has been allowing us to contextualize factors of between-participant variation at a deeper level and with more awareness of their effect on visualization performance [82, 16]. More nuanced information about participants' strategic approach for each individual would allow us to further attribute variation to psychophysical properties or strategic approach. For example, if all strategic approaches are the same, then psychophysical properties are likely to have made the difference; conversely, if each participant focused on the same elementary task, such as assessing area, then their strategy in doing so may carry the most weight.

6.2 Design Recommendations Based On Individual Variation

A CDF-based approach is not feasible on the adjusted log-error approach from prior studies. Since our modelling approach makes it possible, new territory is opened up for providing insight to visualization designers. A designer may be crucially interested in communicating a true proportion of 50%; in this case, we can tell them that a mean absolute error of 35pp (*i.e.* individuals will assess the true proportion to be between 15% and 85%) will capture 98% of responses, and that more than half the population will be within 10pp (*i.e.* a true proportion between 40% to 60%). If the designer feels that a more precise understanding is critical, an alternative to one of the visualizations we have

considered may be necessary; conversely, if the designer is willing to accept that degree of error, they can be confident in moving forward.

Depending on their required degree of accuracy, the CDF plots will provide context as to the impact of selecting different visualizations. A designer who wishes to select a chart type to optimize accuracy to most (say, 90%) of the population) will find that every chart type is likely to be within 5pp of each other. The contribution of between-person variance to this error suggests that, in truth, changing visualizations is likely to provide an increase of accuracy that is between 2 and 3pp and that between-person variance accounts for whatever other differences exist. In circumstances where small amount of precision is essential, bar charts would be recommended; otherwise, the lack of differentiation means that considerations other than accuracy, such as aesthetics or designer preference, should be weighed more heavily than any marginal increase in accuracy that might be gained.

We also note the ease of use associated with directly reporting results as mean absolute error, without needing to wade through conversions from adjusted log metrics. In their paper, Cleveland & McGill stated that "a log scale seemed appropriate to measure relative error". It is not clear what benefit this provides; the response variable does not span multiple orders of magnitude; even with log base 2, they rarely show errors above 3.0, *i.e.* approximately 8. The response variable does not seem to express itself geometrically, in which case a logarithmic approach could reasonably used in an attempt to capture a log-normal distribution, or to exponentiate, in which case a logarithm would provide a linear transformation. The use of midmeans also removes from the best- and worst-performed observations. Because these observations are likely to disproportionately belong to the same individuals, due to the consistency of high- and low-performing individuals found in Figure 5.7, we are especially likely to lose insight regarding these populations. This exacerbates concerns raised by Peck *et al.* [103], who argue that the "data poor" are underrepresented in research findings. Absent any reason to use log error aside from the inertia of tradition, it may be worth pivoting away from that metric.

6.3 Matching Experimental Approach To Designer Expectations

I note that the experimental instructions in Cleveland & McGill, which has persisted across Heer & Bostock's research and our own, were to make "a quick visual judgment". As such, a designer who is creating visualizations meant to be quickly parsed is well-served by this approach, and I conjecture that they are less likely to be deeply concerned with accurate perception. Some designers who do not intend their audience to employ only a quick visual judgment. This does not imply that they expect viewers to sit down with a ruler and protractor to carefully assess their graphics, but they might reasonably assume that participants will spend at least several seconds working to understand what they are seeing. Research studies such as Heer & Bostock's and this one, however, often pay participants based on *expected* time to completion rather than actual time spent. This means that participants are incentivized to proceed as quickly as possible through the experimental tasks, which serves us relatively well for this experimental prompt but less so if deliberation is desired. Research similar to this one but specifically oriented to incentivize some manner of deliberation without excessive time expenditure, *e.g.* a time limit of 15 seconds with some financial incentive to be as accurate as possible, could effectively balance these goals.

An extremely quick assessment at time elapsed as a predictive variable is uninformative; some potentially contributing data, such as the physical size (not just proportional size) of the datasets of interest. It is also realistically possible that considerations such as the strategy employed by the participant will inform both the duration and the error. A focused study would make it more realistic to announce visualization rankings as a matter of fact, without the unspoken disclaimer of "provided the task at hand is a rapid, off-the-cuff judgment".

Chapter 7

Conclusion

By taking advantage of modern computing to address what was once the "substantial chore" of modeling correlation between participant estimates, we have taken a step towards providing nuance to Cleveland & McGill's seminal study. Our work employs Bayesian modeling to yield informative posterior distributions capturing individual variation, allowing us to call into question the conventional rankings around elementary graphical encodings. We conclude that *Bar* is not the optimal encoding in this graphical perception task for at least a fifth of the population. We also conclude that for many choices of encodings, the relevance of between-person variance over between-chart variance suggests that designers should not feel bound by conventional rankings. In circumstances where a few percentage points of error is not crucially important, designers should consider the individual ahead of the chart type, and preferentially assess other design considerations ahead of the encoding. As an additional point of benefit for visualization designers, we propose a shift away from use of log error as a metric, in light of the sheer convenience of working directly with percentage error.

By demonstrating the effectiveness of modeling individual behavior as random effects, we present an alternative to other research that draws conclusions based on the "average" participant and other such aggregation. Ultimately, we believe that after demonstrating possibilities from the posterior distribution obtained from a Bayesian implementation, we may propose that rankings be deprioritized in favor of design approaches that consider individual variation as a primary point of interest. In doing so, we believe

that further research can provide depth to the visualization community's understanding of variation in individual perception, performance, and approach to understanding visualizations, and thereby give more actionable information to visualization designers seeking to effectively communicate.

Bibliography

[1] 538. The 45 best and weirdest charts we made in 2018. https://fivethirtyeight.com/features/the-45-best-and-weirdest-charts-we-made-in-2018/. Accessed: 2019-03-29.

[2] 538. The media has a probability problem. https://fivethirtyeight.com/features/the-media-has-a-probability-problem//. Accessed: 2019-04-20.

[3] 538. Not even scientists can easily explain p-values. https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values//. Accessed: 2019-04-20.

[4] 538. When we say 70 percent, it really means 70 percent. https://fivethirtyeight.com/features/when-we-say-70-percent-it-really-means-70-percent//. Accessed: 2019-04-20.

[5] Alper, B., Riche, N. H., Chevalier, F., Boy, J., and Sezgin, M. Visualization literacy at elementary school. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (2017), pp. 5485–5497.

[6] Ausburn, L. J., and Ausburn, F. B. Visual literacy: Background, theory and practice. *Programmed Learning and Educational Technology 15*, 4 (1978), 291–297.

[7] Baird, J. C., Lewis, C., and Romer, D. Relative frequencies of numerical responses in ratio estimation1. *Perception & Psychophysics 8*, 5 (1970), 358–362.

[8] Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[9] Beecham, R., Dykes, J., Meulemans, W., Slingsby, A., Turkay, C., and Wood, J. Map lineups: Effects of spatial structure on graphical inference. In *IEEE Transactions on Visualization and Computer Graphics* (Los Alamitos, 2017), IEEE Computer Society, pp. 391–400.

[10] Belia, S., Fidler, F., Williams, J., and Cumming, G. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods 10*, 4 (2005), 389.

[11] Bergen, B., and Chan, T. T. Writing direction influences spatial cognition. In *Proceedings of the annual meeting of the cognitive science society* (2005), vol. 27.

[12] Berinato, S. *Good charts: The HBR guide to making smarter, more persuasive data visualizations*. Harvard Business Review Press, 2016.

[13] Bernardo, J. M. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological) 41*, 2 (1979), 113–128.

[14] Börner, K., Bueckle, A., and Ginda, M. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences 116*, 6 (2019), 1857–1864.

[15] Börner, K., Maltese, A., Balliet, R. N., and Heimlich, J. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization 15*, 3 (2016), 198–213.

[16] Boy, J., Rensink, R. A., Bertini, E., and Fekete, J.-D. A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1963–1972.

[17] Briggs, W. M. It is time to stop teaching frequentism to non-statisticians. *arXiv preprint arXiv:1201.2590* (2012).

[18] Brunson, J. C., and Read, Q. D. ggalluvial: Alluvial plots in 'ggplot2', 2023. R package version 0.12.5.

[19] Buerkner, P. Special family functions for brms models. http://paul-buerkner.github.io/brms/reference/brmsfamily.html. Accessed: 2023-04-06.

[20] Bujack, R., Turton, T. L., Samsel, F., Ware, C., Rogers, D. H., and Ahrens, J. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan 2018), 923–933.

[21] Bureau, C. F. P. Consumer financial proection bureau: By the numbers. https://files.consumerfinance.gov/f/documents/201701_cfpb_CFPB-By-the-Numbers-Factsheet.pdf. Accessed: 2022-03-01.

[22] Burgmann, I., Kitchen, P. J., and Williams, R. Does culture matter on the web? *Marketing Intelligence & Planning 24*, 1 (2006), 62–76.

[23] Bürkner, P.-C. Advanced bayesian multilevel modeling with the r package brms. *arXiv preprint arXiv:1705.11123* (2017).

[24] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software 76*, 1 (2017).

[25] Casella, G. Conditional inference from confidence sets. *Lecture Notes-Monograph Series* (1992), 1–12.

[26] CFPB. Cfpb and nyc: How the consumer financial protection bureau empowers and protects new yorkers. https://comptroller.nyc.gov/reports/cfpb-and-nyc-how-the-consumer-financial-protection-bureau-empowers-and-protects- Accessed: 2019-03-30.

[27] CFPB. Cfpb design manual, chart types. https://cfpb.github.io/design-manual/data-visualization/chart-types.html#bar-or-column-charts. Accessed: 2019-03-29.

[28] CFPB. Find out your financial well-being. https://www.consumerfinance.gov/consumer-tools/financial-well-being. Accessed: 2017-03-29.

[29] Cleveland, W., and McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association 79* (1984).

[30] Cohen, J. The earth is round (pį. 05). In *What if there were no significance tests?* Routledge, 2016, pp. 69–82.

[31] Davis, R., Pu, X., Ding, Y., Hall, B. D., Bonilla, K., Feng, M., Kay, M., and Harrison, L. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[32] Dillon, A., and Watson, C. User analysis in hcithe historical lessons from individual differences research. *International journal of human-computer studies 45*, 6 (1996), 619–637.

[33] Dougherty, J., and Ilyankou, I. *Hands-On Data Visualization.* " O'Reilly Media, Inc.", 2021.

[34] Dragicevic, P. Fair statistical communication in hci. *Modern Statistical Methods for HCI* (2016), 291.

[35] Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., and Chevalier, F. Increasing the transparency of research papers with explorable multiverse analyses. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (2019).

[36] Ekstrom, R. B., and Harman, H. H. *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service, 1976.

[37] Fernandes, M., Walls, L., Munson, S., Hullman, J., and Kay, M. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Conference on Human Factors in Computing Systems - CHI '18* (2018).

[38] Few, S. Save the pies for dessert. `https://www.perceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf/`, 2006. Accessed: 2019-03-30.

[39] Few, S. Is the avoidance of 3-d bar graphs a knee-jerk reaction? `https://www.perceptualedge.com/blog/?p=2362/`, 2016. Accessed: 2019-03-30.

[40] Firat, E. E., Joshi, A., and Laramee, R. S. Interactive visualization literacy: The state-of-the-art. *Information Visualization 21*, 3 (2022), 285–310.

[41] for Education Statistics, N. C. Kids' zone: Create a graph. `https://nces.ed.gov/nceskids/createagraph/`. Accessed: 2019-04-07.

[42] Friel, S. N., Curcio, F. R., and Bright, G. W. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education 32*, 2 (2001), 124–158.

[43] Galesic, M., and Garcia-Retamero, R. Graph literacy: A cross-cultural comparison. *Medical decision making 31*, 3 (2011), 444–457.

[44] Gelman, A. The bayesian cringe. `https://statmodeling.stat.columbia.edu/2021/09/15/the-bayesian-cringe/`. Accessed: 2023-03-01.

[45] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[46] Gelman, A., and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[47] Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics 2*, 4 (2008), 1360–1383.

[48] Gelman, A., Simpson, D., and Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy 19*, 10 (2017), 555.

[49] Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. Bayesian workflow. *arXiv preprint arXiv:2011.01808* (2020).

[50] Gigerenzer, G. Mindless statistics. *The Journal of Socio-Economics 33*, 5 (2004), 587–606.

[51] Green, T. M., and Fisher, B. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 203–210.

[52] Harrison, L., Yang, F., Franconeri, S., and Chang, R. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1943–1952.

[53] Hedeker, D., Mermelstein, R. J., and Demirtas, H. An application of a mixed-effects location scale model for analysis of ecological momentary assessment (ema) data. *Biometrics 64*, 2 (2008), 627–634.

[54] Heer, J., and Bostock, M. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 203–212.

[55] Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res. 15*, 1 (2014), 1593–1623.

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

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

[58] Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. *Ecological monographs 54*, 2 (1984), 187–211.

[59] Jardine, N., Ondov, B. D., Elmqvist, N., and Franconeri, S. The perceptual proxies of visual comparison. *IEEE transactions on visualization and computer graphics 26*, 1 (2019), 1012–1021.

[60] Jaynes, E. T. On the rationale of maximum-entropy methods. *Proceedings of the IEEE 70*, 9 (1982), 939–952.

[61] Kale, A., Kay, M., and Hullman, J. Visual reasoning strategies for effect size judgments and decisions. *IEEE transactions on visualization and computer graphics 27*, 2 (2020), 272–282.

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

[63] Kalinowski, P., et al. Identifying misconceptions about confidence intervals. In *Proceedings of the eighth international conference on teaching statistics* (2010), vol. 50.

[64] Kaptein, M., and Robertson, J. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 1105–1114.

[65] Kay, M. *ggdist: Visualizations of Distributions and Uncertainty*, 2023. R package version 3.2.1.

[66] Kay, M., and Heer, J. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 469–478.

[67] Kay, M., Kola, T., Hullman, J., and Munson, S. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, New York, USA, 2016), ACM Press, pp. 4521–4532.

[68] Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI '16, ACM, pp. 5092–5103.

[69] Kay, M., Nelson, G., and Hekler, E. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, New York, USA, 2016), ACM Press, pp. 4521–4532.

[70] Klasnja, P., Hekler, E. B., Korinek, E. V., Harlow, J., and Mishra, S. R. Toward usable evidence: Optimizing knowledge accumulation in hci research on health behavior change. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 3071–3082.

[71] Kong, N., Heer, J., and Agrawala, M. Perceptual guidelines for creating rectangular treemaps. In *IEEE Transactions on Visualization and Computer Graphics* (Piscataway, NJ, USA, 2010), IEEE Educational Activities Department, pp. 990–998.

[72] Kosara, R. An empire built on sand: Reexamining what we think we know about visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (New York, NY, USA, 2016), BELIV '16, ACM, pp. 162–168.

[73] Kosara, R. The eagereyes starter pack. `https://eagereyes.org/starter-pack/`, 2019. Accessed: 2019-03-30.

[74] Kosara, R. Evidence for area as the primary visual cue in pie charts. In *2019 IEEE Visualization Conference (VIS)* (2019), IEEE, pp. 101–105.

[75] Kosara, R., and Haroz, S. Skipping the replication crisis in visualization: Threats to study validity and how to address them: Position paper. In *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)* (2018), IEEE, pp. 102–107.

[76] Kruschke, J. *Doing Bayesian Data Analysis*. Elsevier, 2011.

[77] Kruschke, J. K. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General 142*, 2 (2013), 573.

[78] Kruschke, J. K., and Liddell, T. M. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review 25*, 1 (2018), 178–206.

[79] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics 18*, 9 (2012), 1520–1536.

[80] Lambert, B. *A students guide to Bayesian statistics*. Sage, 2018.

[81] Laufer, D. Small business entrepreneurs: A focus on fraud risk and prevention. *American Journal of Economics and Business Administration 3*, 2 (2011), 401–404.

[82] Lee, S., Kim, S.-H., and Kwon, B. C. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 551–560.

[83] Lewandowski, D., Kurowicka, D., and Joe, H. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis 100*, 9 (2009), 1989–2001.

[84] Lipkus, I. M., and Hollands, J. G. The visual communication of risk. *JNCI monographs 1999*, 25 (1999), 149–163.

[85] Liu, L., Boone, A. P., Ruginski, I. T., Padilla, L., Hegarty, M., Creem-Regehr, S. H., Thompson, W. B., Yuksel, C., and House, D. H. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE transactions on visualization and computer graphics 23*, 9 (2017), 2165–2178.

[86] Liu, Z., Crouser, R. J., and Ottley, A. Survey on individual differences in visualization. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 693–712.

[87] Lynch, S. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, 2007.

[88] Mackinlay, J. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog) 5*, 2 (1986), 110–141.

[89] Maltese, A. V., Harsh, J. A., and Svetina, D. Data visualization literacy: Investigating data interpretation along the noviceexpert continuum. *Journal of College Science Teaching 45*, 1 (2015), 84–90.

[90] Masicampo, E., and Lalande, D. R. A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology 65*, 11 (2012), 2271–2279.

[91] Mazza, R. *Introduction to Information Visualization*, 1st ed. Springer Publishing Company, Incorporated, London, 2009.

[92] McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[93] Miller, J., and Ulrich, R. Interpreting confidence intervals: a comment on hoekstra, morey, rouder, and wagenmakers (2014). *Psychonomic bulletin & review 23*, 1 (2016), 124–130.

[94] Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review 23*, 1 (2016), 103–123.

[95] Morey, R. D., Hoekstra, R., Rouder, J. N., and Wagenmakers, E.-J. Continued misinterpretation of confidence intervals: response to miller and ulrich. *Psychonomic Bulletin & Review 23*, 1 (Feb 2016), 131–140.

[96] Moritz, D., Wang, C., Nelson, G. L., Lin, H., Smith, A. M., Howe, B., and Heer, J. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 438–448.

[97] Moser, C., Phelan, C., Resnick, P., Schoenebeck, S. Y., and Reinecke, K. No such thing as too much chocolate: Evidence against choice overload in e-commerce. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 4358–4369.

[98] Munzner, T., and Maguire, E. *Visualization analysis and design*. AK Peters visualization series. CRC Press, Boca Raton, FL, 2015.

[99] Muth, C., Oravecz, Z., and Gabry, J. User-friendly bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quantitative Methods for Psychology 14*, 2 (2018), 99–119.

[100] Ondov, B. D., Yang, F., Kay, M., Elmqvist, N., and Franconeri, S. Revealing perceptual proxies with adversarial examples. *IEEE transactions on visualization and computer graphics 27*, 2 (2020), 1073–1083.

[101] Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., Han, P. K., and Chang, R. Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 529–538.

[102] Pashler, H., and Wagenmakers, E. Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science 7*, 6 (2012), 528–530. PMID: 26168108.

[103] Peck, E. M., Ayuso, S. E., and El-Etr, O. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.

[104] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[105] Reinecke, K., Minder, P., and Bernstein, A. Mocca-a system that learns and recommends visual preferences based on cultural similarity. In *Proceedings of the 16th international conference on Intelligent user interfaces* (2011), pp. 453–454.

[106] Rigby, R. A., and Stasinopoulos, D. M. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics 54* (2005), 507–554.

[107] Schad, D., betancourt, m., and Vasishth, S. Toward a principled bayesian workflow: A tutorial for cognitive science, Feb 2019.

[108] Schmidt, F. L. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods 1*, 2 (1996), 115.

[109] Shah, P., and Hoeffner, J. Review of graph comprehension research: Implications for instruction. *Educational psychology review 14*, 1 (2002), 47–69.

[110] Shrout, P. E., and Rodgers, J. L. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology 69*, 1 (2018), 487–510. PMID: 29300688.

[111] Smithson, M., and Verkuilen, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods 11*, 1 (2006), 54.

[112] Stone, M., and Bartram, L. Alpha, contrast and the perception of visual metadata. *Color and Imaging Conference 2008*, 1 (Jan 2008), 355–359.

[113] Talbot, J., Setlur, V., and Anand, A. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2152–2160.

[114] Team, S. D. Stan modelling language users guide and reference manual v. 2.31, 2021.

[115] Times, N. Y. 2018: The year in visual stories and graphics. `https://www.nytimes.com/interactive/2018/us/2018-year-in-graphics.html`. Accessed: 2019-03-29.

[116] Tremmel, L. The visual separability of plotting symbols in scatterplots. *Journal of Computational and Graphical Statistics 4* (1995).

[117] Vanpaemel, W., and Lee, M. D. Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review 19*, 6 (2012), 1047–1056.

[118] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis 16*, 2 (2021), 667–718.

[119] Velez, M. C., Silver, D., and Tremaine, M. Understanding visualization through spatial ability differences. In *VIS 05. IEEE Visualization, 2005.* (2005), IEEE, pp. 511–518.

[120] Viegas, F. B., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov 2007), 1121–1128.

[121] Ware, C. *Information Visualization: Perception for Design*, 2nd ed. Morgan Kaufmann Publishers Inc., San Francisco, 2004.

[122] Wattenberg, M. Visualizing the stock market. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 1999), CHI EA '99, ACM, pp. 188–189.

[123] Wigdor, D., Shen, C., Forlines, C., and Balakrishnan, R. Perception of elementary graphical elements in tabletop and multi-surface environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 473–482.

[124] Zhang, H., and Maloney, L. T. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience 6* (2012), 1.

[125] Ziemkiewicz, C., and Kosara, R. Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 911–918.

[126] Ziemkiewicz, C., Ottley, A., Crouser, R. J., Chauncey, K., Su, S. L., and Chang, R. Understanding visualization by understanding individual users. *IEEE computer graphics and applications 32*, 6 (2012), 88–94.