Cognitive Mechanisms for Reasoning with Uncertainty Visualizations

Alex Kale

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jessica Hullman, Chair

Amy Ko, Chair

Jevin West

Jeffrey Heer

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Cognitive Mechanisms for Reasoning with
Uncertainty Visualizations

Alex Kale

Co-Chairs of the Supervisory Committee:
Associate Professor Jessica Hullman
Northwestern University Department of Computer Science

Professor Amy Ko
University of Washington Information School

The joint proliferation of data-driven interfaces in public life and data science in organizations makes reasoning with uncertainty in data visualizations critically important. Lay people and data analysts alike make visual judgments about data almost daily—whether relying on a deluge of Covid-19 visualizations to manage risks to personal and public health, or using exploratory data analysis to drive business decisions. In order to design data visualization software that supports statistically rigorous judgments in these contexts, the visualization community must understand *how people reason with uncertainty visualizations.* My dissertation addresses cognitive mechanisms that chart users rely on when reasoning with uncertainty: (1) automatic perceptual processing, through which the visual system makes intuitive inferences; (2) heuristic strategies, used to interpret visualizations and make consequential decisions; and (3) model-based thinking, whereby analysts compare observed patterns in data with counterfactual predictions from models (either mental or realized in software) that might explain the data. As a capstone to my thesis, I present Exploratory Visual Modeling (`EVM`), a prototype visual data analysis tool that deploys these cognitive mechanisms to support more rigorous exploratory data analysis. The tool enables analysts to express their provisional mental models of data generating process as formal statistical models and to check predictions from these models against observed patterns in data. I

present insights from the design process of `EVM`, as well as considerations for evaluating the design hypothesis that the model checks enabled by `EVM` facilitate improvements in generative thinking during exploratory data analysis. `EVM` deploys automatic and heuristic cognitive mechanisms in service of model-based thinking, providing a proof-of-concept for new ways of designing visualization software.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to express my gratitude to the people who've helped and supported throughout the process of earning my PhD.

I've been lucky to have many great *mentors* on my academic journey. These include the members of my supervisory committee: **Jessica Hullman**, my PhD advisor, took a chance on working with me, knowing that I did not know the tools of computer science yet, but recognizing and cultivating my potential to become a computer scientist. Jessica is an incredible mentor and scientist, and she has become a close friend and confidant to me. I owe her so very much. **Matthew Kay** spent many hours and more than a few late nights working with me on statistics and visualization design. Although I was not his graduate student, he mentored me as if I was on many occasions. Our collaborations did so much to build my skills as a visualization researcher. **Jeffrey Heer** opened his lab to me when I arrived in my PhD program, giving me a home on UW campus where I had peers focused on data visualization. He made me feel like I belonged in this discipline, and our collaborations consistently expanded my horizons. **Amy Ko** stepped up for me in a huge way, when Jessica left the UW for Northwestern at the end of my first year in the PhD program, by volunteering her time to become my supervisory committee chair. Not only did Amy help me navigate the logistics of the graduate program—making it possible for me to finish my degree at the UW—she consistently gave me valuable feedback on my research and guidance on how to teach undergraduates. **Jevin West** mentored my first teaching experience in the iSchool and has always made time to discuss my research interests. **Katharina Reinecke** consistently gave me opportunities to present my research and receive valuable feedback, whether through the milestones of my degree program or informally in her group's lab meetings.

Before I started the PhD program, I had a series of mentors as an undergraduate at the

UW who helped me get started in research and in adult life. **Steve Buck** mentored my first research project as an undergraduate, inviting me to participate in his lab despite my lack of experience and welcoming my "fresh perspective". In doing so, he changed my life by setting me up to pursue a career in research. **Laura Little** cultivated my interest in statistics, teaching the courses that got me interested in statistics and offering me opportunities to grow this interest through peer-teaching and an advanced seminar on statistical reform. She played a critical role in helping me choose a research topic for graduate school and is someone I could always turn to for advice. **Lynn Hankinson-Nelson** cultivated my interest in philosophy of science, inviting me to take upper-division courses which set me up to think deeply about issues of epistemology and uncertainty. Her introductory course helped me find my intellectual identity; her upper division courses cemented my interests with depth and encouragement. **Fred Radke** took me under his wing and mentored me in trumpet playing, which was a major interest of mine in undergraduate. His generosity gave me access to the UW School of Music, improved my musicianship and emotional self-regulation, and kept me invested in my college education early in my college experience when I had doubts about my future. **Scott Murray** hired me as a post-baccalaureate research assistant on a set of visual neuroscience projects that involved a massive data collection effort. In this role, I honed my skills in experimental design, human subject research, programming, and statistical analysis. **Michael-Paul Schalmo** was a postdoctoral researcher in Scott's lab at the time, and Michael-Paul gave me tremendous one-on-one mentorship that set me up for graduate school in a way that few other experiences could have. To this day, I go to him for advice in my career and life and consider him a dear friend.

My wonderful *peers* enriched my PhD journey through collaboration, friendship, and mutual intellectual growth. There are too many individual people to name, so I'll talk about groups instead. One particular strength of the iSchool is the solidarity and support among the graduate students. For me this support came especially from the **i17 cohort**, who provided social and moral support to each other throughout the PhD experience and especially in the first two years. I was also fortunate to find community in UW CSE, where

I participated in and later led the **HCI seminar**, an amazing opportunity to learn from and share ideas with my peers. Jeff Heer and Leilani Battle's group, the **Interactive Data Lab**, gave me a community around visualization research at UW. At every turn they offered valuable feedback, intellectual discussions, and friendship. Jessica Hullman and Matthew Kay's group, the **Midwest Uncertainty Collective** gave incredible feedback and provided a scholarly community for me, especially in the later years of my PhD. Working with them, I was able to grow into a leader, and they frequently challenged me to broaden my perspective.

At times when pursuing a PhD was difficult, I could always turn to my *friends*. Again, there are too many people to mention, but I'll name a few. **Luke Kreikemeier and Jade Highleyman** have been my closest and most steadfast friends for many years. The closeness of our bond kept me grounded through the PhD, even from hundreds or thousands of miles away. **Graham Fluster, Dylan Fluster, Megan Taylor**, and my broader group of Seattle friends from undergraduate have been a huge part of of my social circle, especially in the early years of graduate school. From dance parties in their living room to pandemic game nights on Discord, they kept me in good spirits and lifted me up. **Nick Gendreoux, Jimmy Austin**, and the broader group of musicians I played with in Seattle are wonderful people, whose friendship and creativity has so enriched my life. They helped me to maintain an identity outside of my work, which I think was an incredible gift. **Xavi DeLeon and Abbie Waite** are old friends that I reconnected with during the pandemic, through book club and game nights. They picked me up when I was down and have always encouraged me. To all my friends, I am forever grateful to you for our time together.

Thank you to my *partner*, **Kelsey McCormick**, who shares in life's joys and struggles with me. In our recent times of crisis and trepidation, she helped me build a life in Chicago, to manifest my authentic self, and to more deeply appreciate the virtue of dedication.

To *my parents*, **Steven Kale and Pat Mainella**, thank you for being a consistent source of love, guidance, and support. None of this would be possible without you.

DEDICATION

For my father and step-mom, Steve and Pat,

and for my late mother Marla.

Chapter 1

INTRODUCTION

Reasoning with data often involves grappling with uncertainty, writ broadly. This can mean making judgments and decisions in the face of incomplete or unreliable information. It can mean accounting for the natural variability in stochastic processes (e.g., a game of chance). It can mean attributing changes or variance in observed events to different causes or sources, trying to fill in the gaps in one's notion of "what is real".

In the modern world, most people's experiences with data are mediated through visualizations. For example, members of the public engage in civic discourse around journalism and media, which increasingly features data visualizations and statistical concepts. Government entities, non-profits, and private companies rely on teams of analysts to make uncertain data actionable by using statistical tools to drive informed decisions and/or predictions. In all these contexts, people will almost inevitably encounter data visualizations that seem to show a pattern and wonder to themselves, "Is this pattern real, just a coincidence, or a special case that might be cherry picked to argue a point?" How we answer these questions can influence widespread beliefs among the public, resource allocation, and decision making at the individual and organizational level. The consequences of misunderstandings about data can be dire, as we've seen repeatedly during the covid-19 pandemic.

We often struggle to reason about uncertainty in data, in no small part because many of our statistical tools fail to account for people's natural tendencies when reasoning with uncertainty. Data visualizations conventionally omit uncertainty information, especially public facing visualizations [94]. This encourages well-known tendencies to ingore or downplay uncertainty [106, 199]. When visualization authors do show uncertainty, it tends to be communicated either: implicitly, e.g., through a set of questionably-representative exemplars chosen by a journalist (Fig. 1.1); or using presentation formats like confidence intervals (e.g., Fig. 1.2), which laypeople and scientists alike struggle to interpret [16, 39, 86, 98, 156, 182].

Compounding these problems, the uncertainty information we do present tends to be *contingent* [74] on choices and assumptions made during data collection, analysis, and statistical modeling. Data visualizations and the interfaces around them seldom expound these sources of uncertainty, effectively veiling information from the user that would help to resolve potential misinterpretations.

My thesis presents and argues for an approach to visualization research that can address these problems around reasoning with uncertainty by accounting for *cognitive mechanisms*, the processes of perceiving and thinking that underly human behavior. Drawing on empirical findings, models, and theories from psychology, economics, statistics, human-computer interaction (HCI), and information visualization, I develop mechanistic accounts of how people interpret uncertainty visualizations. I demonstrate through experiments and tool-building how understanding the cognitive mechanisms behind specific types of judgments with uncertainty can support nuanced visualization design guidelines, improved models of visualization "effectiveness" [140], and theoretically-

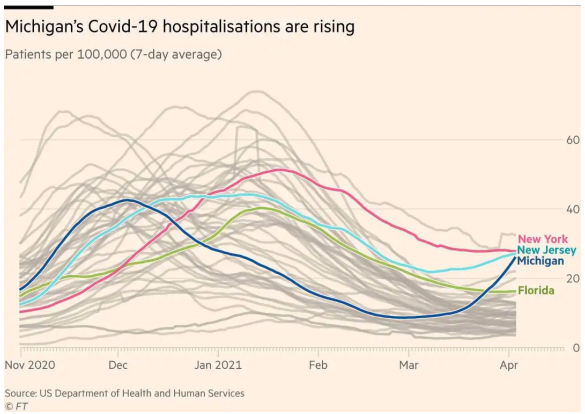


Figure 1.1: Covid-19 visualization from the Financial Times [30]. The chart shows hospitalization rates over time contrasting the trends in Michigan and a handful of other states (in color) with a set of unidentified reference states (in gray). Are the selected groups of states similar enough to be comparable in a way that is statistically meaningful?

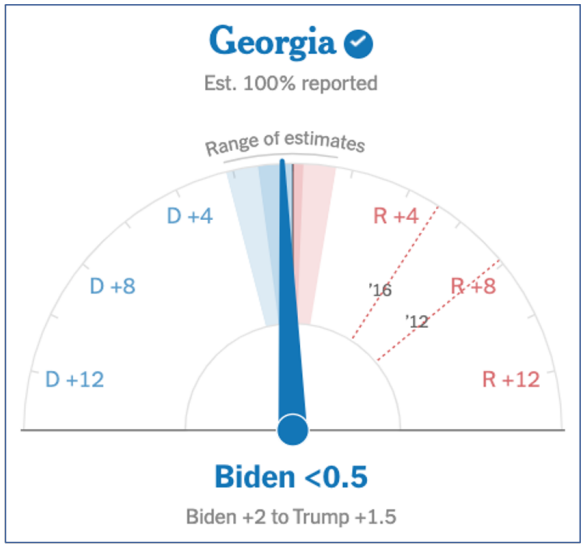


Figure 1.2: Election needle from the New York Times [6] showing the estimated vote margin for Georgia in the 2020 US presidential election. The needle shows the point estimate for vote margin, and the colored wedge behind the needle shows a confidence interval around the estimate.

motivated design patterns for data analysis tools. *I assert that the three distinct cognitive mechanisms examined in this thesis help to explain how people use uncertainty visualizations, in ways that can be useful for creating formalisms to drive interactions with data.*

1.1 Cognitive mechanisms

In my dissertation, I present in-depth research on three mechanisms for visual reasoning with uncertainty visualizations. Although these mechanisms are distinct, chart users can deploy them in concert to perform complex tasks like sensemaking [31,78,164,170]. I aim to study each of these processes separately and to design and build software that helps chart users deploy them rigorously during data analysis.

The set of mechanisms I examine here are not exhaustive. By investigating these cognitive processes in the context of visualization interpretation, I demonstrate a particular way of thinking about and evaluating how people use visualizations. This approach opens up new directions for visualization research, so I think of the mechanisms presented here as a starting point rather than a complete accounting of visual cognition.

1.1.1 Ensemble processing.

Ensemble processing is an unconscious behavior and a fundamental mechanism for sensory learning, by which the brain automatically extracts summary statistical information from the environment [3]. Ensemble processing is most often implicated in inferring the mean of a visual property of objects in a scene, for example, the average size of a sequence of dots [92] or the average lifelikeness of a set of objects [129]. However, evidence from behavioral experiments and neuroscience, respectively, suggest that humans can similarly learn the variance of a stream of events [18] and that variance information about the environment is stored by the brain in order to calibrate sensory sensitivity [55].

These visual summary statistics seem to form the basis for expectations conditional on our environment. For example, that we automatically internalize the color statistics of different natural environments accounts for otherwise hard-to-explain behaviors such as color constancy [24, 212], a phenomenon where colors have relatively stable appearances across environments despite dramatically different sources of illumination. In other words, *the*

visual system is constantly calibrating itself to form the most accurate possible internal representations of the external environment, while compensating for environmental constraints. Some cognitive scientists analogize this sort of automatic association learning and calibration to a Bayesian update over possible states of the world [48], and the results of behavioral experiments seem consistent with this interpretation [18].

Elsewhere, I reviewed the literature on learning empirical priors through sensory adaptation and commented on the challenges and opportunities that this behavior poses for visualization software [107]. I am not the first to propose that such adaptation may play a role in visual analytics tasks. For example, early research with medical professionals shows that visual calibration can improve judgments of radiological images [124, 125]. Prior work on graphical statistical inference proposes a "visualization Rorschach" procedure [28], where analysts look at possible realizations of a data generating process in order to internalize what predictions from that process might look like. The work presented in Chapter 3 [110] tests these ideas empirically, demonstrating that experience-based animated visualizations (a.k.a. hypothetical outcome plots, HOPS [98]) help chart users internalize distributions better than more conventional ways of visualizing distributions. In Chapter 6, I present a software interface for exploratory data analysis which leverages HOPs for visual calibration in order to help analysts reason about the fit between data and possible explanatory models.

1.1.2 Heuristic perceptual proxies.

Heuristics are mental shortcuts where one substitutes a simple judgment as a proxy for a more complex and intractable one [199]. Although heuristics can lead to systematic errors in judgments—a myriad of cognitive biases which researchers have described and named (e.g., [50])—these biases only manifest under specific conditions. Among some but not all HCI researchers there is a misconception that heuristics always lead to biases, that these things go hand-in-hand. However, it is more appropriate to think of heuristics as useful strategies worth relying on and to see resulting biases as edge cases to avoid where possible.

In his 2011 book "Thinking Fast and Slow" [105], Daniel Kahneman conceptually divides cognition into two disjoint systems, System 1 for fast heuristic judgments and System

2 for slow deliberate reasoning. Although this *dual process model* of cognition is an interesting framework with important implications for visualization (e.g., [154]), it oversimplifies cognition in ways that invite misinterpretation. The dual process model encourages us to overlook the ways in which heuristic judgments can be used deliberately and strategically. For example, a data analyst might need to reason carefully through a decision making problem using information they extract almost effortlessly from from a visualization. I argue this analyst is thinking both fast and slow within the context of a single task. Whatever benefits the visualization format confers to this analyst probably have to do with enabling quick heuristic judgments in service of thoughtful and goal-directed behavior, whereas many interpretations of the dual process framework by HCI researchers cast heuristics as the antithesis of careful reasoning. Given the limited capacity of human attention, visualization effectiveness depends in part on whether the visual representation supports reasoning by visual proxy without requiring complex mental transformations [79, 100, 225]. Though this is often overshadowed by the spectre of "heuristics and biases", our biological and cognitive limitations necessitate heuristic reasoning even when we are being deliberate.

However, using perceptual proxies to make heuristic judgments can cause trouble when we chose the wrong visual reasoning strategy for a task. For example, imagine a scenario where a chart user must use a chart to solve a multi-attribute optimization task with real-world consequences. We could think of them as balancing a set of incentives in an internal utility function [106,206]. If all that differentiates between the optimal strategy and suboptimal strategies in most circumstances are small differences in payoffs, people may "satisfice" by choosing suboptimal strategies that work well enough most of the time [178]. Prior work adopts the design philosophy that visualization researchers should catalogue descriptions of specific biases that result from frequently adopted heuristics and design interventions to steer chart users toward better strategies (e.g., mitigating the attraction effect [49]). However, this whack-a-mole approach to heuristics will not scale beyond the limited set of tasks that researchers consider. In Chapter 4, I present an alternative design philosophy that centers the relative utility of visual reasoning strategies, rather than cognitive biases, as a consideration in visualization effectiveness [109]. I empirically test the hypothesis that chart users ignore uncertainty, I demonstrate that this seems to be a widespread form of

satisficing, and I address challenges for creating models of visualization effectiveness that make visual reasoning strategies into an explicit consideration.

1.1.3 Model-based reasoning.

The third type of reasoning mechanism I address is fundamentally model-based, in the sense that it involves sensemaking about information in reference to a mental model of the causal processes in the environment [126]. Whereas ensemble processing and heuristics are canonically [105] chartacterized as model-free or System 1 mechanisms, model-based reasoning is the cognitively demanding, System 2 foil to these other mechanisms. However, in line with my argument above, evidence from neuroscience suggests these model-free and model-based mechanisms are tightly integrated in tasks such as decision making [47].

Multiple lines of research posit model-based reasoning as a core part of data analysis, including accounts of sensemaking [31, 78, 164, 170], statistical inference [66, 67, 96], Bayesian cognition [76, 77], and empirical work on visual analytics interfaces for causal inference [12, 207, 208, 222, 224]. These various accounts agree that chart users often reference mental models of possible explanations for their data during analysis tasks. Hullman and Gelman [96] offer perhaps the best articulation of this phenomenon to date, arguing that these mental models are seldom expressed in graphical user interfaces for exploratory data analysis and thus remain implicit in the mind of the analyst. The same argument casts statistical modeling during programmatic data analysis as a more explicit version of the same activity, iteratively checking provisional models (e.g., as described by Gabry et al. [61]). This view of the role of models in data analysis hearkens back to Tukey's original notion of exploratory data analysis [197] (EDA) and blurs the line between exploratory and confirmatory modes of analysis in ways that, as Hullman and Gelman argue [96], open opportunities for new design patterns in data analysis tools.

Consider how chart users reason about causal inferences as an example of how model-based reasoning affords differentiation between alternative hypothetical *data generating processes* (DGPs). Chart users compare observed patterns in data to counterfactual patterns which are the logical consequences of various hypothesized causal scenarios, and they at-

tempt to rule out hypotheses that are incompatible with their data [12, 75, 224]. In Chapter 5, I empirically investigate how well chart users can make this sort of counterfactual comparison using typical visual analytics (VA) tools, where counterfactual causal patterns are implicit in the mind of the user and not externalized in charts [111]. Borrowing models of causal induction from the Bayesian cognition literature [77], I show that common VA tools fail to adequately support high-quality causal inferences in part because they rely on the analyst to intuitively know what conclusions the patterns in data support. Echoing Hullman and Gelman [96], I argue VA tools should provide external representations of such *model checks*—i.e., comparisons between data and predictions from provisional explanatory models. In Chapter 6, I demonstrate how typical VA tools can be extended to make model-based reasoning more explicit by enabling analysts to express provisional models and check the compatibility of their predictions against patterns in data.

1.2 Dissertation overview

I present three finished research projects and one work-in-progress on cognitive mechanisms for reasoning with uncertainty visualizations. The three finished projects entail carefully designed experiments and extensive statistical modeling of data cognition. Along with my co-authors, I published these works at IEEE VIS, the premier venue for visualization research. The work-in-progress entails an implemented prototype VA application, which puts design implications from my empirical work and theories of statistics into practice, and a detailed evaluation plan to be conducted in the near future.

In Chapter 2, I present related work on uncertainty visualization, graphical statistical inference, human judgment and decision making, visual causal inference, and cognitive modeling. These focused reviews serve to situate my dissertation within a broader body of work and to provide the reader with necessary background to understand what follows.

Chapter 3 focuses on the affordances of various presentation formats for conveying information about uncertainty due to sampling error. Inspired by the prevalence of statistical concepts in journalism, I present an analysis of common judgments that seem to be implicated in a small corpus of online articles. I use this content analysis to motivate a pair of experiments based on a piece from the New York Times (NYT) about "How not to be

mislead by the jobs report" [99]. In these experiments, I rally methods and models from perceptual psychology to investigate the potential benefits of HOPs [98], an uncertainty visualization technique which I argue has unique affordances for supporting statistical reasoning via ensemble processing. I show that HOPs lead to improvements in untrained laypeople's ability to judge the underlying trend in noisy timeseries data compared to more conventional uncertainty representations.

Chapter 4 posits that visual heuristics play a major role in the way laypeople use visualizations to make informed decisions. Expanding on experimental designs and formalisms from behavioral economics and decision theory [10, 206], I present an experiment testing how visualization design choices influence people's reliance on different heuristics and as a consequence their decision making behaviors. I show that people tend to rely on suboptimal heuristics for decision making with uncertainty visualizations, even when we design the charts we show them with better heuristics in mind. The findings of this experiment imply not only that chart users might often need more explicit direction on how to use visualizations, but they also point to shortcomings in prevailing notions of what makes a visualization "effective" [140], which underly visualization recommender systems deployed in popular analysis software (e.g., in Tableau).

In Chapter 5, I investigate the cognitive difficulties that people encounter around reasoning about causation with visualizations. Borrowing from prior work in mathematical psychology [77], I develop a pair of experiments and a method for evaluating the quality of chart users' causal inferences without knowing the true data generating process (DGP). I show that people struggle to make causal inferences with visualizations, not performing reliably better with visualization and interaction techniques common in VA than with simple text contingency tables. I explore possible explanations for these difficulties with model-based reasoning, including insensitivity to sample size and to evidence that seems to verify (rather than falsify) a causal claim. The findings of this experiment suggest new research directions for causal inference with visualizations and novel design patterns for VA tools.

Chapter 6 culminates my thesis by applying insights from previous chapters to the design and construction of Exploratory Visual Modeling (`EVM`), a prototype VA tool that enables analysts to express and check statistical models against their data during exploratory

data analysis (EDA). `EVM` elicits the user's provisional assumptions about DGP as formal statistical models and leverages the cognitive mechanisms described in previous chapters to help the user gauge the compatibility of these assumptions with observed patterns in data. `EVM` provides a proof-of-concept for encouraging users to check statistical models during data exploration, following recommendations from a recent synthesis of statistical theories [96] and defying orthodoxies which insist on a separation between exploratory and confirmatory analysis, demarcated by the application of formal statistical models. This work demonstrates how the ideas within my dissertation can be applied to the building of analysis tools, creating an opportunity for greater impact.

Chapter 7 summarizes the findings and implications of each previous chapter. I present plans for future research directions that follow the trajectory set in this dissertation.

Chapter 2

RELATED WORK

To contextualize my research, I offer an overview of related literature with special attention to the topics and scholarly debates that are core to my dissertation. This chapter contains preliminaries on visualizing uncertainty, tasks with uncertainty visualization, and behavioral models of people's data cognition.

2.1 Visualizing Uncertainty

Construed broadly, uncertainty can mean many different things including but not limited to statistical variability, incomplete information, or reasons to doubt the veracity of evidence from a dataset and analysis. Statistical variability is often the focus of uncertainty visualization because it can be quantified and parsed by analysts using statistical models to help make sense of and separate out patterns in data. However, other forms of uncertainty such as ambiguity, doubt, unknowns, and missing information are much more difficult to account for using mathematical formalisms. This means that most uncertainty visualization research, including some of my own work, is limited in scope to consider mostly quantifiable uncertainty that can be visualized as distributions.

Though it is beyond the scope of this review to present an exhaustive design space of uncertainty visualizations, I highlight a few *distinctions* that might help to characterize such a design space.

Frequency framing vs direct encoding of uncertainty. The distinction I explore most thoroughly in my work is between uncertainty visualizations that show uncertainty as a visual property of graphical markings (e.g., spatial extent or color) versus uncertainty visualizations that show uncertainty as possible outcomes represented by discrete markings. For example, this distinction describes the core difference between continuous probability densities and quantile dotplots [115], a technique invented by Kay and Hullman which quan-

tizes a distribution and represents a set of percentiles as discrete dots. Framing probability in terms of frequencies of discrete events (a.k.a. *frequency framing*) can lead to more accurate statistical reasoning in many contexts—e.g., learning about sampling error [34], medical risk assessment [63, 89], decision making about bus arrival times [56, 115], inferences in the framework of Bayesian updating [69, 87, 119, 145], and graphically elicited predictions from laypeople about stochastic processes [70,97]. The benefits of frequency framing may have to do with the fact that the visual system automatically extracts frequency information from our environment [3, 85], even when people are not consciously aware of doing so [81].

Animated vs static uncertainty. Another distinction is between static visualizations that show distributions in aggregate versus dynamic visualizations that show possible realizations from a distribution over time, dubbed hypothetical outcome plots or HOPs by Hullman [98].[1] Static visualizations (e.g., intervals [44,192], boxplots [196], histograms [161], probability densities [11,183]) are currently the predominant way of visualizing uncertainty. Static visualizations align well with the design goal of cognitive efficiency in communicating uncertainty [32,95,128,195], whereas animations are thought to be less efficient because they require temporal integration and pose challenges related to memory and attention [201].

However, static uncertainty visualizations force chart users to confront numerous cognitive difficulties. Designers of static uncertainty visualizations tend to assume that the audience has some baseline statistical knowledge [16] and a propensity to connect abstract concepts such as sampling distributions to graphical representations [43]. Error bars in particular can be difficult to interpret due to inconsistency of conventions governing what constructs they represent [39, 86] (e.g., X% confidence intervals, standard errors, X% predictive intervals), although interval representations of uncertainty are often recommended by experts [44, 141, 192]. For charts with enclosed shapes, within-the-bar bias can impact estimates of likelihood such that values within the shape tend to be judged as more likely than values outside [39,148]. Uncertainty visualizations which use separate marks to encode expected value and uncertainty can elicit a heuristic bias of attention toward expected value

[1] Although Hullman coined and popularized the term HOPs, earlier work applied a similar approach to show uncertainty in medical volume rendering [139] and estimates of geospatial surfaces [52, 54, 72]. Hullman and colleagues' [98] distinct contribution was to recognize and empirically evaluate the benefits of HOPs, formalizing it as a visualization technique.

and away from uncertainty [98], in keeping with a well-documented tendency for people to underweight or ignore uncertainty [198].

Prior to the research presented in Chapter 3 [110], little previous empirical work had evaluated the potential for animation to provide an effective visual metaphor for probabilistic information. The exceptions are a few early studies of animated uncertainty visualizations in cartography [52, 54, 72], a study on HOPs of hurricane location forecasts [132], and previous work by Hullman [98] establishing that HOPs enable more accurate inferences about joint probabilities than static uncertainty visualizations and that HOPs support comparable performance to static representations when users estimate properties of univariate distributions. In recent years, news organizations used HOPs and similar interactive simulations to communicate uncertainty in the job market [99] and in political elections [1, 23, 167]. This suggests potential for HOPs to communicate uncertainty information to the public in informal contexts, a possibility that we investigate in Chapter 3.

Expressiveness of underlying distribution. Both of these distinctions about frequency framing of probability and animated uncertainty define different kinds of *"information format"* [69], which may be more or less in line with the information processing abilities of the human visual system or the structure of a reasoning problem. However, information format is not the only aspect of an uncertainty visualization that impacts its utility for a given task. A further important consideration that I wish to distinguish here is the *expressiveness* of uncertainty visualizations (in the sense of MacKinlay [140]) or the granularity with which they show a distribution. For example, interval versus probability density representations of a distribution both encode a range of possible outcomes as spatial extent, but probability densities are more expressive of the underlying distribution.

The design characteristics I highlight here distinguish between the kinds of uncertainty representations that feature prominently in my dissertation.

2.2 Tasks with uncertainty visualizations

Typically, uncertainty visualizations support tasks that involve reasoning about a space of possibilities, whether these are possible outcomes that a chart user needs to consider or different ways that a data scientist might analyze a given data set. Common tasks with

uncertainty visualizations include:

- **Inferences:** What is a reasonable estimate of an imperfectly measured value? Is there a real difference between these two groups of observations? Is it likely that a given data generating model explains the relationships in a data set?

- **Decisions:** Should I invest in a particular intervention? Should I choose one product or service over another? What strategy should I adopt to maximize monetary gain or minimize monetary loss? What is the most accurate way to analyze this data?

- **Sensemaking:** Is my interpretation of data accurate? Do I see the counterfactual patterns in my data that I expected to see under a hypothetical data generating process (DGP)? What could I be missing about this data set that would help to explain an unanticipated pattern?

Notice that these tasks are highly overlapping. For example, inference is often a sub-task in decision making or sensemaking contexts. Also note that while uncertainty visualization and statistics can be valuable tools for such tasks, merely seeing distributions of data or model outputs is not enough to solve these kinds of problems. Technological representations of these problems tend to underdetermine their solutions, so *complex human judgments are inexorable from many data science applications that involve reasoning with uncertainty.*

2.2.1 Graphical statistical inference

Graphical statistical inference, broadly construed, refers to visual methods for judging whether a particular statistical model fails to describe the observed pattern in data. For example, from the early days of exploratory analysis, Tukey called for analysts to examine model residuals in order to inspect where a model might be wrong [197]. Similarly, common model diagnostic tools such as QQ-plots [137] create visual tests that a model's assumptions are satisfied, which I refer to as *"model checks"* following Hullman and Gelman [96].

Although there are conceptualizations of visual inference that do not fit into the model check framing, the inference tasks I examine in my dissertation do. The task we ask participants to perform in the experiments presented in Chapter 3 [110] is a form of model

check—i.e., chart users judge whether samples of jobs added to the economy each month of a year are more compatible with models that assume a growth trend or no growth in the job market. Similarly, the causal inferences about data generating process (GDP) that we ask participants to make in the experiments presented in Chapter 5 [111] are essentially model checks, where we've asked chart users to imagine counterfactual predictions from alternative DGPs and gauge their compatibility with samples of data from a fake clinical trial.

The best-known formulation of model checks is the visualization lineup protocol [28,216], in which many plots of fake data are generated from a "null model" (i.e., representing a null hypothesis) and an impartial observer is asked to pick out a plot of the real data among the set of fake plots. Although this lineup procedure has been analogized to a formal statistical test, this comparison is fraught because it is difficult to create a lineup of fake plots that share all visual properties with the real data except for those which would violate the model's assumptions [202,203]. Hullman and Gelman [96] argue that, thankfully, we do not need to create a graphical inference procedure that is a perfect analog to a statistical test in order to benefit from model checks. They give motivating examples of how imperfect model checks can be useful, demonstrating that without evoking the analogy to a statistical test or attempting to achieve guarantees about error rates, model checks still show analysts precisely *for which cases a model is wrong about the data.* This might serve to pique the analysts curiosity and spur new discoveries about factors that may explain patterns in data. However, Hullman and Gelman's theoretical account leaves many unanswered questions about how imperfect model checks ought to be realized in interactive systems.

Statistical best-practice tells us to construct and check models iteratively in a process called *model expansion* [61], making incremental changes to a model that represents our provisional understanding of the data and inspecting model checks and diagnostics to see if these incremental changes improve our accounting of the data. I see deep parallels between this model expansion workflow and both Tukey's conception of exploratory data analysis and graphical inference methods such as lineups. All three approaches to analysis pose "what if" questions for the analyst such as, "What if we add or remove this variable to our model? What if we relax this assumption?" Whether we rely on the lineup or any other model check such as Tukey's plots of residuals, the power of model checks is that they

prompt the analyst for their creative and interpretive input as to what will be the next step in an analysis, whether that is an exploratory query, the relaxing of an assumption, or the acceptance of a provisional model. The work presented in Chapter 6 synthesizes these ideas about graphical statistical inference into a software prototypes for incorporating model checks into exploratory data analysis.

2.2.2 Incentivized decisions

Whether considering problems of resource allocation or risky choice, decision making is fundamentally about selecting an optimal course of action. In order to choose an optimal course of action, people need to consider uncertainty information. *Decision theory* [206] tells us that an optimal course of action $a_o \in A$ is one that maximizes expected utility $\mathrm{E}(u_a(e))$ across possible events $e \in E_s$ and possible states of the world $s \in S$,

$$\mathrm{E}(u_a(e)) = \Sigma_{s \in S} \Sigma_{e \in E_s} \Pr_a(e) \cdot u(e)$$

$$a_o = \mathrm{argmax}_{a \in A} \mathrm{E}(u_a(e))$$

where $u()$ is the utility or value of a given event e in the space of possible events E_s, given a state of the world s among possible states of the world S, and $\Pr_a(e)$ is the probability density function of events given a particular course of action a among possible courses of action A. In this formalism, uncertainty manifests in two ways: (1) as a probability distribution $\Pr_a(e)$ of possible events E_s; and (2) as a conceptual space of possible states of the world S. Decision theory helps us appreciate how ignoring possible events or states of the world might result in mis-calibrated judgments of expected utility $\mathrm{E}(u_a(e))$ and thus suboptimal choices $a \neq a_o$.

Behavioral economists investigate a person's ability to make optimal decisions using gambling problems [10]. A typical setup for such experiments is to ask participants to choose between two different gambles with different expected payoffs, manipulating the probabilities of a payoff, the amount of the payoff, or both. Often, one of the two gambles is a coin-flip with a 50% chance of winning, which serves as a common baseline to enable comparisons across decision scenarios. The experiment I present in Chapter 4 [109] follows

a similar paradigm to investigate decisions from visualized uncertainty. Results from such experiments describe the conditions under which people make more or less optimal choices.

When people make suboptimal choices, economists historically call this behavior "irrational", but many contemporary debates in cognitive science concern how we should actually interpret these behaviors. Nobel prize winning work on *prospect theory* [106] posits that both probability and utility are subjective judgments—at least as these constructs exist in the mind of a decision maker—and conceptualizes errors in these judgments as the reason for suboptimal decisions. This gives rise to traditional explanations of bias in decision making as risk-seeking or -aversion [209]. However, recent efforts in cognitive modeling (e.g., [118,194]) problematize explanations in terms of risk attitudes, suggesting that apparent economic irrationality is an artifact of lossy signal processing in the brain, which varies in its imprecision from person to person. Although it is far beyond the scope of this review to resolve this debate, these issues play a key role in Chapter 4 [109].

One thing that does seem clear is that departures from normative decision making sometimes result from reliance on *heuristics* [106, 178, 198], simple mental shortcuts in lieu of probabilistic thinking. For example, consider a decision scenario like the one in Chapter 4 [109], where a chart user must compare visualized event distributions under alternative courses of action $\Pr_a(e)$ in order to make a decision. This chart user needs to judge *effect size* [37], or the reliability of the difference between two distributions, in order to make an optimal choice. However, when people don't know how to interpret uncertainty, prior work [98] suggests that they may substitute a judgment of the mean difference between distributions for more complicated judgments of effect size, a visual reasoning strategy that we investigate in Chapter 4. This *visual distance heuristic* motivates visualization design principles—e.g, that the quantitative axis on a bar chart should always start at zero [26, 93], or that axis scales should make visual distance commensurate with effect size [219]. Experiments show that axis scale impacts the perceived importance of effect size regardless of chart type (e.g., lines versus bars) and despite attempts to signal that an axis does not start at zero (e.g., breaking the axis) [40]. Expanding the axis scale on a chart that displays inferential uncertainty (e.g., 95% confidence intervals) to the scale implied by descriptive uncertainty (e.g., 95% predictive intervals) can reduce bias in impressions of

effect size [88].[2] In Chapter 4, I extend this line of work by investigating how the visual distance heuristic impacts decision making from visualizations.

2.2.3 Casual inference

In Chapter 5, I present a pair of experiments looking at how people make causal inferences about data generating process (DGP) when they explore visualized data. This is in some sense a model check (see Section 2.2.1) where users judge the compatibility of data with different provisional models. However, in this scenario the models represent parameterized causal explanations for the patterns in the dataset, and that makes this a sensemaking task [31, 78, 164, 170] with the goal of inferring the underlying DGP in order to understand the data. The software tool presented in Chapter 6 is an attempt to support this type of thinking and to address shortcomings in conventional tools for visual causal inference.

Much of the psychology and statistics literature on visual aids for causal reasoning focuses on contingency tables (e.g., [4,12,35,75,77,180]). Contingency tables support causal inferences by using layout to encode conditional probabilities, the same way trellis plots afford grouping by factors during visual data analysis [15, 196]. Whether or not a factor seems to be *"collapsible"*—whether or not patterns the data seem to change depending on whether the data are grouped by that factor—can be a visual signal for reasoning about causal relationships such as confounding [75]. However, empirical research on interpretation strategies for contingency tables [12] suggests that analysts often misinterpret signals like collapsability because they don't ascertain the mapping between these visual signals and hypothesized causal relationships. Tools like Tableau enable users to explore collapsability by interactively grouping data. Some of the experimental manipulations in Chapter 5 [111] investigate whether the ability to interactively control data aggregation improves the quality of causal inferences.

Research on visual analytics (VA) employs a broader range of representations to support causal reasoning, including parallel coordinates [207,208], bar charts [224], "diff bar charts"

[2]Predictive intervals are by definition wider than confidence intervals from the same model, so showing confidence intervals on a scale expanded to include the range plausible predictions makes confidence intervals look smaller.

showing counterfactual outcomes under different conditions as layered bars [222], and novel techniques using animation to show event sequences (e.g., [53, 101, 104]).

Some of these tools also incorporate directed acyclic graphs (DAGs) as interfaces to models and visualizations (e.g., [207,208,222,224]). *DAGs* are a graphical representation used for causal reasoning—with nodes representing variables and arrows between nodes representing posited causal relationships—which have garnered attention in recent years [160]. DAGs encode hypothesized relationships among variables (e.g., Fig.5.1Ⓑ), making causal relationships and the assumptions they entail explicit and in some cases testable [9, 157–159]. We use DAGs in Chapter 5 [111] to present differences between alternative causal explanations for data sets that we ask participants to judge (Figs. 5.2 & 5.9).

VA systems frequently use interaction techniques such as cross-filtering linked views of data (e.g., [207, 208]) and click- or drag-and-drop-based chart construction (e.g., [187, 222, 224]). The most similar prior research[3] to the work presented in Chapter 5 [111] tests whether constructing charts by clicking on variables versus dragging variables onto a DAG makes a difference in analysts' ability to differentiate between different kinds of causal relationships [224], specifically identifying mediating variables. Although they do not find an effect of interaction technique on causal inferences, the authors provide a detailed strategy analysis extending evidence from psychological studies [12] that analysts struggle to reason about the exact set of visual signals they should look for to verify or falsify a causal relationship. I extend this line of work in Chapter 5 [111] by studying whether the ability to (un)facet charts or cross-filter coordinated multiple views impacts the quality of untrained analysts' causal inferences.

Prior work in visualization [145,151] and risk communication [63,183] suggests that icon arrays can improve Bayesian inferences such as those implicated in causal induction [77], perhaps in part because of cognitive benefits of framing probabilities as frequencies of events [69, 87, 98, 115]. However, the performance improvements associated with these visualizations seem to depend on the use of text in the display [145], the chart user's spatial ability [151], and the strategies users rely on [109]. In Chapter 5 [111], we compare icon

[3]Also, see the class project from my collaborator Yifan Wu that kicked off this study `https://logical-interactions.github.io/causal2020/`

arrays to text tables and bar charts since these visualizations span the design space for showing count data (from a fake clinical trial) and are also easy to create in VA software like Tableau.

2.3 Behavioral models of data cognition

In order to describe how people use uncertainty visualizations to perform tasks such as inferences and decisions, I draw on behavioral modeling techniques from mathematical psychology and behavioral economics. I introduce these models as they have been applied in their disciplines of origin, and I describe how I use and extend them in my dissertation.

2.3.1 Psychometric functions

One type of model that makes multiple appearances in my dissertation are *psychometric functions* (PFs) [121]. PFs describe performance on two-alternative forced choice (2AFC) tasks, such as the judgment of whether the jobs market is growing or not in Chapter 3 and the decision about whether or not to pay to add a new player to a fantasy sports team in Chapter 4. The sigmoid (a.k.a. S-shaped) functional form of PFs is derived from *signal detection theory* (SDT) [73], a model of how we make choices by weighing evidence in favor of two alternative interpretations of a stimulus (Fig. 2.1). Where SDT describes noisy perception of signal strength in a given stimulus, the PF extrapolates the functional relationship between signal and classification accuracy across a range of signals. For our purposes the stimuli are always displays of data, and the underlying strength of evidence is a statistical attribute that describes the difficulty of the 2AFC judgment. Traditionally, psychologists use PFs to estimate perceptual sensitivity to simple stimuli—e.g., our ability to distinguish between sounds of different loudness where the sounds are the stimuli and their physical amplitude is the signal or strength of evidence [73].

Different parameterizations of the PF enable us to estimate different summary statistics describing a person's performance. In both Chapter 3 [110] and Chapter 4 [109], I use PFs to estimate *just-noticeable differences* (JNDs), the change in signal which would be discernible to a subject 75% of the time. JNDs are the most common summary statistic derived from

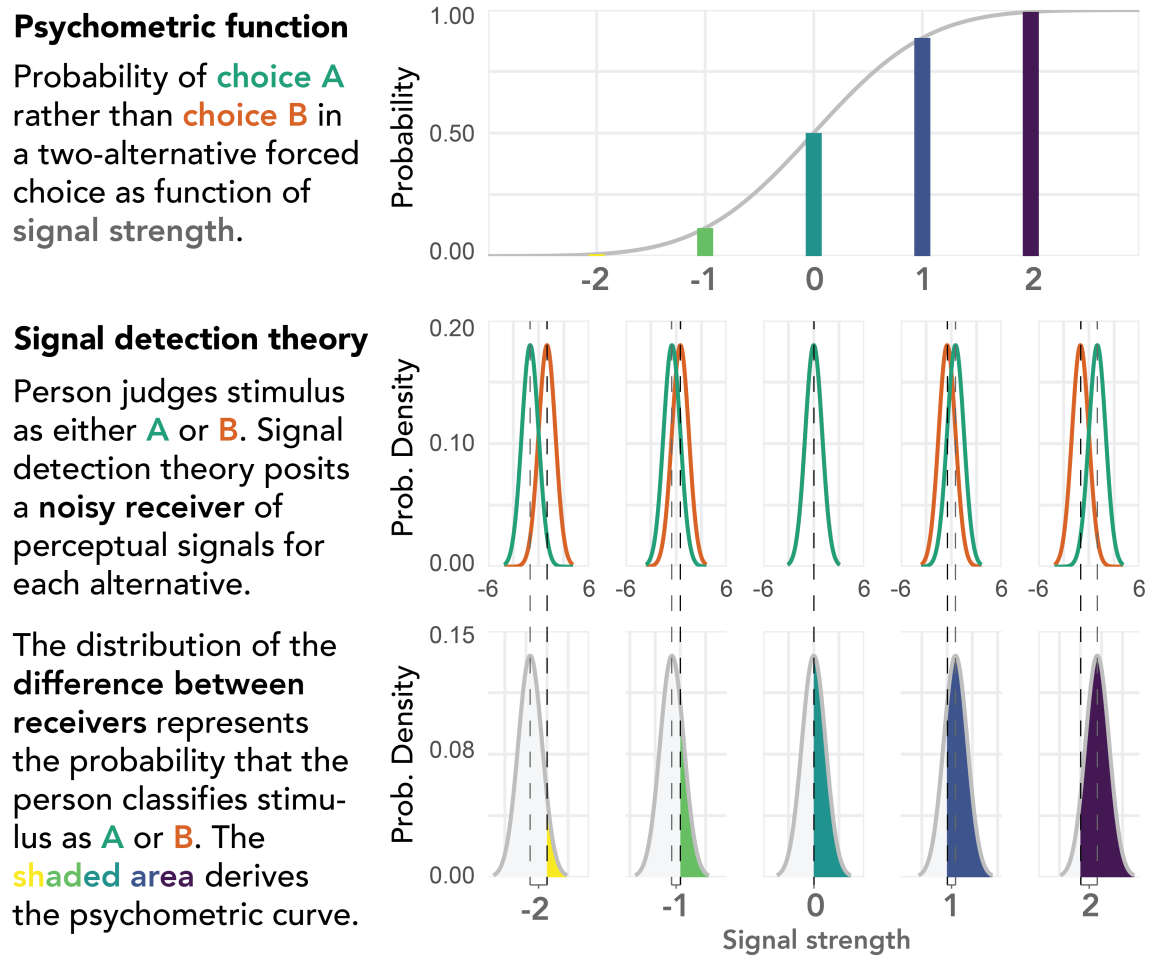


Figure 2.1: The relationship between a psychometric function (PF) and signal detection theory (SDT), demonstrated for a two-alternative forced choice (2AFC) task with five different stimuli (x-axis in the top row, columns across the lower rows). The PF (top row) estimates the performance of an observer classifying any such stimulus as one of two alternative interpretations, choice A or B, as a function of the signal conveyed in a stimulus or the strength of evidence for the 2AFC judgment. *The form of the PF is derived from SDT*. SDT posits "noisy receivers" in the mind of the person performing the task for each of the two alternative interpretations of a stimulus, choices A and B. The responses of these receivers to a stimulus are probabilistic, representing the perceived strength of evidence in favor of each choice. The five stimuli (shown in columns of the lower rows) each convey levels of signal equal to the mean difference between the receiver distributions (middle row), annotated with the black and gray vertical lines on the x-axes. The bottom row depicts the distribution of the difference between receiver for choices A and B. We derive the PF by taking the integral of the difference distributions at a given level of signal strength.

PFs, often used to describe sensitivity to signal under different conditions. In visualization research, JNDs have been applied to create perceptually-uniform color palettes (i.e., CIELAB [60]) and perceptually-driven color palette selection tools [130, 189], among other applications. In Chapter 3 [110], I use a custom parameterization of PFs from perceptual psychology [121] that enables me to derive *PF spreads*, a measure of noise in the "receivers" posited by SDT (Fig. 2.1, standard deviation of the distribution in the middle row) which I use to analyze the calibration of participants' confidence ratings (following Sanders et al. [171]). In Chapter 4 [109], I parameterize PFs as logistic regression in order to estimate a bias parameter called the *point of subjective equality* (PSE), the level of signal where a person is expected to split their responses evenly between two alternative interpretations of the data (e.g., expected to pay for the new player 50% of the time).

The primary thing I do differently from most prior work is that I use PFs in scenarios where the definition of "signal" is more abstract and conceptual, such as a likelihood ratio representing the strength of a statistical inference in Chapter 3 [110] or a utility function representing the expected payoff for a decision in Chapter 4 [109]. The most similar applications of PFs to date in visualization research use correlation as a measure of signal [80, 114, 168].

2.3.2 Causal support

In Chapter 5 [111], I draw on and extend a model of causal reasoning called causal support, first proposed by Griffiths and Tenenbaum [77]. *Causal support* formulates causal inferences as a Bayesian update on the log odds of a finite set of causal explanations given some observed data. Mathematically, causal support has similar properties to a Chi-squared test (i.e., Are the data in each cell of a contingency table likely generated by the same process?), which prior work analogizes to the kind of comparisons between data and model predictions that analysts visualize in "model checks" [66, 67, 96] such as QQ-plots. However, unlike a Chi-squared test, causal support relies on Monte Carlo simulations to assign likelihoods under alternative causal explanations, making causal support extensible to any finite set of generative causal models. For instance, Pacer and Griffiths extend causal support to handle continuous data [153] and event streams [152]. Similarly, in the second experiment

of Chapter 5 [111], we present an extension of causal support to evaluate inferences about more than two possible data generating models.

Previous cognitive models of causal inference explored in psychology share more in common with parameter estimation than statistical inference per se, a subtle but important distinction. One such model *delta p* posits that that people judge differences in conditional proportions of observed events when making causal inferences about count data [4]. Another such model *causal power* posits that people judge the magnitude of effect size when making causal inferences [35]. Both of these predecessors to causal support make the assumption that causal inferences are fundamentally a perceptual judgment, however, causal power rescales delta p based on the potential to detect any signal whatsoever within the observed data. In contrast, causal support assumes that the signal for causal inferences depends on the possible data generating models that the analyst has in mind and represents these alternative models explicitly. This makes causal support more flexible, with higher predictive validity for human judgments than delta p, causal power, and even Chi-squared [77]. Causal support reflects analysts' natural tendency to dichotomize, for better or worse, reasoning about whether or not causal relationships exist rather than how strong they are.

2.3.3 Other behavioral models

Other behavioral models of data cognition presented in my dissertation fall more within the framework of regression. A few principles guide the construction of these models:

- We want models whose assumptions preclude responses that are impossible or out of bounds on the scale the participants use to answer the questions we ask them. This often means using a certain link function in a generalized linear model or selecting certain priors for Bayesian regression.

- We want to account for the fact that people's perception of most quantities is logarithmic [185, 204], in the sense we expect to see multiplicative (rather than additive) distortions in the perception of visual properties of a display such as length, area, brightness, count, or frequency.

- We want to account for the fact that people's perception of probabilities tends to follow a functional form that is linear in log odds units [71, 228].

- We want to account for the fact that judgements on a given scale are often made relative to some reference value(s) in ways that produce local patterns of biases in response near the reference point(s) [91].

I will often make reference to these principles of behavioral modeling in motivating the choice of models in subsequent chapters.

Chapter 3

HYPOTHETICAL OUTCOME PLOTS HELP UNTRAINED OBSERVERS JUDGE TRENDS IN AMBIGUOUS DATA

One of the cognitive mechanisms I highlight in my dissertation is *ensemble processing* or the visual system's ability to automatically, quickly, and accurately decode summary statistical information from what we see. In this Chapter, I present the first project from my doctoral research, where the study team[1] and I investigated how well hypothetical outcome plots [98] support inferences about the underlying data generating process (DGP) that might have produced noisy samples of timeseries data. This work [110] was initially published at IEEE VIS and has been edited for coherence with the overall dissertation.

3.1 Calibrating inferences about visualizations in the news

Members of the public often encounter visualizations in the news and in online media, and journalists invite them to assess the underlying trends in these data. This kind of judgment inherently involves reasoning about uncertainty, e.g., "How likely is it that I would see this data if a particular trend was true?" In this study, we examine a scenario inspired by a piece in the New York Times (NYT) called "How not to be mislead by the jobs report" [99], where chart users were shown visualizations of what the monthly jobs report would look like if the economy was stagnant versus growing at a given rate (Fig. 3.1). For an individual news reader, the stakes of judging whether the job market is growing might be that these jobs report numbers inform their short-term investment decisions (e.g., in their retirement portfolio). For the broader public, interpretation of the jobs report impacts sentiments about the economy, which in turn might impact voting behavior or political capital.

In public facing data presentations, visualization authors often omit uncertainty information [94], in part because of distinct challenges around modeling and communicating

[1]The original study team for this work was me, Francis Nguyen, Matthew Kay, and Jessica Hullman.

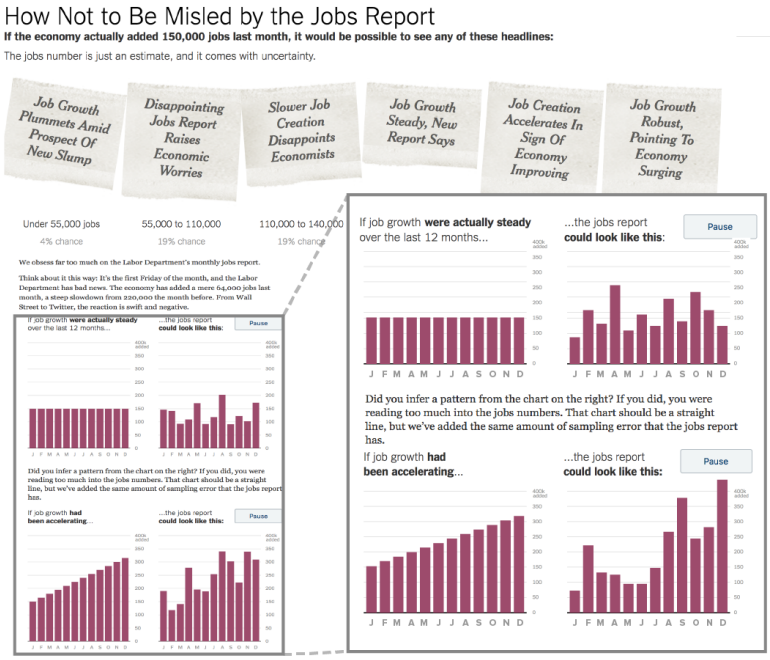


Figure 3.1: A condensed view of the NYT article [99] that inspired our study design.

uncertainty.[2] For this reason, uncertainty visualizations in the news deserve special attention from researchers as a use case for visualization. In this study, we perform a light-weight content analysis of uncertainty visualizations in the news, finding that the NYT visualization on "How not to be mislead by the jobs report" [99] asks audiences to think about multiple statistical concepts commonly implicated in uncertainty visualizations in the news. We argue that this makes the NYT example ideal for uncertainty visualization research.

Design choices in uncertainty visualizations make it more or less possible for chart users to rely on different information processing approaches (a.k.a. cognitive mechanisms). For example, hypothetical outcome plots or HOPs [98] should enable ensemble processing because they encourage the user to integrate information about a distribution of possible outcomes across a series of examples. On the other hand, more conventional displays such as error bars do not make information about individual samples available in the same way, so chart users should not be able to internalize these distributional representations using en-

[2]This work on "Why Authors Don't Visualize Uncertainty" [94] came after the study presented in this chapter but nonetheless provides important context. Back in 2016, when we were designing the study presented here, the study team had the sense that uncertainty visualization in the news was relatively uncommon and that approaches to visualizing uncertainty in this context were somewhat scattershot.

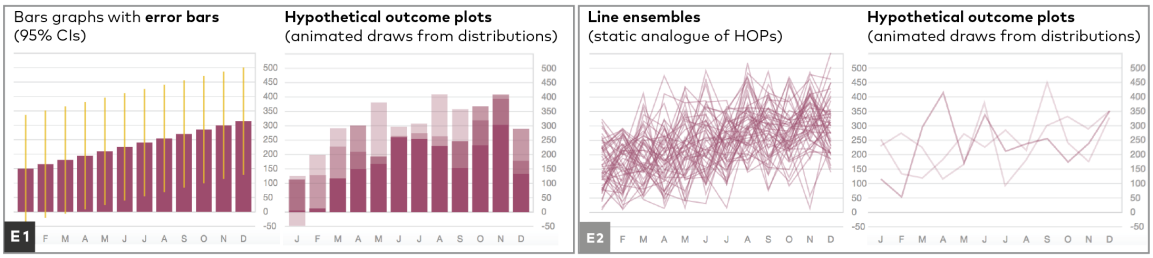


Figure 3.2: We present two experiments (E1 and E2) evaluating four different uncertainty visualizations (from left to right): bar graphs with error bars, bar hypothetical outcome plots (HOPs), static line ensembles, and line HOPs.

semble processing. Visualization approaches that better support ensemble processing might improve the user's ability to internalize expectations about variability due to sampling error, potentially leading to better calibrated visual inferences about the data generating process underlying noisy samples of data like the monthly jobs report numbers.

In this chapter, I present two crowdsourced experiments investigating (1) whether visualization designs that promote ensemble processing lead to better-calibrated visual inferences, and (2) whether static versus animated displays of hypothetical outcomes differ in their ability to support calibrated visual inferences through ensemble processing. In both experiments, Mechanical Turk workers infer which of two possible underlying trends is more likely to have produced a sample of time series data by referencing uncertainty visualizations which depict the two trends with variability due to sampling error (Fig. 3.3). We use this scenario in both experiments to test people's performance on the task when using error bars, HOPs, or line ensembles as a reference (Fig. 3.2). We find that chart users can accurately infer the underlying trend in samples of data that are more ambiguous when using HOPs rather than error bars or line ensembles.

3.2 Identifying uncertainty judgments in the news

We conducted a lightweight, formative content analysis investigating the use of uncertainty visualizations in the news and how people might interpret visualized uncertainty in this context. We use this analysis to inform our experimental task.

3.2.1 Selection of articles

We gathered a convenience sample of 22 articles from mainstream news in the US and UK. We started with recent visualizations from publishers like the NYT, and supplemented this set by conducting Google News searches containing terms like 'uncertainty' or 'probability' and 'visualization' or 'chart' (see Supplemental Materials[3]). Articles were included in our analysis (1) if they are primarily concerned with communicating uncertainty information and (2) if they contain at least one visualization which conveys that there are a *distribution of possible values* which an observation or prediction might take.

While we started with 32 articles which met our first inclusion criterion, we had to exclude ten of these articles because they failed to meet our second criterion. In these cases, the visualizations only encoded single values of each estimate (e.g., a central tendency), and visualized distributions were contextualized as separate events plotted together by geographic location or time (e.g., an estimate for each of multiple countries). That the media use distributions of outcomes for separate events to illustrate uncertainty (e.g., Fig. 1.1) speaks to a disconnect between public concerns about uncertainty and the notion of uncertainty employed in information visualization research, in which uncertainty is most often defined as an expression of error in an estimate [166,192]. The sample of articles considered reflects the time frame of the study and should be considered representative of uncertainty visualizations published in the popular press from 2014 to early 2018.

3.2.2 Qualitative coding

Our goal was to identify a set of distinctions that could help visualization researchers understand the types of tasks implied by uncertainty visualizations in the news. We used a bottom-up iterative coding process in order to see what distinctions emerged from close analysis of the examples in our corpus. Two authors separately coded each article initially, listing for each a set of concrete questions about the visualized data that the article posed to the audience. We identified these questions by considering how the uncertainty visualizations support the narrative in the text of each article. Through several additional passes on

[3] `https://github.com/kalealex/jobs-report-hops`

the corpus, the two authors iteratively reviewed and refined the codes. We arrived at a set of yes-or-no distinctions used to label and characterize each question posed by the articles.

3.2.3 Results and discussion

The results of this analysis characterize the ways that uncertainty visualizations are used in the news and ultimately justify our choice to adopt the NYT piece on "How not to be misled by the jobs report" [99] as a task scenario for our experiments.

Codebook: Conceptual distinctions

The first three labels we coded characterize the *low-level visual tasks* implied by the text and the visual encodings. These codes answer the question: Is the relevant information from the uncertainty visualization obtainable through (1) direct reading of encoded values, (2) ensemble processing [3], or (3) shape perception [57]? While direct reading is highly deliberate, ensemble processing is an automatic visual computation about sets of graphical encodings. Shape perception occurs when the contours of graphical elements (e.g., tops of bars, trend lines) are recognized as patterns which have a specific meaning to the user. Most visualizations can be read in more than one way, so these labels are not mutually exclusive.

The second set of labels describe *what information is needed* to answer each question. Does the task primarily require the audience to consider (1) central tendency or (2) variance of a distribution? These labels are not mutually exclusive. To differentiate tasks that emphasize uncertainty from those that emphasize central tendency, we only assign a question with both codes if it is not possible to identify which aspect of the data dominates the task.

The final pair of labels concern the *high-level cognitive operations* which the observer performs with visually extracted information in order to address each question posed by the article. Do the concrete task and visual encoding require observers to make (1) comparisons or (2) inferences? The reader must compare values or distributions when the question posed by the article requires a relative judgment (e.g., differences of value or conditional likelihood). Inference occurs when a user is asked to reason about an aspect of the data that is not directly encoded (e.g., the central tendency of a set of points).

Frequency of codes

We identified a total of 87 questions that were implied in the 22 articles we analyzed. The information the reader needed to glean from the visualizations in order to answer these 87 questions could be read directly from the visualization in 60 (69.0%) cases, could be acquired through ensemble processing in 56 (64.4%) cases, and could be read through shape perception in 60 (69.0%) cases. This shows that designers often encode relevant information in multiple different ways (e.g., providing a trend line as well as individual data points).

We observed 19 (21.8%) questions that were primarily about central tendency and 34 (39.1%) questions that were primarily about uncertainty. Only 33 of the questions (37.9%) were concerned with both central tendency and uncertainty without one clearly dominating the task. This suggests that judgments about uncertainty in news contexts are often framed as separate tasks from judgments of central tendency. Empirical evidence in uncertainty visualization [98, 132, 155, 156] and cognitive science [103, 199] suggests that effective uncertainty visualization should promote the integration of both central tendency and uncertainty information because otherwise people tend to substitute potentially error prone heuristics for more nuanced judgments of uncertainty.

Of the 87 questions examined, 64 (73.6%) involved comparisons between individual data points and/or distributions, and 72 (82.8%) involved inferences about information not directly encoded in the visualizations. Both comparison and inference seem to be common cognitive operations required to interpret uncertainty visualizations in the news. This suggests that when designing and evaluating uncertainty visualizations for a broad audience, it is not sufficient to simply measure ability of observers to extract directly encoded values from a given chart.

Selecting a representative uncertainty judgment task

We wanted a task for our experiments that required more sophisticated forms of judgment than simply reading directly encoded values. We also wanted to choose a task that required considering both uncertainty and central tendency. We used the results of our coding to identify a concrete question from a visualization and article originally presented by the NYT:

"How not to be misled by the jobs report" [99][4]. In our adaptation of this task (Fig. 3.3), participants decide which of two underlying trends, growth or no growth in the job market, is more likely to have produced an observed sample of job growth numbers. The choice between trends requires the participant to consider how variance due to sampling error impacts samples of hypothetical jobs numbers. The participant must visually extract the trend in a hypothetical sample of jobs numbers through either ensemble processing or shape perception. Then, considering the central tendency and variance of the two possible trends based on uncertainty visualizations, participants make inferences about the likelihood of a given sample under each trend. Thus, our task entails reasoning spanning all of our distinctions except for the direct reading of individually encoded values.

The task we selected is representative not only of how uncertainty visualization is used in the news but of tasks that are implicated in statistical literacy. By asking participants to make inferences from samples of noisy data, our task engages core statistical competencies. In order to understand applications of statistics in science, people must recognize that samples may not always resemble the model or population from which they are produced [62, 65]. In our task and in public-facing presentations of statistics (e.g., election models), there are components of variance in sampling which are accounted for by underlying trends (signal) and components of variance which are due to sampling error and other random processes (noise). The ability to recognize and parse these sources of variance is implicated in our task and is important to the comprehension of statistics [62, 65].

3.3 Experiment 1

Experiment 1 mirrors the visualization design and data interpretation context of the NYT article "How not to be misled by the jobs report" [99]. We evaluate how well HOPs facilitate well-calibrated visual inferences about the underlying trend in noisy data compared to a more conventional encoding of uncertainty—error bars. We choose error bars as the control visualization for this task because (1) they are a common static encoding of uncertainty and (2) they can be used to add uncertainty information to the bar encoding used in the NYT,

[4] `https://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html`

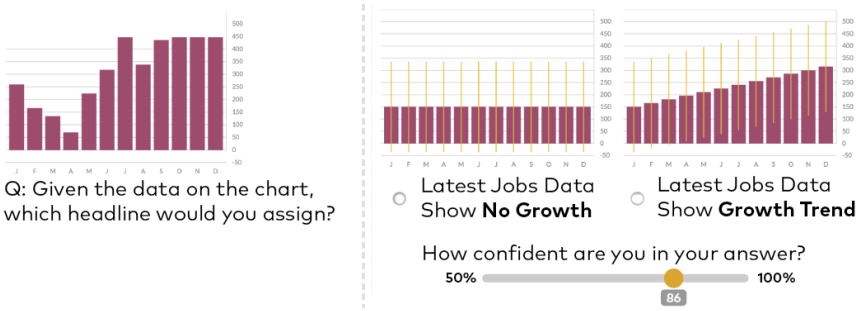


Figure 3.3: A depiction of the task interface used in our studies (stimuli for Experiment 1 are shown). The chart that participants judge on the current trial is on the left side of the display. The reference uncertainty visualizations for the "no growth" and "growth" trends are on the right side of the display. Underneath the uncertainty visualizations, the participant uses radio buttons to indicate which trend is more likely and the slider to rate their confidence.

enabling a controlled comparison that remains faithful to the original presentation.

3.3.1 Methods

In a crowdsourced experiment on Mechanical Turk, we ask participants to discriminate which of two possible underlying trends is more likely to have produced hypothetical samples of jobs added to the economy each month of a simulated year (Fig. 3.3). Thus, we embed the visualizations from the NYT [99] into a traditional two-alternative forced choice (2AFC) experiment, styled after traditional methods in psychophysics [73, 121].

Procedure

Upon accepting the Human Intelligence Task (HIT), participants were redirected to a web page containing instructions on the task. Participants were told they would play the role of a newspaper editor who is presented with a bar chart of jobs added to the economy each month of a simulated year and asked decide on a headline about the growth trend in the job market for that year (Fig. 3.3). On each trial, the participant made a 2AFC judgment ("growth" or "no growth") about one bar chart and provided a rating of their confidence in this judgment on a scale of 50 (random guess) to 100 (absolutely certain). Each participant completed a block of 60 trials using error bars as a reference for the task and another block of 60 trials with HOPs, where the order of the visualization conditions

was counterbalanced across observers. At the end of 120 trials, each participant completed a brief demographic survey, including questions about familiarity with statistics and with the specific visualizations shown in the task. The HIT carried a reward of $8 ($29.32 per hour on average).

Measures and hypotheses

Our dependent measures are parameter estimates derived from psychometric functions (PFs, Fig. 3.5) and a related approach for modeling confidence data. These are models of forced-choice behavior based on signal detection theory (SDT, see Section 2.3.1, Fig. 2.1).

Psychometric Functions (PFs). For each participant's 2AFC responses under each visualization condition, we fit a PF [121] estimating two parameters: (1) the JND, which is the level of evidence at which the participant is expected to perform with mean accuracy on the task; and (2) the spread of the PF, which describes the noise in the participant's perception of evidence in the task (Fig. 3.5). We predicted that when uncertainty due to sampling error was visualized using HOPs, as compared to error bars, subjects would have smaller estimated JNDs on average, indicating that observers require less evidence to distinguish the trend underlying a noisy sample. We also predicted that users would have smaller estimated PF spreads, indicating that they find stimuli ambiguous across a narrower range of evidence.

Confidence Fitness. Existing approaches to confidence analysis in uncertainty visualization evaluation tend to either assume that more confidence is better (e.g., [19]) or analyze the correlation between confidence and accuracy over a set of judgments (e.g., [39]). However, both approaches may be inadequate to deal with the complexity of confidence as a construct; for example, research in judgment and decision making describes how confidence reporting is often noisy and may not adhere to the laws of probability [150]. To overcome these limitations, we adapt an approach from psychophysics [171] which explicitly models the noise in an observer's confidence reporting process.

Confidence fitness can be interpreted as an estimate of the degree to which confidence ratings are coherent with a probabilistic interpretation. The model is based on SDT [73]

and assumes that confidence ratings from the ideal observer should reflect the accuracy of their judgments at the given level of evidence [171]. To estimate a ground truth for confidence, we simulated many trials for each observer based on their PF. We estimate the ideal confidence of an observer as the accuracy of these simulated noisy judgments at each level of evidence. Confidence fitness is a latent parameter of the model ranging from 0 to 1 which estimates the degree to which actual confidence reporting is random or ideal. We had no strong *a priori* predictions about confidence fitness.

Stimuli and trial generation

Stimuli, Units of Evidence, and Task Difficulty. The PF fitting process requires a single measure of stimulus intensity which approximates the difficulty of the judgment task for each stimulus. Stimulus intensity is used as a ground truth to determine whether or not judgments are correct. In our task, the intensity measure should quantify the strength of the evidence in favor of a growth or no growth interpretation of a given sample of jobs numbers. We calculate the intensity based the probability of each possible trend having produced the sample of jobs numbers. For any given stimulus (set of hypothetical job numbers), strength of evidence is the log likelihood ratio describing the relative likelihood that the stimulus was produced by the no growth or growth trend in the job market (Fig. 3.4). This produces a log-linear intensity scale where positive values represent evidence in favor of no growth (e.g., bottom row of Fig. 3.4), negative values represent evidence in favor of growth (e.g., top row of Fig. 3.4), and zero is the point of maximum uncertainty. Because our measure of intensity should be consistent whether the evidence more strongly favors the growth or the no-growth trend, we take the absolute value of this log likelihood ratio.

$$evidence = \mid \log_{10}\left(\frac{\Pr(D|\theta_{noGrowth})}{\Pr(D|\theta_{growth})}\right) \mid$$

Here, D is a given set of job growth numbers, $\theta_{noGrowth}$ is a trend in which the jobs added each month are normally distributed about 150k with a standard deviation (SD) of 95k, and θ_{growth} is a trend in which there is a linear increase of 15k jobs per month from 150k in January to 350k in December, with a SD of 95k jobs each month (Fig. 3.3, right side).

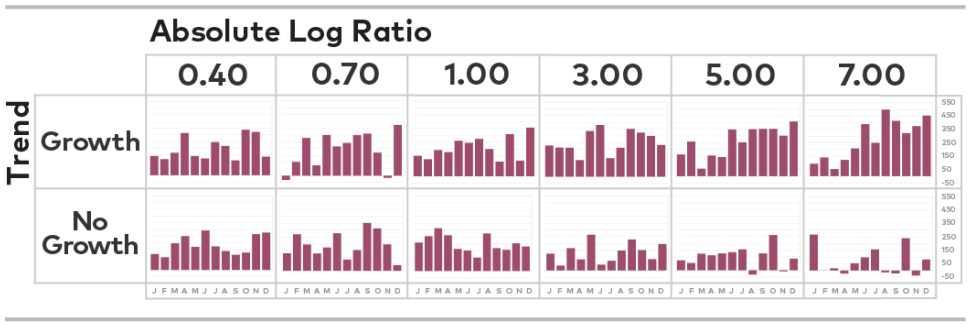


Figure 3.4: Example stimuli judged by participants in our task. Scan across the figure to get a concrete sense of how our units of evidence translate into the appearance of a stimulus. Units of evidence are the log ratio of the probability that each stimulus was produced by the growth vs no growth trends, given shared variability due to sampling error. We take the absolute value of the log likelihood ratio so that units of evidence have the same sign regardless of which trend produced the stimulus.

For both trends, we matched the mean job growth for each month to the NYT article [99]. However, we differed from the NYT article by using a SD of 95k jobs instead of 55k. We chose this SD to guarantee that there were many visually distinct stimuli to sample for which the underlying trend was ambiguous.

Staircase Sampling Procedure. A major challenge in obtaining valid PF fits is making sure that the observer completes enough trials at stimulus intensities which are ambiguous. Choosing stimuli which the participant can judge correctly, but which are not easy, reduces the uncertainty in the fitting process [177]. However, the researcher must not present too many trials, otherwise issues of participant attrition and fatigue arise. Adaptive sampling procedures are the best solution to this problem. These algorithms sample stimuli at levels of evidence which are ambiguous but not uninformative based on the participant's past performance. We used a three-down, one-up staircase (suggested in [64]) in which the level of evidence was incremented (i.e., made easier) by 3 absolute log likelihood ratio units each time the participant guessed wrong, and the level of evidence was decremented (i.e., made harder) every third time the participant was correct. In order to avoid autocorrelation in performance resulting from participants noticing the sampling procedure, we randomly interleaved 25 trials each from two different staircases (suggested in [38]) as well as 10 gold standard trials at an absolute log likelihood ratio of 9 (very easy). The two staircases differed only in their decrementing step sizes of 2.22 and 1.65 absolute log likelihood ratio

units, respectively. Step sizes were chosen based on pilot data and recommendations from simulation studies on how to create staircases with stable convergence [64] and how to sample in order to minimize uncertainty in the parameter estimates from PFs [177]. These staircases promoted meaningful PF fits in a minimal number of trials.

Uncertainty Visualizations. In our task, participants use different uncertainty visualizations as a reference showing the no growth and growth trends (Fig. 3.3, right side). HOPs were generated by repeatedly sampling 12-month sets of jobs added to the economy from each underlying trend, plotting these numbers in bar charts, and animating transitions between frames. Animated transitions between bar values were 500 ms in duration with a 10 ms delay between each bar and a frame rate of 45 hz. Each sample was displayed for 1500 ms in between the animated transitions. Animation parameters followed those used by the NYT. We used a cubic easing function for animated transitions, consistent with the NYT.

Participants

We recruited 62 MTurk Masters workers, each located in the United States and with a HIT approval greater than 95%, to participate in the study. A power analysis on pilot data suggested our target sample size should be 50 within-subjects comparisons of HOPs and error bars, where each participant completes a block of 60 trials for each visualization condition and the condition order is counterbalanced across participants. To achieve this, we iteratively sampled small batches of participants ($n < 10$), alternating the starting visualization condition, and applying a set of *a priori* exclusion criteria to the data collected. This was procedure was repeated until we had 50 participants in our sample after exclusions, and condition order was counterbalanced across participants. During the iterative sampling procedure (see Preregistration[5]), one participant was excluded due to exceptionally poor performance, and five participants were excluded from analysis due to poorly converging psychometric function fits. We accidentally collected six participants more than intended, so six participants were chosen at random and excluded from analysis thereafter.

[5] `https://osf.io/975us/register/5771ca429ad5a1020de2872e`

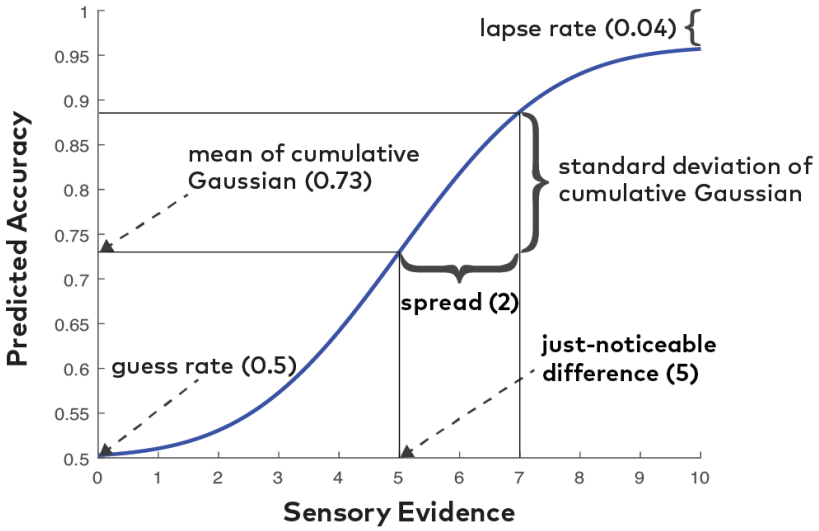


Figure 3.5: An illustration of the psychometric function with relevant parameters.

Analysis approach

Psychometric Function Estimation. We used a maximum likelihood optimization algorithm to fit a cumulative Gaussian distribution to model accuracy as a function of the level of evidence presented in each stimulus (Fig. 3.5). Recall that evidence is the degree to which the data in a stimulus informs the forced choice between the two possible underlying trends of growth and no growth. The lower asymptote of this cumulative Gaussian was set to a *guess rate* of 0.5 (Fig. 3.5, lower left), which would be achieved in theory if the participant guessed on each trial. The upper asymptote of the cumulative Gaussian was set to one minus a *lapse rate* (Fig. 3.5, upper right), the base rate of stimulus-independent errors due to lapses of attention. Following recommendations from a simulation study on PF fitting [214], we estimate lapse rate as a constrained free parameter between 0 and 0.06 in order to reduce bias in other parameter estimates. The level of evidence corresponding to the mean of the fitted cumulative Gaussian is the *JND* (Fig. 3.5, lower right). The standard deviation of the cumulative Gaussian is the *spread* parameter of the PF (Fig. 3.5, middle right). Per SDT (Fig. 2.1), the spread parameter represents noise in perception of the level of evidence, where PFs with greater spreads indicate that the trend in the jobs report is ambiguous across a larger range of evidence.

Confidence Fitness Estimation. We used maximum likelihood estimation to find an

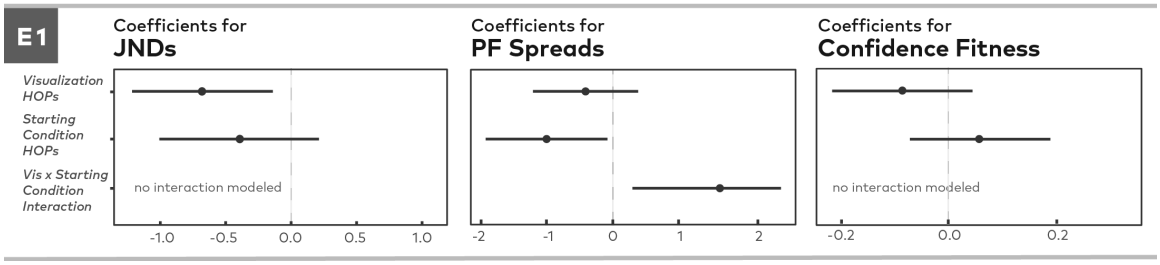


Figure 3.6: Estimated regression coefficients for mixed-effects linear models of each psychophysical response variable in Experiment 1 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

optimal value of confidence fitness for each PF. This algorithm was transcribed from Sanders et al. [171]. We provide a detailed description of the confidence fitness algorithm and our implementation in Supplemental Materials.

Statistical Inference. We used mixed-effects linear models in lme4 for R to estimate each response variable (i.e., JND, PF spread, and confidence fitness) as a function of fixed effects of visualization condition and condition order and a random intercept effect of participant. Further, we believed that an interaction between visualization condition and condition order would explain variance in parameter estimates due to practice or learning. Following the procedure in our preregistered analysis plan[6], for each measure we used ANOVA to compare model residuals with and without the interaction between visualization condition and condition order, and we chose the most parsimonious mixed-effects model which still minimized the variance of residuals. We provide analysis scripts and data in a project repository[7].

3.3.2 Results and discussion

We report results separately for each of our metrics of interest, JNDs, PF spreads, and confidence fitness.

Just-noticeable differences (JNDs)

We modeled fixed effects for visualization condition and condition order and a random intercept term to account for individual differences, as described above. We did not include

[6] `https://osf.io/975us/register/5771ca429ad5a1020de2872e`

[7] `https://github.com/kalealex/jobs-report-hops`

the interaction between visualization and starting condition in our model of JNDs because including the interaction did not reduce residual variance.

We found that JNDs were lower in the HOPs condition (Fig. 3.6, left) than in the error bars condition (Est = -0.68; 95% CI: [-1.19, -0.14]; $t = -2.49$; $p = 0.02$). The second column of Figure 3.4 illustrates an effect of roughly this size. This effect suggested that on average when participants used HOPs rather than error bars they required less evidence about the underlying trend in a noisy time series to achieve mean accuracy (around 75%). Put differently, participants could successfully discriminate the underlying trend in the job market for more ambiguous stimuli when using HOPs.

Psychometric function spreads

We modeled visualization condition, condition order, and their interaction as fixed effects and a random intercept for individual differences. We found no reliable differences in PF spread estimates between visualization conditions (Fig. 3.6, middle). However, PF spread estimates were lower among participants who started in the HOPs condition (Est = -1.00; 95% CI: [-1.91, -0.09]; $t = -2.12$; $p = 0.04$) compared to participants who started with error bars. Recall that narrower PF spreads indicate less noise in the perception of evidence. We speculate that participants may have developed mental representations for the the two possible growth trends during the first block of trials and relied on the uncertainty visualizations less in later trials. There was also an interaction between visualization condition and starting condition (Est = 1.46; 95% CI: [0.34, 2.58]; $t = 2.54$; $p = 0.01$), indicating a learning or practice effect. Regardless of the visualization condition under which data was collected, PF spread estimates tended to be larger in the first block of trials than in the second block. We speculate that participants became more sensitive to evidence in the task with practice.

Although we intentionally avoided the use of practice trials or feedback in order to test the behavior of untrained observers, it seems that practice or learning introduced differences across blocks which were not necessarily related to visualization condition. While the effect of starting condition on users' sensitivity to evidence could mean that HOPs facilitate

superior learning about sampling error, this effect could also be explained as an artifact of testing visualizations within-subjects and keeping the task consistent across blocks. The ambiguity of this result points to considerations about experimental design in visualization evaluation. If the goal of the experiment is to measure learning, a longitudinal experiment which employs practice trials to control for practice effects is ideal. In studies like ours where the goal is to measure visualization effectiveness, future experiments in this paradigm should employ a between-subjects design, as we do in Experiment 2, in order to mitigate ambiguity about which visualization condition is informing a user's mental representation of the task.

Confidence fitness

We modeled fixed effects of visualization and condition order and a random intercept for individual differences. We did not model the interaction of visualization and starting condition. Although confidence fitness within individuals was seldom consistent across blocks (see Supplemental Materials), we found no reliable effect of visualization or condition order on confidence fitness (Fig. 3.6, right).

Since the results of the confidence fitness analysis were inconclusive and difficult to interpret, we conducted a more conventional *exploratory analysis* on confidence reporting. We used a mixed-effects linear model on trial-level data to estimate reported confidence based on fixed effects of visualization, starting condition, and their interaction as well as fixed effects of the level of evidence conveyed in each stimulus, whether or not a participant's answer was correct, and their interaction. We also included a random intercept effect of participant.

This model (Fig. 3.7) showed a small effect of visualization condition (Est = 1.62; 95% CI: [0.80, 2.43]; $t = 3.90$; $p < 0.001$) such that users report slightly higher confidence on average when using HOPs. There was also an interaction between visualization and condition order (Est = -1.33; 95% CI: [-2.48, -0.18]; $t = -2.28$; $p = 0.02$). Simple effects showed that users were more confident on average when they used HOPs in the second block than in any other combination of visualization or block order. More than practice or learning, visualization condition seemed to play a role in confidence reporting, although

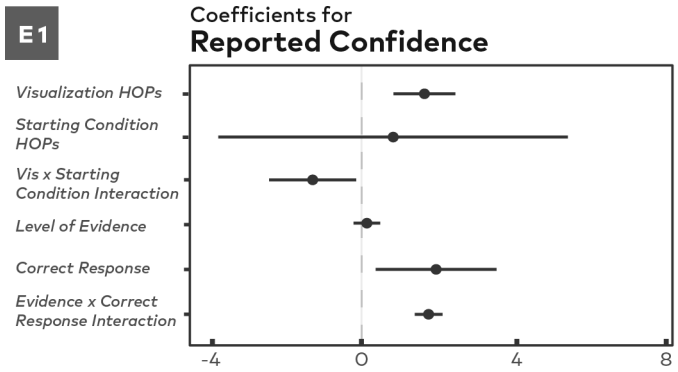


Figure 3.7: Estimated regression coefficients for our exploratory mixed-effects linear model of confidence reports in Experiment 1 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

these effects are of doubtful practical significance. Looking at trial-level predictors, we found that users were slightly more confident on average when they answered correctly (Est = 1.93; 95% CI: [0.33, 3.52]; $t = 2.37$; $p = 0.02$). We also found an interaction between correctness and the level of evidence presented in a stimulus (Est = 1.73; 95% CI: [1.36, 2.09]; $t = 9.25$; $p < 0.001$). Confidence increased with increasing evidence for trials where the user was correct, but confidence decreased with increasing evidence for trials where the user was incorrect. Sanders et al. [171] found that this interaction was predicted by the estimates of expected confidence produced by their confidence modeling. However, we failed to replicate this finding. The expected confidence ratings produced by our model of confidence fitness failed to predict reported confidence very well.[8]

3.4 Experiment 2

In experiment 1 we find that participants are more sensitive to the trend underlying samples from a noisy time series when using HOPs rather than error bars. However, in light of prior work demonstrating that people struggle to interpret error bars [16, 39, 86] and that bar encodings induce biased judgments of probability [148], this result is not entirely surprising. How might these results differ if we compare HOPs to a more effective static visualization than error bars and remove the influence of the within-the-bar bias from our

[8]See Supplemental Materials for more information on the confidence fitness analysis. `https://github.com/kalealex/jobs-report-hops`

visual encodings? Also, the differences between HOPs and error bars inferred from the results of experiment 1 conflate the effect of *animated vs static* encodings with the effect of *discrete sampling vs summary statistical* representations of probability. In experiment 2, our primary aim is to use the evaluation framework we developed to isolate the effect of animated vs static encodings of uncertainty. Thus, we use a similar study design to compare HOPs composed of lines (instead of bars) to static line ensembles (Fig. 3.2, middle right). We choose line ensembles as our control condition because they use a frequency representation of uncertainty which is superior to error bars, and they aggregate the discrete samples shown in HOPs into one static view.

3.4.1 Methods

With the exception of the following changes, the methods used in our second experiment were identical to the methods used in the first.

Uncertainty Visualizations

We tested three uncertainty visualization conditions in our second experiment. Two of these were HOPs with lines instead of bars (Fig. 3.2, right). We test two HOPs conditions to examine whether the benefits of HOPs persist at very high frame rates where the participant cannot process individual samples. The *fast HOPs* condition displayed each sample for only 100 ms in order to test the limits of visual ensemble processing. The *regular HOPs* condition displayed each sample for 400 ms, matching the presentation rate chosen by Hullman et al. [98]. Recall that HOPs in Experiment 1 displayed each frame for 1500 ms, following the visualization design of the NYT. In contrast, the faster frame rate of the regular HOPs condition in Experiment 2 reflects our study team's preferred animation parameters.

The static visualization condition in our second experiment was a *line ensemble* (Fig. 3.2, middle right). The line ensemble visualizations displayed the same samples as HOPs using the same line encoding, but lines were aggregated into one static view. We chose to use 50 lines per ensemble with representative sampling from each month in order to maintain the perception of a discrete representation of uncertainty. We did not match the lines per

ensemble to the lines shown across all frames of the HOPs because we cannot know which lines users attended to in the HOPs condition. By comparing HOPs to line ensembles we investigate the impact of animating vs aggregating discrete sampling-oriented representations of uncertainty.

Experimental design

We tested visualization conditions between-subjects, whereas visualization condition was a within-subjects comparison in Experiment 1. We changed to a between-subjects comparison in order to avoid ambiguity about which visualization condition was informing participants' mental representation of the two possible underlying trends. Just like the first experiment, each participant completed two blocks of 60 trials under the same sampling procedure as before. However, in Experiment 2 both blocks of trials were completed under one visualization condition. Thus, we sampled twice the number of trials per PF fit in the second experiment as we did in the first. We chose to sample 120 trials per PF in order to improve the quality of our JND and PF spread estimates and reduce the role of sampling error in our results.

Analysis approach

We used the same maximum likelihood estimation procedures as in Experiment 1. For statistical inferences on JNDs, PF spreads, and confidence fitness, we used a linear model with visualization condition as a predictor. For Experiment 2, we preregistered[9] a secondary analysis of confidence reporting on each trial, similar to the exploratory analysis in our first experiment. We used a mixed-effects linear model of reported confidence with fixed effects of visualization condition, level of evidence in the stimulus, whether or not the participant answered correctly, and the interaction between evidence and correctness. We also included a random effect for individual differences.

[9] `https://osf.io/gw4cj/register/5771ca429ad5a1020de2872e`

Figure 3.8: Estimated regression coefficients for mixed-effects linear models of each psychophysical response variable in Experiment 2 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

3.4.2 Results and discussion

Again, we report results separately for JNDs, PF spreads, and confidence fitness.

Just-noticeable differences (JNDs)

We found that JNDs were lower in the regular HOPs condition (Fig. 3.8, left) than in the line ensembles condition (Est = -1.22; 95% CI: [-2.18, -0.26]; $t = -2.52$; $p = 0.01$). This suggested that animating hypothetical outcomes across frames instead of aggregating them into one static view facilitated correct judgments of the underlying trend in samples conveying a lower level of evidence. JNDs were also lower for participants who used fast HOPs rather than line ensembles (Est = -0.74; 95% CI: [-1.69, 0.22]; $t = -1.52$; $p = 0.13$). However, this effect was small and unreliable. When the frame rate of HOPs was faster, gains in performance were attenuated. Most individual JND estimates were in approximately the same range regardless of visualization condition. The effects we found were largely driven by a subset of individuals, mostly in the line ensembles condition, who showed little sensitivity on the task (JNDs > 9 absolute log likelihood units). It seems that animating hypothetical outcomes at a particular frame rate leads to more consistent levels of sensitivity across observers in judging samples from a noisy time series.

Psychometric function spreads

In line with the results from our first experiment, we found that visualization condition did not seem to impact PF spreads. On average, spread estimates were not reliably different

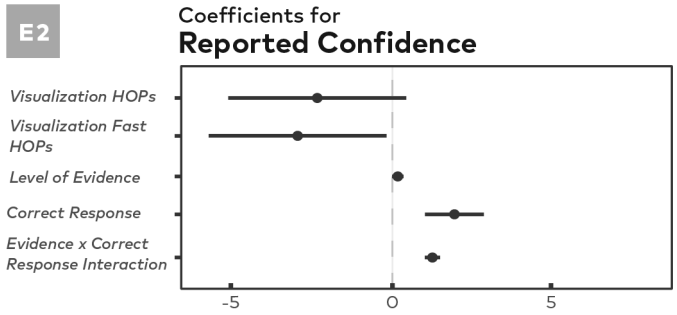


Figure 3.9: Estimated regression coefficients for our mixed-effects linear model of confidence reports in Experiment 2 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

for participants in the regular HOPs, fast HOPs, and line ensemble conditions (Fig. 3.8, middle).

Confidence fitness

Confidence fitness was not reliably different across visualization conditions (Fig. 3.9, right). We speculate that animation of hypothetical outcomes, compared to a static depiction of the same outcomes, may have little distorting effect on the efficacy of confidence monitoring.

We ran a mixed-effects linear model on confidence reporting data (Fig. 3.9) similar to the exploratory analysis in Experiment 1. On average participants reported lower confidence in the fast HOPs condition than in the line ensembles condition (Est = -2.94; 95% CI: [-5.70, -0.17]; $t = -2.07$; $p = 0.04$). Participants in the regular HOPs condition also reported lower confidence on average (Est = -2.34; 95% CI: [-5.09, 0.44]; $t = -1.64$; $p = 0.10$), although this effect was not reliable. These results were in contrast with those of Experiment 1, in which participants reported slightly *higher* confidence on average when using HOPs rather than error bars. This apparent reversal of effect may have occurred because line ensembles establish a higher baseline for confidence than error bars, but it could also be noise.

Our findings regarding trial-level predictors were consistent with those of the first experiment. We found that participants were more confident on trials where they were correct (Est = 1.96; 95% CI: [1.04, 2.88]; $t = 4.18$; $p < 0.001$) and that there was an interaction between correctness and the level of evidence presented (Est = 1.34; 95% CI: [1.15, 1.53]; $t = 13.77$; $p < 0.001$). We also found a minuscule effect of the level of evidence in a stim-

ulus on confidence (Est = 0.19; 95% CI: [0.01, 0.37]; $t = 2.04$; $p = 0.04$). Similar to our first experiment, our raw confidence data followed the predictions of Sanders et al. [171], who created the confidence fitness algorithm. Their model predicted that the relationship between confidence reporting and the level of evidence presented in a stimulus is modulated by correctness of participant responses. However, our implementation did not produce this predictive behavior (see Supplemental Materials).

3.5 Discussion and future work

In Experiment 1, we find that users are more sensitive to the underlying trend in samples from a noisy time series when using HOPs rather than error bars as a reference for what no growth versus growth in the jobs report could look like. This suggests that sampling-oriented presentations of uncertainty lead to better calibrated visual inferences about underlying trends in noisy data than summary statistical representations of uncertainty. This finding is consistent with the literature on frequency presentations of probability [69, 70, 85, 87], sampling-oriented uncertainty visualizations [56,97,98,115,132], and the ensemble processing abilities of the human visual system [2, 3, 81, 92, 129]. Our first experiment extends this line of inquiry by (1) applying vision science methodology to uncertainty visualization evaluation and (2) addressing the question of whether HOPs can improve sensitivity to uncertainty information in the public domain.

In Experiment 2, we isolate the effect of animation on user perceptions of probability. We compare HOPs, at two different frame rates, showing hypothetical sets of 12 outcomes as lines to static line ensembles, which show the same samples using the same line encoding but aggregated into one static display. We find that participants in the regular HOPs condition (400 ms per sample) are more consistently sensitive to the underlying trend in samples from a noisy time series than participants in the line ensembles condition. However, this effect is attenuated for fast HOPs (100 ms per sample). Perhaps, the ability of the visual system to automatically process ensembles breaks down beyond a certain frequency. Future work in information visualization and vision science should systematically test ensemble processing abilities across a more granular set of frame rates to try to pinpoint the limitations of the visual system in processing this kind of display.

Our analyses of confidence fitness suggest that the efficacy of participants' confidence monitoring was not systematically impacted by visualization conditions. Confidence fitness is a unitless latent parameter of a model describing the degree to which reported confidence matches predictions from a Monte Carlo simulation based on signal detection theory. As such, it is difficult to interpret the practical significance of these null results.[10]

3.5.1 Limitations

Confidence fitness. Confidence as a construct is difficult to interpret. Confidence fitness contextualizes confidence reporting data in comparison to a theoretical ground truth, quantifying the degree to which confidence reporting fits a particular statistical definition. In principle, this tells us whether reported levels of confidence are warranted based on the presented evidence and the perceptual sensitivity of the user. However, we find that the algorithm does not consistently predict reported confidence, and this suggests that there may be wrong assumptions in the model (see Supplemental Materials). Future work should investigate these assumptions in search of better models to establish a ground truth for confidence. Alternatively, perhaps a normative account of confidence, which assumes that confidence means the same thing to participants and researchers, is not appropriate considering the intersubjectivity of confidence reporting.

Psychometric functions. A typical approach when using psychometric functions (PFs, Fig. 3.5) to model forced-choices is to estimate JNDs and use statistical inference to detect shifts in the sensitivity of participants' perceptions under different conditions, quantified as differences in JNDs. In doing this, we fit a descriptive model but throw out all information other than a point estimate in the process of statistical inference.

Although, we can leverage other aspects of the PF for inference. We did this by examining often ignored PF spreads, which represent noise in the perception of evidence. Judging by the null effects of visualization on PF spreads in this study, one might think they are an insensitive measure of visualization effectiveness. However, the change in PF spreads over

[10] In retrospect, it's likely that this study was not designed with high statistical power to detect interface effects on confidence fitness, considering we had no *a priori* expectations about how these visualizations would impact confidence fitness as a construct.

time within an individual measures of how quickly people learn a task on a given interface. Since our experiments were not designed to measure such a learning effect, this remains a promising direction for future work. We also demonstrated how estimates of the PF spread can be leveraged, along with signal detection theory (Fig. 2.1), to bootstrap estimates of expected confidence reporting [171].

The core problem with the traditional use of JNDs is that PF fitting and statistical inference are relegated to separate computational models. In future work, we intend to rethink perceptual modeling using PFs and incorporate the entire procedure into a hierarchical Bayesian modeling approach [143].[11]

3.5.2 Design guidelines

Our findings imply that uncertainty visualizations in public-facing venues such as the news have a measurable impact on perceptions of uncertainty. Designers of uncertainty visualizations for public audiences have a responsibility to use uncertainty representations which promote accurate perceptions of probability.

HOPs and other sampling-oriented displays are particularly helpful when understanding the process that produced data is critical. In such cases, summary statistical encodings like error bars conceal important information about variability behind opaque statistical constructs like standard error [16, 39, 43, 86]. In contrast, showing discrete outcomes from the process which produced the data offers a more interpretable rendering of uncertainty [56, 69, 70, 87, 97, 98, 115], especially when viewers are unlikely to have statistical training.

The trade-off between static sampling-oriented visualizations and HOPs is mostly a question of how these techniques interact with the affordances of the visual system. In the case of static sampling-oriented visualizations, designers should consider whether displaying too many outcomes in one view will disrupt the perception of discrete outcomes, resulting in a density encoding rather than the intended frequency encoding. Density encodings for uncertainty are interpreted less consistently than frequency encodings because they describe probability in the abstract rather than showing users an experiential representation

[11] We take this hierarchical modeling approach in the analysis of incentivized decisions in Chapter 4.

of uncertainty [97, 98, 115]. Displaying too many outcomes in one view also might lead to the problem of crowding, the inability of the visual system to generate accurate percepts of cluttered scenes [3]. In the case of HOPs, designers should consider whether the audience will allocate the necessary attention to integrate outcomes over time. Additionally, the designer should choose a frame rate for HOPs which is not so fast that the viewer cannot process individual observations and not so slow that the viewer becomes impatient. Interestingly, we and colleagues in our lab independently chose a similar frame rate to the one tested in prior work on HOPs [98], 400 ms per sample. Future work should examine a range of possible parameters and empirically establish the optimal design for HOPs.

Although communicating uncertainty is recommended in order to emphasize that outcomes may not always resemble their expected value [62, 65], in some cases uncertainty visualizations may not be well-matched to the intended task. If uncertainty is unlikely to aid interpretation (e.g., when reading a point estimate), a direct encoding of central tendency might be preferred. Prior work shows that HOPs are inferior to point estimates with error bars or violin plots if the task is to estimate central tendency of highly variable outcomes [98]. In some scenarios involving decision making, communicating the uncertainty in the process which produced the data might introduce confusion, anxiety, or indecision [131]. However, failing to show uncertainty may lead to a false sense of security in visualized outcomes [5, 182]. Designers should consider the needs and background of their audience in order to determine how and whether or not to visualize uncertainty.

3.6 Summary

In this Chapter, I present a study demonstrating that animated, sampling-oriented uncertainty visualizations (a.k.a. hypotherical outcome plots, HOPs [98]) support better-calibrated visual inferences among untrained laypeople than more conventional, static uncertainty representations. The task scenario in this study was based on judgments people might make when encountering uncertainty visualizations in the news. It required users to integrate information about the central tendency and variance of distributions to make inferences about the likelihood of samples, a realistic and complex judgment incorporating statistical competencies implicated in many news articles presenting uncertainty informa-

tion. Specifically, performance on this task reflects the ability to infer the underlying data generating process (DGP) in noisy samples of data, a task that I try to support through the tool-building work presented in Chapter 6.

3.7 Acknowledgements

The roles of the first and second authors were approximately equal. Alex Kale contributed domain knowledge about psychophysics and experimental design and was responsible for implementing the data analyses. Francis Nguyen contributed domain knowledge about system development and graphical design and was responsible for implementing the experimental interface. Jessica Hullman and Matthew Kay advised the project, contributing expertise in statistical inference and uncertainty visualization. Alex wrote the paper with iterative feedback from Jessica and editing help from other authors. Francis created the figures for the paper with guidance from other authors. Alex and Jessica conducted the qualitative analysis of visualizations in the news. Alex, Jessica, and Francis designed the experimental procedure. All authors selected visualization conditions to evaluate.

We would like to thank Geoffrey Boynton, Steve Franconeri, Michael-Paul Schallmo, and the members of the Interactive Data Lab for their feedback throughout the study.

Chapter 4

VISUAL REASONING STRATEGIES FOR EFFECT SIZE JUDGMENTS AND DECISIONS

The second cognitive mechanism I highlight in my dissertation is actually a set of cognitive processes called *heuristics* or shortcuts for thinking whereby a person substitutes a simple judgment for a more complex or intractable one, often in a goal directed way. In this Chapter, I present an experiment where the study team[1] and I investigated how visualization design choices impact the heuristics that laypeople use to make incentivized decisions with charts. This work [109] was published at IEEE VIS, where it won the 2020 InfoVis Best Paper Award.

4.1 Heuristics in uncertainty visualization mis)interpretation

Many visualization authors perceive visualizing uncertainty as an exception, rather than a norm [94]. However, the common practice of omitting uncertainty information from visualizations and focusing attention on point estimates leads to "incredible certitude" [142], the unwarranted impression that error is minimal or not important. To enable informed judgments and decisions, experts commonly suggest showing uncertainty information alongside point estimates, e.g., by showing intervals in which estimates could fall [44, 45, 141, 192].

However, presenting uncertainty alongside point estimates may not lead users to incorporate uncertainty information into their judgments. A large body of work on biases due to heuristics (e.g., [105, 198, 200]) shows that people often avoid or discount uncertainty information. This suggests that chart users may ignore uncertainty in favor of means even when both are presented [98]. We might call this behavior *satisficing* [178]—opting for a suboptimal strategy when better strategies are available—to the extent that uncertainty information is essential for making the judgment at hand.

[1]The original study team for this work was me, Matthew Kay, and Jessica Hullman.

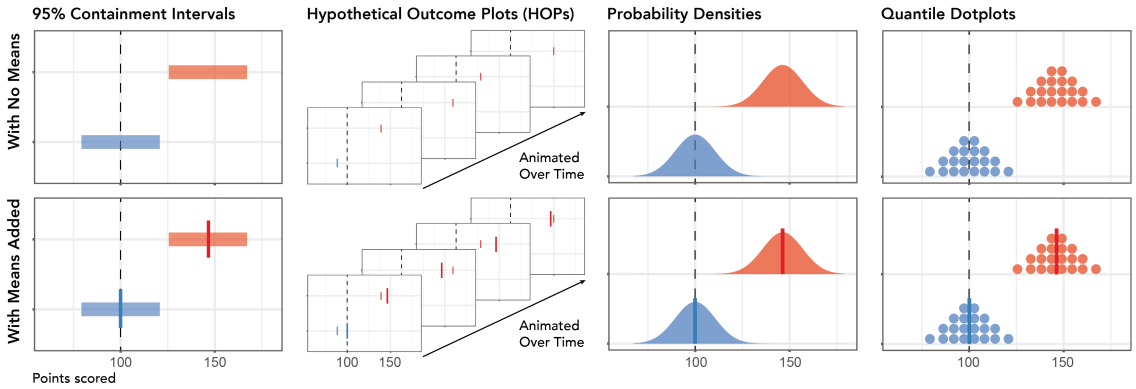


Figure 4.1: Visualization designs evaluated in our experiment.

Different visualization design choices make the mean more or less salient. Imagine a continuum of uncertainty visualization designs representing how perceptually difficult it is to decode the mean from a chart. At one extreme are hypothetical outcome plots or HOPs [98,110] where the mean is only encoded implicitly as the average of a set of outcomes presented across frames of an animation. At the other extreme are direct encodings of point estimates presented alongside uncertainty (e.g., represented as error bars). We expect that the salience of the mean in uncertainty visualization designs and other factors such as frequency-framing of probability [56, 98, 110, 115] influence the degree to which users focus on means and ignore uncertainty.

How might chart users who focus on means judge *effect size*, the effectiveness of an intervention such as a medical treatment or policy? Imagine a user viewing visualizations like those in Figure 4.1, showing distributions of outcomes predicted under two alternative courses of action. Discounting uncertainty may manifest as using *distance between means or gist estimates of distance between distributions as a proxy for effect size* and not judging distance relative to the width of distributions. Using only distance as a proxy for effect size may be misleading (Fig. 4.2) because the distance between distributions depends on a number of factors, including the variance of distributions and the visualization author's choice of axis scale as noted by previous work [40, 88, 219].

We investigate a scenario where distance heuristics lead to a predictable pattern of bias in order to measure how different visualization designs impact users' reliance on distance as a proxy for effect size. Users are shown charts depicting various effects on a fixed axis

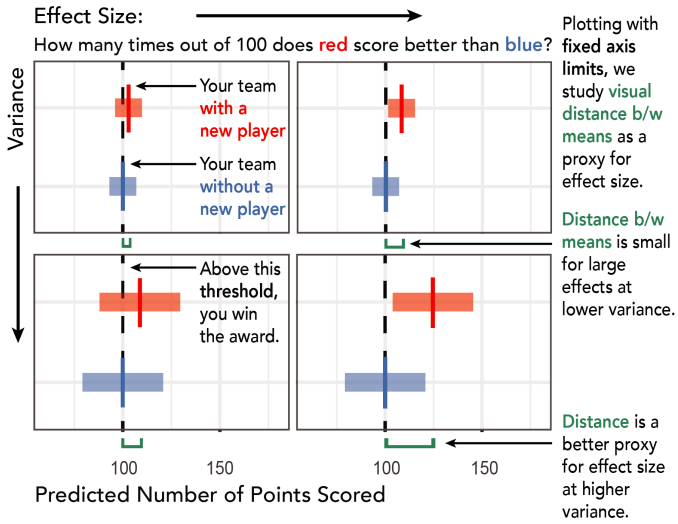


Figure 4.2: Intervals with means showing two levels of effect size (72% and 95% probability of superiority, $\Pr(S)$) at low and high variance. Using visual distance between means as a proxy for effect size should result in greater bias toward underestimating effect size at lower variance than at higher variance.

(Fig. 4.2) such that when distributions have lower variance, visual distance between means is small regardless of effect size, but distances correspond to effect size more consistently at higher variance. In this scenario, we expect that adding means to uncertainty visualizations leads users to underestimate effect size at lower variance. Conversely, adding means may reduce this underestimation bias at higher variance.

We contribute a pre-registered experiment on Mechanical Turk investigating how uncertainty visualization design impacts lay users' judgments and decisions from effect size. We find that *visualization designs which support magnitude estimation are not necessarily best suited as decision aids*. Quantile dotplots lead to the least bias in magnitude estimation, but other visualizations lead to the least bias in decision making, such as intervals without means when there was high variance. On a fixed axis scale, densities without means support unbiased decisions at lower variance, and users show substantial bias with all visualizations at higher variance. Visualization effectiveness for decision making depends on the level of variance in data relative to the axis scale. Adding means has a negligible impact on magnitude estimation, but in most cases it leads to less utility-optimal decisions.

In a qualitative analysis of users' strategy descriptions, we find that *few users apply*

the optimal strategy for reading an uncertainty visualization when one exists. Instead, the majority of users appear to *satisfice* [178] by using a small set of heuristics. We find that the *majority of users report relying on visual distance* between distributions regardless of uncertainty information, an observation that is consistent with the biases in our quantitative results. We also find that many users switch between strategies. This suggests that many uncertainty visualizations may not be interpreted in ways that researchers and designers expect, and characterizing possible strategies may lead to design recommendations based on how users reason in practice.

4.2 Method

We tested how adding means to different uncertainty visualizations impacts users' estimates and incentivized decisions from effect size.

4.2.1 Tasks & procedure

Our task was like a fantasy sports game. We showed participants charts comparing the predicted number of points scored by their team with and without adding a new player (e.g., Fig. 4.2). Participants estimated the effect size of adding the new player and decided whether or not to pay to add the new player to their team.

Effect size estimation. We asked participants to estimate a measure of effect size called *probability of superiority* or common language effect size [144]: *"How many times out of 100 do you estimate that your team would score more points with the new player than without the new player?"* We elicited probabilities as "times out of 100" based on literature in statistical reasoning (e.g., [69, 87]) suggesting that people reason more accurately with probabilities when they are framed as frequencies. Probability of superiority, the percent of the time that outcomes for one group A exceed outcomes for another group B, is a proxy for standardized mean difference $\frac{\mu_A - \mu_B}{\sigma_{A-B}}$ [37, 46], the difference between two group means relative to uncertainty in the estimates. Using synthetic data (see Section 4.2.5), we evaluated bias in effect size estimates compared to a known ground truth.

Intervention decisions. We also asked participants to make binary decisions indicating whether they would *"Pay for the new player,"* or *"Keep [their] team without the new player."*

On each trial, the participant's goal was to win an award worth \$3.17M, and they could pay \$1M to add a player to their team if they thought the new player improved their chances of winning enough to be worth the cost. There were four possible payouts in each trial:

1. The participant won without paying for a new player (+\$3.17M).
2. The participant paid for a new player and won (+\$2.17M).
3. They failed to win without paying for a new player (\$0).
4. The participant paid for a new player and failed to win (-\$1M).

The user could only lose money if they paid for the new player.[2] We set up the incentives for our task so that a risk-neutral chart user should pay for a new player only when effect size was larger than 74% probability of superiority or Cohen's d of 0.9, the average effect size in a recent survey of studies in experimental psychology [176]. This enabled us to evaluate intervention decisions compared to a utility-optimal standard.

Feedback. At the end of each trial we told users whether or not their team scored enough points to win an award, using a Monte Carlo simulation to generate a win or loss based on the participant's decision. We split feedback into two tables. One showed the change in account value for the current trial. The other showed cumulative account value and how this translated into a bonus in real money. By showing probabilistic outcomes, instead of the expected value of decisions, feedback gave participants a noisy signal of how well they were doing, mirroring real-world learning conditions for decisions under uncertainty.

Payment. Participants received a *guaranteed reward of \$1* plus a bonus of $\$0.08 \cdot (account - \$150M)$, where \$0.08 per \$1M was the exchange rate from account value to real dollars, *account* was the value of their fantasy sports account at the end of the experiment, and \$150M was a cutoff account value below which they receive no bonus. These values were carefully chosen to result in bonuses between \$0 and \$3, such that participants who guessed randomly and experienced unlucky probabilistic outcomes would receive no bonus, and participants who responded optimally would be guaranteed a bonus.

User strategies. To supplement our quantitative measures with qualitative descriptions

[2] In pilot studies, we tested how framing outcomes as winning versus losing awards impacted user behavior and found that participants had greater preference for intervention when it was described as increasing the certainty of gains, consistent with prior work by Tversky and Kahneman [106, 200].

of users' visual reasoning, at the end of each of the two block of trials, we asked users, *"How did you use the charts to complete the task? Please do your best to describe what sorts of visual properties you looked for and how you used them."*

4.2.2 Formalizing a class of decision problems

Our decision task represents a class of decision problems where one makes a *binary decision* about whether or not to invest in an intervention that changes the probability of an *all-or-nothing outcome*. For example, this class of problems includes medical decisions about treatments that may save someone's life or cure them of a disease, organizational decisions about hiring personnel to reach a contract deadline, and personal decisions such as paying for education to seek a promotion. Previous decision making literature examines similar problems in the context of salting the road in freezing weather [102, 103], voting in presidential elections [213], and willingness to pay for interventions in a fictional scenario [88]. The key similarity between these decision problems is that their incentive structures imply a *common utility function*.

A utility function defines optimal (i.e., utility maximizing [206]) decisions for a risk-neutral observer, providing a normative benchmark used to measure bias in decision making. Comparing behavior to a risk-neutral benchmark is a common practice in judgment and decision making studies [10], often used to measure risk preferences [209] or attitudes that make a person more or less inclined to take action than they should be based on a cost-benefit analysis. In the class of decision problems we investigate, the implied utility function depends on both the amount of money one stands to win or lose (e.g., the value of an award and the cost of a new player) and the effect size (e.g., the difference in team performance with versus without a new player).

Let v be the value of an award. Let c be the cost of adding a new player to the team. The utility-optimal decision rule is to intervene if

$$v \cdot \Pr(award|\neg player) < v \cdot \Pr(award|player) - c$$

where $\Pr(award|\neg player)$ is the probability of winning an award *without* a new player, and $\Pr(award|player)$ is the probability of winning an award *with* a new player. Assuming a

constant ratio between the value of the award and the cost of intervention $k = \frac{v}{c}$, we express the decision rule in terms of the difference between the probabilities of winning an award with versus without a new player:

$$\Pr(award|\neg player) + \frac{1}{k} < \Pr(award|player)$$

The threshold level of effect size above which one should intervene depends on the incentive ratio k and the probability of a payout without intervention $\Pr(award|\neg player)$. In our study, we fixed the incentives $k = 3.17$ and the probability of winning an award without a new player $\Pr(award|\neg player) = 0.5$ so that users would not have to keep track of changing incentives, and effect size alone was the signal that users should base decisions on.[3] This enabled a controlled evaluation of how users translate visualized effect size into a sense of utility. By modeling a functional relationship between effect size and utility, we go beyond prior work which either does not vary the effectiveness of interventions (e.g., [102, 103, 213]) or examines only two levels of effect size as a robustness check for statistical tests (e.g., [88]).

4.2.3 Experimental design

We assigned each user to one of four uncertainty visualization conditions at random, making comparisons of uncertainty visualizations between-subjects. On each trial, users made a probability of superiority estimate and an intervention decision. We asked users to make repeated judgments for two blocks of 16 trials each. In one block, we showed the users visualizations with means added, and in the other block there were no means. We counterbalanced the order of these blocks across participants. Each of the 16 trials in a block showed a unique combination of ground truth effect size (8 levels) and variance of distributions (2 levels), making our manipulations of ground truth, variance, and adding means all within-subjects. The order of trials in each block was randomized. In the middle of each block, we inserted an attention check trial, later used to filter participants who did not attend to the task. Users always saw an attention check at 50% probability of superiority

[3] In pilot studies, we tried manipulating k and $\Pr(award|\neg player)$ and found that these changes had little impact on the effectiveness of different uncertainty visualizations for supporting utility-optimal decision making. In light of prior work showing that Mechanical Turk workers do not respond to changes of incentives [188], we suspect that these manipulations might have an impact in real-world settings which is difficult to measure on crowdsourcing platforms.

with means and at 99.9% without means. Hence, each participant completed 17 trials per block and 34 trials total.

4.2.4 Uncertainty visualization conditions

We evaluated visualizations intended to span a design space characterized by the visual salience of the mean, expressiveness of uncertainty representation, and discrete versus continuous encodings of probability. As described above, we showed four uncertainty visualization formats—intervals, hypothetical outcome plots (HOPs), density plots, and quantile dotplots—with and without separate (i.e., extrinsic) vertical lines encoding the mean of each distribution. We expected that adding means would bias effect size estimates toward discounting uncertainty and that this effect would be most pronounced for uncertainty visualizations in which the mean is *not* intrinsically salient.

Intervals. We showed users intervals representing a range containing 95% of the possible outcomes (Fig. 4.1, left column). In the absence of a separate mark for the mean, the mean was not intrinsically encoded, and the user could only find the mean by estimating the midpoint of the interval. Intervals were not very expressive of probability density since they only encoded lower and upper bounds on a distribution.

Hypothetical outcome plots (HOPs). We showed users animated sequences of strips representing 20 quantiles sampled from a distribution of possible outcomes (Fig. 4.1, left center column), matching the data shown in quantile dotplots. Animations were rendered at *2.5 frames per second with no animated transitions* (i.e., no tweening or fading) between frames, looping every 8 seconds. We shuffled the two distributions of 20 quantiles using a 2-dimensional quasi-random Sobol sequence [181] to minimize the apparent correlation between distributions. Like intervals, HOPs did not make the mean intrinsically salient, as means were implicitly encoded as the average position of an ensemble of strips shown over time. However, HOPs were more expressive of the underlying distribution than intervals and expressed uncertainty as frequencies of events, so they conveyed an experience-based sense of probability.

Densities. We showed users continuous probability densities where the height of the

area marking encoded the probabilities of corresponding possible outcomes on the x-axis (Fig. 4.1, right center column). Unlike intervals and HOPs, the mean was explicitly represented as the point of maximum mark height because distributions were symmetrical, so means were intrinsically salient. Densities were also the most expressive of the underlying probability density function among the uncertainty visualizations we tested.

Quantile dotplots. We showed users dotplots where each of 20 dots represented a 5% chance of a corresponding possible outcome on the x-axis (Fig. 4.1, right column). Like densities, because distributions were symmetrical and dots were stacked in bins to express this symmetry, the mean was explicitly represented as the point of maximum height and was thus intrinsically salient.

4.2.5 Generating stimuli

We generated synthetic data covering a range of effect size, so there were an equal number of trials where users should and should not intervene. Recall that 50% corresponded to a new player who did not improve the team's performance at all, 100% corresponded to a definite improvement in performance, and 74% was the utility-optimal decision threshold. We sampled eight distinct levels of ground truth probability of superiority, four values between 55% and 74% and four values between 74% and 95%, such that there are an equal number of trials above and below the utility-optimal decision threshold. Prior work in perceptual psychology [71, 228] suggests that the brain represents probability on a log odds scale. For this reason, we converted probabilities into log odds units and sampled on this logit-transformed scale using linear interpolation between the endpoints of the two ranges described above. We added two attention checks at probabilities of superiority of 50% and 99.9%, where the decision task should have been very easy, to allow for excluding participants who were not paying attention.

To derive the visualized distributions from ground truth effect size, we made a set of assumptions. We assumed *equal and independent variances* for the distributions with and without a new player σ^2_{team} such that $\sigma^2_{diff} = 2\sigma^2_{team}$ where σ^2_{diff} was the variance of the difference between distributions. We tested two levels of variance, setting the standard

deviation of the difference between distributions σ_{diff} to a low value of 5 or a high value of 15. These levels produced distributions that looked relatively narrow or wide compared to the width of the chart, making visual distance between distributions an unreliable cue for effect size such that at low variance large effect sizes corresponded to distributions that looked close together.

We determined the distance between distributions, or mean difference μ_{diff}, using the formula $\mu_{diff} = d \cdot \sigma_{diff}$ where d were ground truth values as standardized mean differences (i.e., Cohen's d [37,46]). The mean number of points scored without the new player was held constant $\mu_{without} = 100$, which corresponded to a 50% chance of winning the award. We calculated mean for the team with a new player $\mu_{with} = \mu_{without} + \mu_{diff}$. We rendered our chart stimuli using the parameters μ_{with}, $\mu_{without}$, and σ_{team} to define the two distributions on each chart. Holding the chance of winning without a new player constant at 50% (Fig. 4.2, blue distributions) is an experimental control that enables us to compare a user's preference for new players across trials using a coin flip gamble as the alternative choice, which is common in judgment and decision making studies [10].

4.2.6 Modeling

We wanted to measure how much users underestimate effect size in their probability of superiority responses, how much they deviate from a utility-optimal criterion in their decisions, and how sensitive they are to effect size for the purpose of decision making. To measure underestimation bias, we fit a linear in log odds model [71, 228] to probability of superiority responses, and we derive *slopes* describing users' responses as a function of the ground truth (Fig. 4.3). To measure bias and sensitivity to effect size in decision making, we fit a logistic regression to intervention decisions, and we derive *points of subjective equality* and *just-noticeable differences* describing the location and scale of the logistic curve as functions of effect size (Fig. 4.4).

Approach

We used the brms package [29] in R to build Bayesian hierarchical models for each response variable: probability of superiority estimates and decisions of whether or not to intervene. We started with simple models and gradually added predictors, checking the predictions of each model against the empirical distribution of the data. This process of *model expansion* [61] enabled us to understand the more complex models in terms of how they differ from simpler ones.

We started with a *minimal model*, which had the minimum set of predictors required to answer our research questions, and built toward a *maximal model*, which included all the variables we manipulated in our experiment. We specified the minimal and maximal models for each response variable in our preregistration.[4]

Expanding models gradually helped us determine priors one-at-a-time. Each time we added a new kind of predictor to the model (e.g., a random intercept per participant), we honed in on weakly informative priors using prior predictive checks [61]. We centered the prior for each parameter on a value that reflected no bias in responses. We scaled each prior to avoid predicting impossible responses and to impose enough regularization to avoid issues with convergence in model fitting. We documented priors and model expansion in Supplemental Materials.[5]

[4] `https://osf.io/9kpmb`

[5] `https://github.com/kalealex/effect-size-jdm`

Linear in log odds model

We use the following model (Wilkinson-Pinheiro-Bates notation [29, 163, 217]) for responses in the probability of superiority estimation task:

$$\begin{aligned}
\text{logit}(response_{\Pr(S)}) \sim &\text{Normal}(\mu, \sigma) \\
\mu = &\text{logit}(true_{\Pr(S)}) * means * var * vis * order \\
&+\text{logit}(true_{\Pr(S)}) * vis * trial \\
&+\big(\text{logit}(true_{\Pr(S)}) * trial + means * var \big| worker\big) \\
\log(\sigma) = &\text{logit}(true_{\Pr(S)}) * vis * trial \\
&+means * order \\
&+\big(\text{logit}(true_{\Pr(S)}) + trial \big| worker\big)
\end{aligned}$$

Where $response_{\Pr(S)}$ is the user's probability of superiority response, $true_{\Pr(S)}$ is the ground truth probability of superiority, *trial* is an index of trial order, *means* is an indicator for whether or not extrinsic means are present, *var* is an indicator for low versus high variance, *vis* is a dummy variable for uncertainty visualization condition, *order* is an indicator for block order, and *worker* is a unique identifier for each participant used to model random effects. Note that there are submodels for the mean μ and standard deviation σ of user responses.

Motivation. We apply a logit-transformation to both $response_{\Pr(S)}$ and $true_{\Pr(S)}$, changing units from probabilities of superiority into log odds, because prior work suggests that the perception of probability should be modeled as linear in log odds (LLO) [71, 228]. We model effects on both μ and σ because we noticed in pilot studies that the spread of the empirical distribution of responses varies as a function of the ground truth, visualization design, and trial order. However, we are most interested in effects on mean response. The term $\text{logit}(true_{\Pr(S)}) * means * var * vis * order$ tells our model that the slope of the LLO model varies as a joint function of whether or not means were added, the level of variance, uncertainty visualization, and block order (i.e., all of these factors interacted with each other). This enables us to answer our core research questions, while controlling for order effects. The term $\text{logit}(true_{\Pr(S)}) * vis * trial$ models learning effects, so we isolate the impact

of uncertainty visualizations. In both submodels, we added within-subjects manipulations as random effects predictors as much as possible without compromising model convergence.

Logistic regression

We use this model to make inferences about intervention decisions:

$$\begin{aligned} intervene \sim & \text{Bernoulli}(p) \\ \text{logit}(p) = & evidence * means * var * vis * order \\ & + evidence * vis * trial \\ & + \big(evidence * means * var + evidence * trial \big| worker\big) \end{aligned}$$

Where *intervene* is the user's choice of whether or not to intervene, p is the probability that they intervene, and *evidence* is a logit-transformation of the utility-optimal decision rule (see Section 4.2.2):

$$evidence = \text{logit}(\Pr(award|player)) - \text{logit}(\Pr(award|\neg player) + \frac{1}{k})$$

This gives us a uniformly sampled scale of evidence where zero represents the utility-optimal decision threshold. All other factors are the same as in the LLO model (see Section 4.2.6.2).

Motivation. We logit-transform our evidence scale because internal representations of probabilities are thought to be on a log odds scale [71, 228], such that linear changes in log odds appear similar in magnitude. The term $evidence * means * var * vis * order$ tells our model that the location and scale of the logistic curve vary as a joint function of whether or not means were added, the level of variance, uncertainty visualization, and block order. Mirroring an analogous term in the LLO model, this enables us to answer our core research questions, while controlling for order effects. The term $evidence * vis * trial$ models learning effects. As with the LLO model, we specify random effects per participant through model expansion by trying to incorporate as many within-subjects manipulations as possible.

4.2.7 Derived measures

From our models, we derive estimates for three preregistered metrics that we use to compare visualization designs.

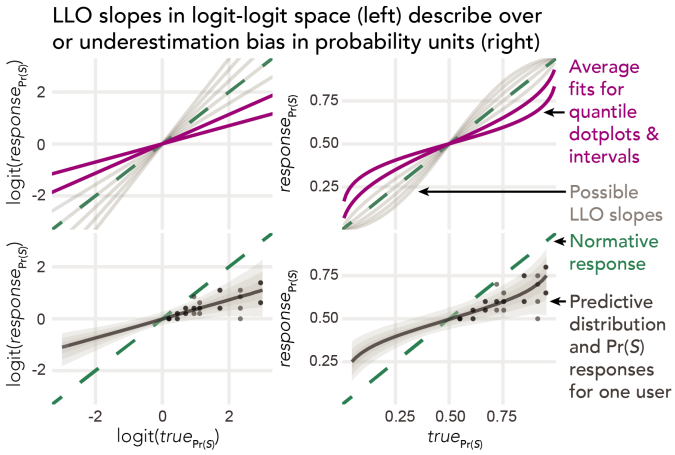


Figure 4.3: Linear in log odds (LLO) model: fits for average user of quantile dotplots and intervals compared to a range of possible slopes (top); predictive distribution and observed responses for one user (bottom).

Linear in log odds (LLO) slopes measure the degree of bias in probability of superiority $\Pr(S)$ estimation (Fig. 4.3). A slope of one indicates unbiased performance, and slopes less than one indicate the degree to which users underestimate effect size.[6] We measure LLO slopes because they are very sensitive to the expected pattern of bias in responses, giving us greater statistical power than simpler measures like accuracy. Specifically, LLO slope is the expected increase in a user's logit-transformed probability of superiority estimate, $\text{logit}(response_{\Pr(S)})$, for one unit of increase in logit-transformed ground truth, $\text{logit}(true_{\Pr(S)})$. Using a linear metric (i.e., slope in logit-logit space) to describe an exponential response function in probability units comes from a theory that the brain represents probabilities on a log odds scale [71, 228]. The LLO model [71, 228] can be thought of as a generalization of the cyclical power model [91] that allows a varying intercept or a modification of Stevens' power law [185] for proportions.

Points of subjective equality (PSEs) measure bias toward or against choosing to intervene in the decision task relative to a utility-optimal and risk-neutral decision rule (see Section 4.2.2). PSEs describe the level of evidence at which a user is expected to intervene 50% of the time (Fig. 4.4). A PSE of zero is utility-optimal, whereas a negative value indicates that a user intervenes when there is not enough evidence, and a positive value

[6] LLO slopes less than one represent bias toward the probability at the intercept, $\text{logit}^{-1}(intercept)$, which is close to $\Pr(S) = 0.5$ in our study.

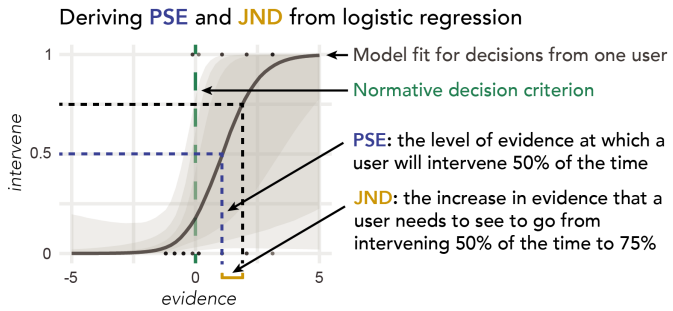


Figure 4.4: Logistic regression fit for one user. We derive point of subjective equality (PSE) and just-noticeable difference (JND) by working backwards from probabilities of intervention to levels of evidence.

indicates that a user doesn't intervene until there is more than enough evidence. In our model, PSE is $\frac{-intercept}{slope}$ where *slope* and *intercept* come from the linear model in logistic regression.

Just noticeable differences (JNDs) measure sensitivity to effect size information for the purpose of decision making (Fig. 4.4). They describe how much additional evidence for the effectiveness of an intervention a user needs to see in order to increase their rate of intervening from 50% to about 75%. A JND in evidence units is a difference in the log probability of winning the award with the new player. We chose this scale for statistical inference because units of log stimulus intensity are thought to be approximately perceptually uniform [185, 204]. In our model, JND is $\frac{\text{logit}(0.75)}{slope}$ where *slope* is the same as for PSE.

4.2.8 Participants

We recruited users through Mechanical Turk. Workers were located in the US and had a HIT acceptance rate of 97% or more. Based on the reliability of inferences from pilot data, we aimed to recruit 640 participants, 160 per uncertainty visualization. We calculated this target sample size by assuming that variance in posterior parameter estimates would shrink by a factor of roughly $\frac{1}{\sqrt{n}}$ if we collected a larger data set using the same interface. Since we based our target sample size on between-subjects effects (e.g., uncertainty visualization), our estimates of within-subjects effects (e.g., adding means) were precise.

We recruited 879 participants. After our preregistered exclusion criterion that users needed to pass both attention checks, we slightly exceeded our target sample size with 643

total participants. However, we had issues fitting our model for an additional 21 participants, 17 of whom responded with only one or two levels of probability of superiority and 4 of whom had missing data. After these non-preregistered exclusions, our final sample size was 622 (with block order counterbalanced). All participants were paid regardless of exclusions, on average receiving $2.24 and taking 16 minutes to complete the experiment.

4.2.9 Qualitative analysis of strategies

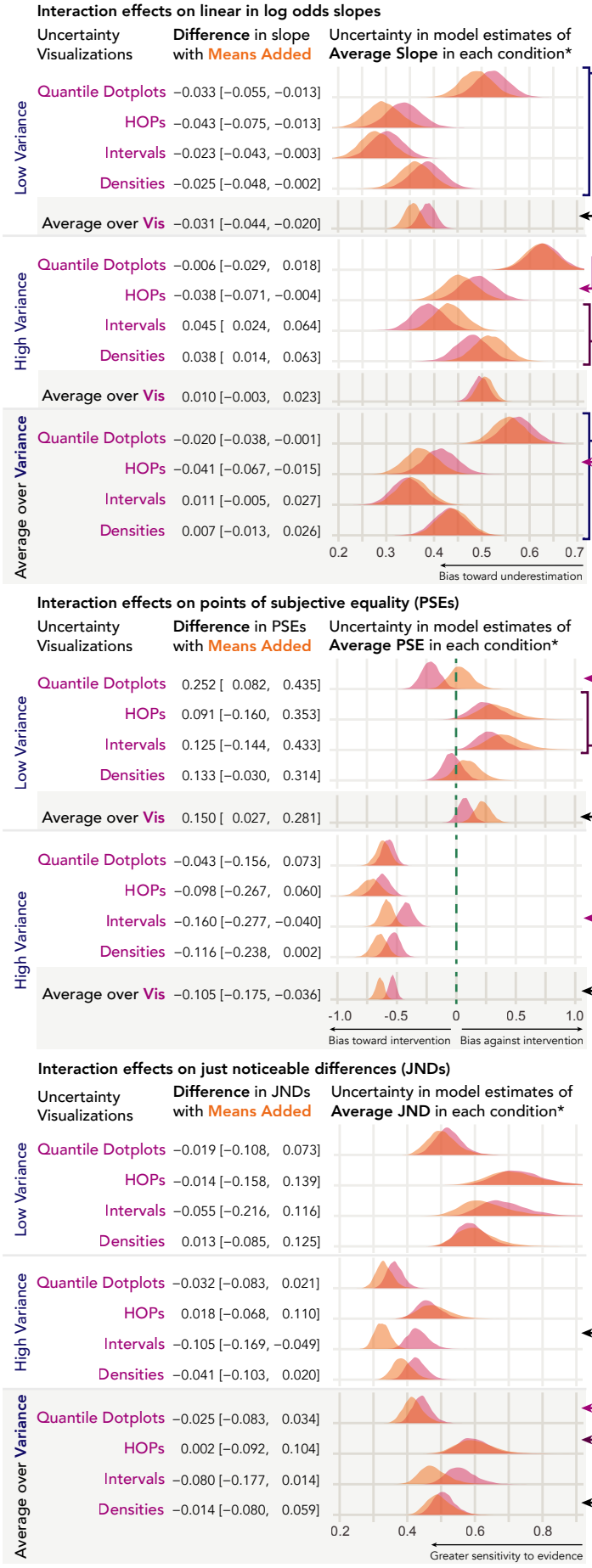
Using the two strategy responses we elicited from each user, we conducted a qualitative analysis to characterize the visual reasoning strategies users relied on with different visualization designs (with and without means) and whether they switched strategies.

I developed a bottom-up open coding scheme for how users described their reasoning with the charts. Since some responses were uninformative about what visual properties of the chart a user considered (e.g., *"I used the charts to estimate the value added by the new player."*), we excluded 180 participants for whom both responses were uninformative from further analysis, resulting in a final sample of 442 for our qualitative analysis.

We used our open codes to develop a classification scheme for strategies based on what visual features of charts users mentioned, whether they switched strategies, and whether they were confused by the chart or task. We coded for the following uses of visual features:

- **Relative position** of distributions
- **Means**, whether users relied on or ignored them
- **Spread** of distributions, whether users relied on variance, ignored it, or erroneously preferred high or low variance
- **Reference lines**, whether users relied on imagined or real vertical lines (e.g., the annotated decision threshold in Fig. 4.1 & 4.2)
- **Area**, whether users relied on the spatial extent of geometries
- **Frequencies**, whether users of quantile dotplots or HOPs relied on frequencies of dots or animated draws

We generated a spreadsheet of quotes, open codes, and categorical distinctions which enabled us to describe patterns and heterogeneity in user strategies.



4 Results

4.1 Probability of superiority judgments

For each uncertainty visualization, **adding means** at **low variance** decreases LLO slopes. Recall that a slope of one corresponds to no bias, and a slope less than one indicates underestimation. When we **average over uncertainty visualizations**, **adding means** at **low variance** reduces LLO slopes for the average user, indicating a very small 0.8 percentage points increase in probability estimation error.

At **high variance**, the effect of **adding means** changes directions for different uncertainty visualizations. **Adding means** decreases LLO slopes for **HOPs**, whereas **adding means** increases LLO slopes for **intervals and densities**. Because differences in LLO slopes represent changes in the exponent of a power law relationship, these slope differences of similar magnitude indicate a very small increase in probability of superiority estimation error of 0.3 percentage points for HOPs and small reductions in error of about 1.5 and 1.0 percentage points for intervals and densities, respectively.

Users of all uncertainty visualizations underestimate effect size. When we **average over variance**, users show an average estimation error of 8.6, 14.0, 14.8, and 12.4 percentage points in probability of superiority units for quantile dotplots, HOPs, intervals, and densities, respectively, each **without means**. In this marginalization, **adding means** only has a reliable impact on LLO slopes for **HOPs**, but the difference is practically negligible.

4.2 Intervention decisions

4.2.1 Points of subjective equality

For each uncertainty visualization, **adding means** at **low variance** increases PSEs. This results in different effects depending on whether the visualization with **no means** has a PSE below or above utility-optimal. Recall that a PSE of zero is utility-optimal, a negative PSE indicates intervening too often, and a positive PSE indicates not intervening often enough. Users of **quantile dotplots** with **no means** have negative PSEs which become unbiased when we **add means**. Users of **HOPs and intervals** with **no means** have positive PSEs, biases which increase when we **add means**. Users of **densities** with **no means** have PSEs near zero and become more biased when we **add means**. Only the effect for quantile dotplots is reliable. When we **average over uncertainty visualizations**, at **low variance** the average user may have a PSE 0.6 percentage points above utility-optimal with **no means**, and **adding means** increases this mild bias by about 1.7 percentage points in terms of the probability of winning.

At **high variance**, **adding means** decreases PSEs. Since PSEs for all uncertainty visualizations with **no means** are below optimal, **adding means** increases biases in all conditions, however, the effect is only reliable for **intervals**. When we **average over uncertainty visualizations**, at **high variance** the average user has a negative PSE 9.5 percentage points below utility-optimal with **no means**, and **adding means** increases this bias by about 2.1 percentage points.

4.2.2 Just noticeable differences

At **low and high variance**, the effects of **adding means** on JNDs are mostly unreliable. Recall that smaller JNDs indicate that a user is sensitive to smaller differences in effect size for the purpose of decision making. **Adding means** only has a reliable effect on JNDs for **intervals** at **high variance**, where it reduces JNDs by 1.2 percentage points in terms of the probability of winning.

When we **average over variance**, **quantile dotplots with means** lead to the smallest JNDs, and users of **HOPs with** or **without means** have the largest JNDs, a difference of about 1 percentage point in terms of the probability of winning. Quantile dotplots **with** or **without means** have reliably smaller JNDs than other conditions, with the exception of unreliable differences between quantile dotplots with **no means** and **densities with** or **without means**.

*Probability densities of model estimates show posterior distributions of means conditional on the average participant.

4.3.3 Discussion

Among the uncertainty visualizations we tested, quantile dotplots lead to the least biased probability of superiority estimates. This is not surprising given previous work (e.g., [69,87, 98,110,115]) showing that frequency-based visualizations are effective at conveying probabilities. However, it is surprising that users do not perform reliably differently with frequency-based HOPs than with intervals or densities. HOPs directly encode probability of superiority by how often the draws from the two distributions change order, whereas in all other conditions users would need to calculate effect size analytically from visualized means and variances to arrive at the "correct" inference, although we doubt that users engage in such explicit mathematical reasoning. In Section 5, we present descriptive evidence of heuristics that users employ with different visualization designs, which helps to explain these results.

In most cases, the small effects on LLO slopes when adding means to uncertainty visualizations are probably negligible. However, they are consistent with the pattern of behavior we expect if users rely on visual distance between distributions as a proxy for effect size. When variance is lower relative the axis scale, distances between distributions look small even for large effects (Fig. 4.2, top), and users tend to underestimate effect size *more* when means are added. When variance is higher relative the axis scale, distances between distributions roughly correspond to effect size (Fig. 4.2, bottom), and users tend to underestimate effect size *less* when means are added, at least for densities and intervals.

Our results suggest that the best visualization design for utility-optimal decision making probably depends on the level of variance relative to the axis scale. At lower variance, when multiple levels of variance are shown on a common scale, densities without means or quantile dotplots with means lead to the least bias in decisions. At higher variance, users are biased toward intervening in all conditions, and both densities without means and intervals without means lead to the least bias. The impact of means also depends on variance and axis scaling, such that when we average across uncertainty visualizations, adding means exacerbates biases that exist when means are absent. The effect of variance on PSEs (see Supplemental Materials) is large, such that users intervene more often at higher variance than at lower variance. One possible explanation for this is that users rely on distance

between distributions as a proxy for effect size and make decisions as if effects are larger when distributions are further apart (Fig. 4.2).

Reported effects of visualization design on JNDs may not be practically important. All differences in JNDs between visualization designs are smaller than the difference between high versus low variance (see Supplemental Material). Smaller JNDs at high variance may reflect the fact that our high variance charts use white space more efficiently.

4.3.4 Comparing magnitude estimation & decision making

Different visualization designs lead to the best performance on our magnitude estimation and decision making tasks. To explore this decoupling of performance across tasks, we calculate average posterior estimates of our derived measures—LLO slope, PSE, and JND—for each individual user and compare them. Figure 4.5 shows that many individuals who are poor at magnitude estimation (i.e., LLO slopes below one) do well on the decision task (i.e., PSEs and JNDs near zero).

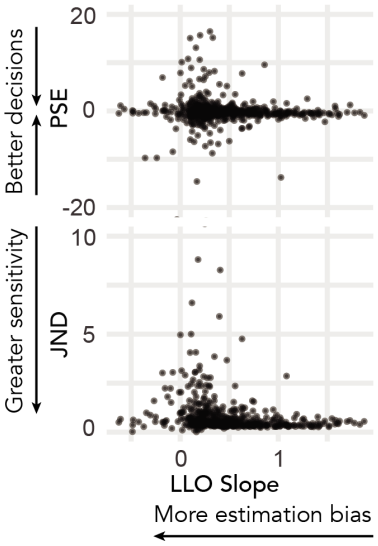


Figure 4.5: PSEs and JNDs vs LLO slopes per user.

One possible explanation for this decoupling of performance on our two tasks is that users may rely on different heuristics to judge the same data for different purposes. This is consistent with Kahneman and Tversky's [106] distinction between *perceiving* the probability of an event to be p and *weighting* the probability of an event in decision making as $\pi(p)$, which suggests that decision weights reflect preferences based on probabilities and risk attitudes [209]. Recent work in behavioral economics [118] suggests that biases in decision making are partially attributable to imprecision in an individual's subjective perception of numbers (i.e., "number sense"). Since JNDs reflect the precision of perceived

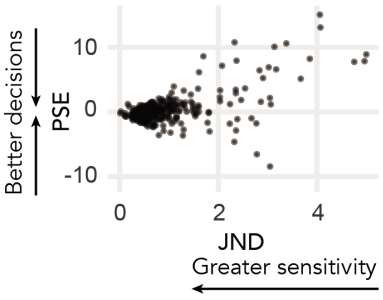


Figure 4.6: JNDs vs PSEs.

effect size implied by one's decisions and PSEs represent bias in decision making, we can investigate this relationship within individual users in our study (Fig 4.6). In agreement with prior work, we see that greater sensitivity to effect size for decision making (i.e., JNDs close to zero) predicts more utility-optimal decisions (i.e., PSEs close to zero). Although, based on the decoupling of LLO slopes and JNDs, it also seems clear that a user's internal sense of effect size is not necessarily identical when they use the same information for different tasks. We should be mindful that perceptual accuracy may not feed forward directly into decision making.

4.4 Visual reasoning strategies

We use qualitative analysis of reported strategies to identify ways that users judge effect size by comparing distributions, giving us a vocabulary for how visualization design choices impact their interpretations.

4.4.1 Prevalent strategies

The strategies we identify are not mutually exclusive. We count a user as employing a strategy if they mention it in either of their responses.

Only distance. About 62% of users (275 of 442) rely on "how far to the right" the red distribution is compared to the blue one *without mentioning that they incorporate the variance of distributions into their judgments* (Fig. 4.2). Roughly 69% of these users (190 of 275) describe making a gist estimate of distance between distributions, with 46% (126 of 275) saying they rely on the mean difference specifically, and 13% (36 of 275) saying they rely on both gist distance and mean difference. Strategies which involve only the distance between distributions should result in a large bias toward underestimating effect size, which is what we see in our aggregated quantitative results.

Distance relative to variance. Only about 8% of users (35 of 442) mention that their interpretations of distance depend on the spread of distributions, suggesting that perhaps very few untrained users are sensitive to the impact of variance on effect size. If users estimate standard deviation and mean difference between distributions, they could use this information to calculate effect size analytically. However, we think it is far more likely

that these users judge the distance between distributions relative to the spatial extent of uncertainty visualizations, which should result in underestimation bias which is similar to but less pronounced than with judgments of *only distance*.

Cumulative probability. A substantial 36% of users (160 of 442) estimate the cumulative probability of winning the award with and/or without the new player. This strategy involves judging the distance, proportion of area, or frequency of markings across the threshold number of points to win (e.g., Fig. 4.7). These users may be confusing cumulative probability of winning the award, which is the best cue in the decision task, with probability of superiority (i.e., probability that team does better with the new player than without), which is what we ask for in the estimation task. However, since the probability of winning increases monotonically with probability of superiority, this strategy should theoretically result in milder underestimation bias than distance-based strategies.

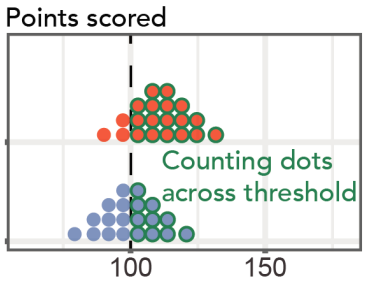


Figure 4.7: Cumulative probability strategy with quantile dotplots.

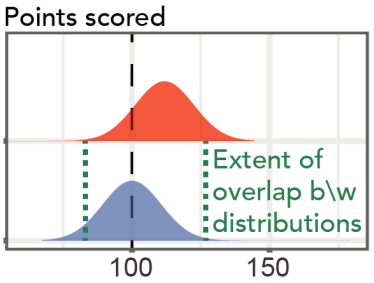


Figure 4.8: Overlap strategy with densities.

Distribution overlap. About 7% of users (31 of 442) describe judging the overlap between distributions. While similar to distance-based strategies, users conceptualize this strategy in terms of area rather than the gap between distributions (Fig. 4.8). For example, one user said they use HOPs *"only to see how much of an overlap [there is] between the two areas,"* suggesting that they imagine contours of distributions over the sets of animated draws. This strategy probably results in underestimation bias similar to judging *distance relative to variance*.

Frequency of draws changing order. This strategy is only relevant to the HOPs condition, where only about 16% of users (19 of 121) employed it. It involves judging the number of animated frames in which the draws from the two distributions switch order (Fig. 4.9). This is the best way to estimate probability of superiority from HOPs [98]. If we think of the user as accumulating information across frames, the precision of their inference is

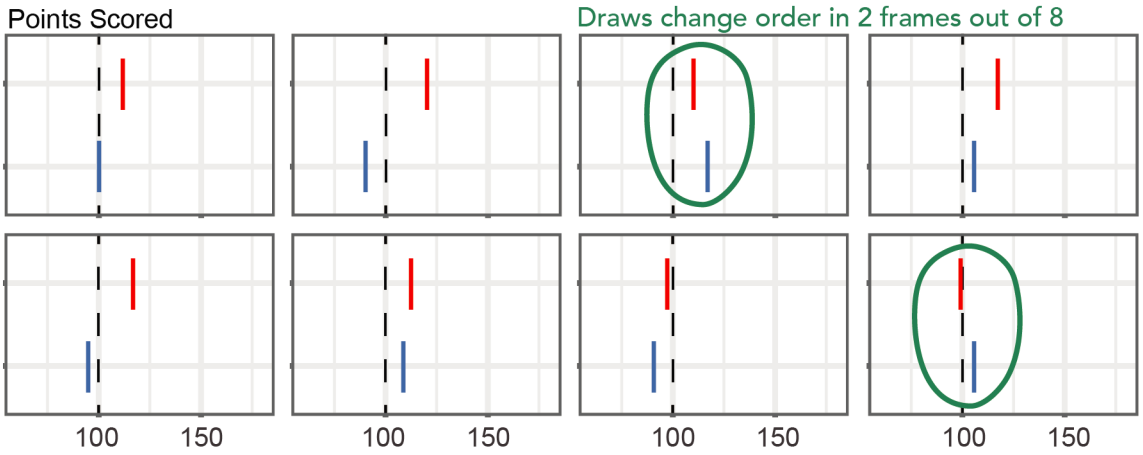


Figure 4.9: Frequency of draws changing order strategy with HOPs.

mostly limited by the number of frames they watch. For example, in Figure 4.9 red scores higher than blue in 6 of the 8 frames, and watching only 8 frames limits the precision of this inference to increments of $\frac{1}{8}$. The fact that only a handful of HOPs users employ this strategy helps to explain why the performance of HOPs users is worse than expected.

Switching strategies. A substantial 29% of users (129 of 442) switch between strategies in the middle of the task. For example, one user of intervals without means described a mix of *cumulative probability* and *distribution overlap* strategies: *"If the red [distribution] was completely past the dotted line then I would buy the new player no matter what. If there were overlaps with blue I would just risk assess to see if it was worth it to me or not."* While more of a meta-strategy, our observation that a significant proportion of users switch is important because it suggests that judgment processes involved in graphical perception may not be consistent within each user.

4.4.2 Impacts of visualization design choices

Users rely on visual features (Section 4.2.9) and strategies (Section 4.4.1) to varying degrees depending on visualization design (Table 4.1).

Intervals. Roughly 75% of intervals users (85 of 112) rely on *relative position* as a visual cue for effect size compared to 69% with densities (68 of 99), 61% with HOPs (74 of 121), and 59% with quantile dotplots (65 of 110). Of intervals users who look at *relative position*,

Table 4.1: Frequency of strategies used per uncertainty visualization.

Strategy	Intervals	HOPs	Densities	Dotplots	Overall
Distance	73	77	61	64	275
Rel. to Var.	11	9	10	5	35
Cumulative	34	50	30	46	160
Overlap	17	2	9	3	31
Draw Order	0	19	0	0	19
Switching	35	48	23	23	129
Total	112	121	99	110	442

about 87% (74 of 85) employ an *only distance* strategy, while only about 13% (11 of 85) judge *distance relative to variance* . In other words, only about 10% of intervals users (11 of 112) incorporate variance into their judgments of distance. About 28% of intervals users (31 of 112) report looking at *area*, with about 55% of these users (17 of 31) employing a *distribution overlap* strategy.

HOPs. About 61% of HOPs users (74 of 121) look at *relative position* to judge effect size. Of HOPs users who rely on *relative position*, merely 3% (2 of 74) use a *distance relative to variance* strategy. However, looking at *relative position* is not mutually exclusive with looking at *frequency* of draws, which 45% of HOPs users (54 of 121) rely on as a visual feature. Among HOPs users who rely on frequencies, about 69% (37 of 54) employ a *cumulative probability* strategy, while about 35% (19 of 54) rely on the optimal strategy of counting the *frequency of draws changing order*. Roughly 40% of HOPs users (48 of 121) mention *switching strategies* compared to 31% with intervals (35 of 112), 23% with densities (23 of 99), and 21% with quantile dotplots (23 of 110). Among HOPs users who switch strategies, about 81% (39 of 48) rely on the mean as a cue. Strategy switching involves the mean for about 30% of HOPs users who rely on *relative position* (22 of 74) compared to 43% of HOPs users who rely on *frequency* (23 of 54). That most HOPs users rely on *relative position*, and that those who do rely on *frequency* are more likely to switch to or from relying on the mean, helps to explain poor performance with HOPs.

Densities. About 69% of densities users (68 of 99) rely on *relative position* as a visual cue. Of densities users who look at *relative position*, only about 13% (9 of 68) employ a *distance relative to variance* strategy. As one might expect, a substantial 36% of densities users (36

of 99) rely on *area* as a cue, compared to 10% of quantile dotplots users (11 of 110). Among densities users who rely on *area*, about 53% (19 of 36) employ a *cumulative probability* strategy, while about 28% (10 of 36) employ a *distribution overlap* strategy. Interestingly, about 27% of densities users (27 of 99) mention relying on the *spread* of distributions as a cue, more than the 21% of users with intervals (24 of 112), 21% with HOPs (25 of 121), and 10% with quantile dotplots (11 of 110) who report relying on the same cue.

Quantile dotplots. Roughly 59% of quantile dotplots users (65 of 110) describe looking at *relative position* to judge effect size, similar to 61% of users with HOPs (74 of 121) and less than the 69% of densities users (68 of 99) and 76% of intervals users (85 of 112) who report using the same cue. Merely 6% of quantile dotplots users who rely on *relative position* (4 of 65) employ a *distance relative to variance* strategy. 37% of quantile dotplots users (41 of 110) rely on *frequency* as a visual cue by counting dots. About 81% of quantile dotplots users who rely on *frequency* (33 of 41) employ a *cumulative probability* strategy.

Adding means. A substantial 35% of users (155 of 442) describe relying on the mean as a cue for effect size. If we split users based on whether or not they start the task with means, about 31% of users (67 of 218) *switch strategies* when means are added to the charts halfway through the task, compared to 10% (23 of 224) who switch strategies when means are removed. This asymmetry in strategy switching suggests that means are "sticky" as a cue: Among the 15% of users (67 of 442) who start with and rely on means, about 66% (44 of 67) attempt to visually estimate means *after means are removed from charts*, almost twice as many as the 34% (23 of 67) who switch to relying on other cues. However, the impact of adding means on performance depends on what other strategies a user is switching between. Among the 20% of users (90 of 442) who rely on means and switch strategies, about 44% (40 of 90) just incorporate the mean into judgments of *relative position* without relying on other visual cues. Other groups of users switch between relying on means and less similar visual cues, with 34% (31 of 90) also mentioning *frequency* and 12% (11 of 90) mentioning *area*. That many users switch between relying on *relative position* and means, and that strategies are heterogeneous, helps to explain why the average impact of means on performance is small in our results.

4.5 General discussion

Our results suggest that *design guidelines for visualizing effect size* should depend on the user's task, the variance of distributions, and design choices about axis scales. To provide concrete design guidelines while acknowledging the inherent complexity of our results, we present high-level take-aways for designers alongside relevant caveats.

Quantile dotplots support the most perceptually accurate distributional comparisons, at least among the visualization designs we tested. **Caveat:** Asking users to perform two tasks may have led users to rely on relatively simple strategies like *cumulative probability* more than strategies which require more mental energy like *frequency of draws changing order*. Conditions of high cognitive load seem to favor uncertainty visualizations like quantile dotplots over HOPs.

Densities without means seem to support the best decision making across levels of variance. On a fixed axis scale, densities without means and quantile dotplots with means perform best at lower variance, while densities without means and intervals without means perform best at higher variance. No visualization design we tested eliminated bias in decision making at higher variance. **Caveats:** The visualization design that leads to the least bias in decision making depends on the variance of distributions relative to axis scale. Future work should investigate bias in decision making over a gradient of variances shown on a common scale, including charts with heterogeneous variances, as this would enable more exhaustive design recommendations.

Adding means leads to small biases in magnitude estimation and decision making from distributional comparisons, leading users to underestimate effect size and make less utility-optimal decisions in most in most cases we tested. **Caveats:** Although the biasing effects of means are mostly negligible, our estimates of these biases are probably very conservative for two reasons: (1) added means were only highly salient in the HOPs condition; and (2) in the absence of added means, users already tend to rely on *relative position*, a cue which the mean merely reinforces. The effects of adding means on decision quality reverse at high versus low variance, so these biases may disappear for specific combinations of variance and axis scale.

Users rely on distance between distributions as a proxy for effect size, so designers should note when this will be misleading and encourage more optimal strategies. Our quantitative analysis shows that adding means induces small but reliable biases in magnitude estimation, consistent with distance-based heuristics. Our qualitative analysis of strategies verifies that the majority of users (357 of 442; 80.8%) rely on distance between distributions or mean difference to judge effect size. **Caveats:** Subtle design choices probably impact the tendency to rely on distance heuristics versus other strategies. For example, including a decision threshold annotation on our charts (Fig. 4.2) may have encouraged users to judge effect size as *cumulative probability*, rather than probability of superiority, contributing to underestimation bias.

4.5.1 Limitations

We only tested symmetrical distributions, and this may limit the generalizability of our inferences. Although we speculate that chart users may rely on central tendency regardless of the family of a distribution, reasoning with multi-modal distributions in particular may involve different strategies not accounted for in the present study.

Because we rely on self-reported strategies in our qualitative analysis, our findings only reflect *conscious* strategies. This leaves out implicit or automatic information processing such as visual adaptation [107] and ensemble processing [110, 190], except in rare cases where users report trying to *"roughly average"* predictions presented as HOPs. Our choice to incentivize the decision making task but not magnitude estimation may have contributed to the decoupling of performance on our two tasks. We cannot disentangle this possible explanation from evidence corroborating Kahneman and Tversky's [106] distinction between perceived probabilities and decision weights (see Section 4.3.4).

We control the incentives for our decision task rather than manipulating them, in part because it is not feasible to test dramatically different incentives on Mechanical Turk. As such the risk preferences that we measure as PSEs are representative of users optimizing small monetary bonuses, and they may not capture how people respond to visualized data in crisis situations when lives, careers, or millions of dollars are at stake. However, by devising

a task that is representative of a broad class of decision problems (see Section 4.2.2), we make our results as broadly applicable as possible. We speculate that the *relative impacts* of visualization designs on risk preferences should generalize to decision problems with similar utility functions.

4.5.2 Satisficing & heterogeneity

The visual reasoning strategies that chart users rely on when making judgments from uncertainty visualizations may not be what visualization designers expect. We present evidence that, in the absence of training, users satisfice by using suboptimal heuristics to decode the signal from a chart. We also find that not all users rely on the same strategies and that many users switch between strategies. Satisficing and heterogeneity in heuristics make it difficult both to anticipate how people will read charts and to study the impact of design choices. Conventionally, visualization research has characterized visualization effectiveness by ranking visualization designs based on the performance of the average user (e.g., [36]). However, in cases like the present study where users are heterogeneous in their strategies, these *averages may not account for the experience of very many users* and are probably an oversimplification. Visualization researchers should be mindful of satisficing and heterogeneity in users' visual reasoning strategies, attempt to model these strategies, and try to design ways of training users to employ more optimal strategies.

4.5.3 Toward better models of visualization effectiveness

Because some users seem to adopt suboptimal strategies or switch between strategies when presented with an uncertainty visualization, models of visualization effectiveness which codify design knowledge and drive automated visualization recommendation and authoring systems should represent these strategies. We envision a new class of behavioral models for visualization research which attempt to enumerate possible strategies, such as those we identify in our qualitative analysis, and learn how often users employ them to perform a specific task when presented with a particular visualization design. Previous work [100] demonstrates a related approach by calculating expected responses based on a set of alter-

native perceptual proxies for visual comparison and comparing these expectations to users' actual responses. Like the present study, this work describes the correspondence between expected patterns and user behavior. Instead, we propose incorporating functions representing predefined strategies into predictive models which estimate the proportion of users employing a given strategy.

In a pilot study, we attempted to build such a model: a Bayesian mixture model of alternative strategy functions. However, because multiple strategies predict similar patterns of responses, we were not able to fit the model due to problems with identifiability. This suggests that the kind of model we propose will only be feasible if we design experiments such that alternative strategies predict sufficiently different patterns of responses. The approach of looking at the agreement between proxies and human behavior [100] suffers the same limitation, but there is no analogous mechanism to identifiability in Bayesian models to act as a fail-safe against unwarranted inferences. Future work should continue pursuing this kind of strategy-aware behavioral modeling.

We want to emphasize that the proposed modeling approach is not strictly quantitative, as the definition of strategy functions requires a descriptive understanding of users' visual reasoning. As such this approach offers a way to formalize the insights of qualitative analysis and represent the gamut of possible user behaviors inside of visualization recommendation and authoring systems.

4.6 Summary

We contribute findings from a mixed design experiment on Mechanical Turk investigating how visualization design impacts judgments and decisions from effect size. Our results suggest that visualization designs which support the least biased estimation of effect size do not necessarily support the best decision making. We discuss how a user's sense of the signal in a chart may not necessarily be identical when they use the same information for different tasks. We also find that adding means to uncertainty visualizations induces small but reliable biases consistent with users relying on visual distance between distributions as a proxy for effect size. In a qualitative analysis of users' visual reasoning strategies, we find that many users switch strategies and do not employ an optimal strategy when one

exists. We discuss ways that canonical characterizations of graphical perception in terms of average performance gloss over possible heterogeneity in user behavior, and we propose opportunities to build strategy-aware models of visualization effectiveness which could be used to formalize design knowledge in visualization recommendation and authoring systems beyond context-agnostic rankings of chart types.

4.7 Acknowledgements

Alex Kale led this study and contributed all the primary research artifacts (e.g., the experiment interface, statistical analysis, qualitative analysis, figures, and written report). Jessica Hullman and Matthew Kay gave many rounds of feedback on experimental design, statistical modeling, figures, and writing. This project involved five pilot studies, and iterating through these was a huge investment for each of us involved. The initial idea to investigate the visual distance heuristic by manipulating variance and controlling axis scale was suggested by Jessica as a follow-up to her 2015 study on HOPs [98].

We thank the members of the UW IDL and Vis-Cog Lab, as well as the MU Collective at Northwestern for their feedback. This work was supported by a grant from the Department of the Navy (N17A-T004).

Chapter 5

CAUSAL SUPPORT: MODELING CAUSAL INFERENCES WITH VISUALIZATIONS

The third cognitive mechanism I highlight in my dissertation is *model-based thinking* or reasoning about what we experience in terms of mental models that could explain that experience. For example, when exploring a dataset in order to make sense of its underlying causal relationships, a person engaged in model-based thinking would compare what they see in the data to what they would expect under different possible data generating processes—i.e., compare observed patters to counterfactual predictions. In this Chapter, I present a pair of experiments where the study team[1] and I investigated what makes it difficult to make causal inferences about data generating process from a visualized dataset. This work [111] was published at IEEE VIS, where it won an Honorable Mention Award.

5.1 Challenges of visual causal inference

Data analysts engaged in sensemaking [31, 78, 164, 170] infer the compatibility of different causal explanations with their data. During exploratory analysis or provisional statistical modeling, analysts implicitly or explicitly compare their data to patterns they expect as logical consequences of hypothesized data generating processes [8, 66, 67, 94, 96]. Visualizations play a critical role in causal inference both because externalization reduces cognitive load [43] and because human capabilities for inference rely heavily on sensory expectations (e.g., mental imagery) and comparisons between expectations and experiences [25, 107].

Data analysts and software designers need to anticipate how human capabilities for causal inference may be error prone. For instance, perceptual biases such as underestimation of sample size (e.g., [119]) contribute to errors in causal inferences insofar as perceived associations seem to be the basis for causal inferences [4, 223]. Analysts also err in their

[1]The original study team for this work was me, Yifan Wu, and Jessica Hullman.

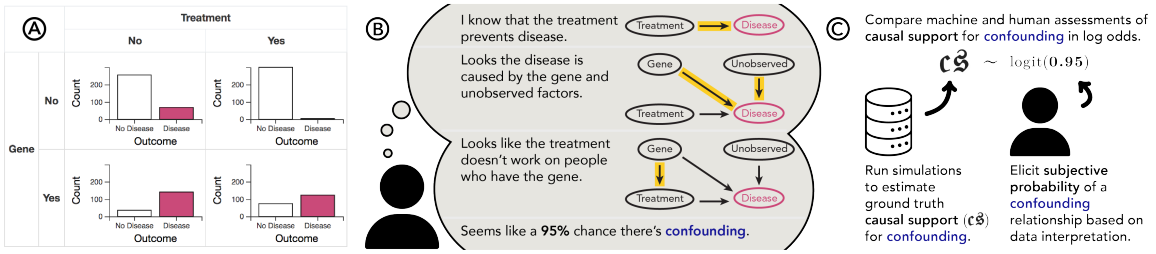


Figure 5.1: Modeling causal inferences with visualizations: Ⓐ Users view and may interact with data visualizations; Ⓑ Ideally, users reason through a series of comparisons that allow them to allocate subjective probabilities to possible data generating processes; and Ⓒ We elicit users' subjective probabilities as a Dirichlet distribution across possible causal explanations and compare these causal inferences to a computed benchmark of causal support, which we derive from Bayesian inference across possible causal models.

causal interpretations of data when the mapping between a potential causal explanation and an expected pattern in the data is unclear [12, 224]. For example, imagine an analyst trying to detect confounding in experiment results on the effectiveness of a treatment at preventing disease (Fig. 5.1 Ⓐ). To detect that 'gene' is a confounding factor, the analyst must see effects of gene on both treatment effectiveness (i.e., a difference between the top and bottom cells in the right column of table Ⓐ) and overall rate of disease (i.e., a difference between the top and bottom rows of Ⓐ). Attributing these signals in the data to confounding requires the analyst to know what they are looking for, rather than passively detecting the appropriate patterns.

To assess how well visual analytics (VA) tools support such causal inferences, we need to compare analysts' inferences to a benchmark that is roughly 'normative', that captures important aspects of good causal inference. We adopt causal support, a model from mathematical psychology [77], as a benchmark for evaluating causal inferences from visualizations. *Causal support* models one's belief in a set of possible data generating processes as a Bayesian update. Causal support is a good normative model for three reasons. (1) Causal support has numerous proprieties of valid statistical inference in light of the analyst's prior knowledge. It captures the fact that belief in evidence should be stronger as sample size increases. It accounts for unknown unknowns about the space of possible models, such that causal support assigns no posterior probability to models that the analyst does not explicitly consider. (2) Prior work [77] shows through a system of experiments that causal

support accounts for otherwise hard-to-explain patterns in human causal inferences, e.g., that subjective belief in a causal relationship varies as a function of the potential to detect that relationship in a given data set. (3) Since causal support is *extensible to any generative model* (i.e., models that can assign likelihood to data), it can be applied in a wide range of VA applications to evaluate causal inferences.

We contribute two experiments using causal inference problems involving count and proportion data to (1) study the utility of causal support for gaining insight into inferences from visualizations and (2) evaluate how well visualizations common in VA software support causal inferences about possible data generating processes. We compare three common visual encodings for count and proportion data—bar charts, icon arrays, and text tables as a baseline—and we investigate how the ability to interactively aggregate or cross-filter data in bar charts impacts causal inferences. In Experiment 1, we ask participants to differentiate whether a treatment is effective by allocating probability across two different data generating processes. We find that chart users' causal inferences are far from normative with all visualizations we tested, and interacting with visualizations can improve sensitivity to signal in predictable ways. Ultimately, however, no visualization reliably outperforms text tables. We also see that chart users are more sensitive to evidence against a treatment effect than evidence in favor of one, suggesting an unequal weighting of falsifying versus verifying evidence. In Experiment 2, we replicate the main findings Experiment 1 but with a confounding detection task where participants allocate probability across four alternative data generating processes. Experiment 2 demonstrates how casual support can be extended to study causal inferences about more complex data generating processes.

5.2 Experiment 1

How well do different visualization designs that are common in visual analytics (VA) software support causal inferences about possible data generating processes? We evaluate visualizations of count data including text contingency tables, icon arrays, grouped bar charts, bars that users can interactively aggregate, and linked bars that users can interactively cross-filter. In Experiment 1, we investigate chart users' ability to infer whether a treatment prevents a disease. By asking chart users about a treatment effect in count data, we

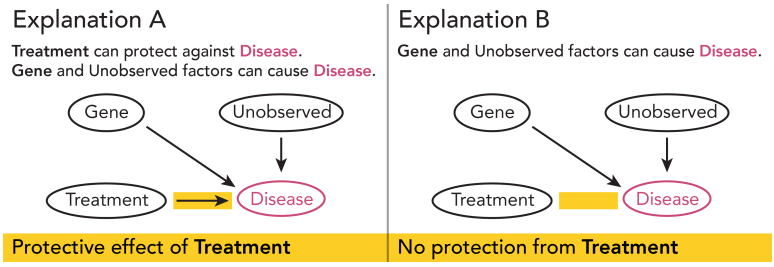


Figure 5.2: DAGs representing possible causal explanations participants were asked to consider in Experiment 1.

build on the task and structural equation models used by Griffiths and Tenenbaum [77] to propose and validate causal support. Count data are also ideal for evaluating bar charts. The design requirements for supporting our causal inference task with count data are that visualizations should express both the *proportion* of people with disease and *sample size*. Based on these requirements, we ruled out testing pie charts and heatmaps, common ways of encoding proportions that do not encode sample size.

5.2.1 Method

We set out to study how well different visualizations support causal reasoning by using *causal support* as a benchmark for causal inferences.

Task scenario & response elicitation

Participants played the role of an analyst hired by a company to interpret samples of data on the effectiveness of experimental treatments at preventing various diseases. We showed participants visualizations of the number of people in each sample who did or did not receive *treatment*, get a *disease*, and have a *gene* known to cause the disease.

We asked participants to judge the underlying causal relationships in the data, rating their degree of belief that treatment protects against disease by allocating probability across the two DAGs in Figure 5.2. We chose to study causal inferences about treatments, genes, and diseases in order to create a scenario where users would find the possible causal explanations feasible, coherent, and memorable.

Question & elicitation. We asked participants the following question:

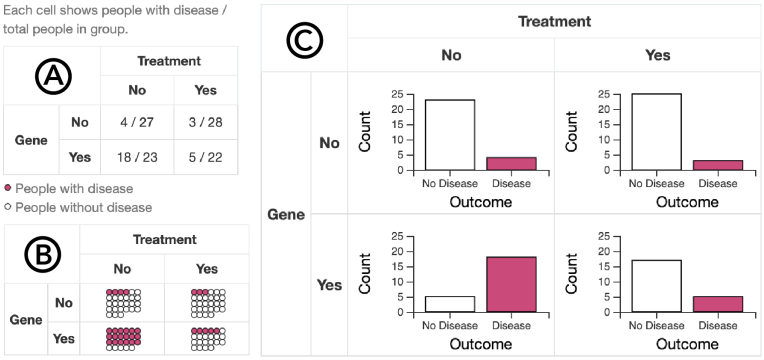


Figure 5.3: Non-interactive visualizations evaluated in our study: Ⓐ text contingency tables; Ⓑ faceted icon arrays; and Ⓒ faceted bar charts.

> ***How much do you believe in each of the causal explanations described below?*** *Imagine you have 100 votes to allocate across the two possible explanations. Split your 100 votes between explanations based on your degree of belief. For example, if you think one explanation is twice as likely as the other, you might give 67 votes (roughly two thirds) to that explanation and 33 votes (roughly one third) to the other. Assume no other explanations are possible.*

Participants responded with two complementary probabilities. We used form validation to make sure their responses were both numbers between 0 and 100 that summed to 100. Following prior work on eliciting Dirichlet distributions [33, 150] (i.e., probabilities allocated across alternatives), when participants gave their first response, we imputed what the second response would need to be in order for their responses to sum to 100. This imputed value and a corresponding prompt, *"Adjust your responses until both numbers reflect your beliefs."*, were both highlighted with the same color to indicate this imputation. We elicited probabilities as "votes out of 100" because *frequency framing* tends to reduce bias in probability estimates [69, 87, 150]. Participants received no feedback on their responses. We transformed these responses into perceived causal support, which we compared to our benchmark.

Perceived causal support. The dependent variable in our study was a measure of the perceived log odds of a *target explanation* over other possible causal explanations. Specifically, in Experiment 1 we targeted explanation A, which posited a treatment effect, requiring us

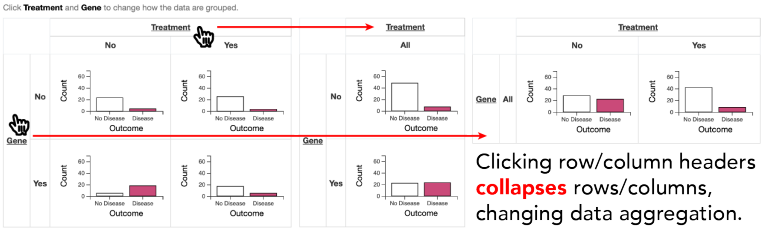


Figure 5.4: Aggregating bars mimic shelf construction and faceting.

to transform participants' responses into a log response ratio (lrr_A),

$$lrr_A = \log\left(\frac{response_A}{response_B}\right)$$

where $response_A$ and $response_B$ were the probabilities participants allocated to causal explanations A and B, respectively, on each trial. We used a log odds scale in order to make participants' perceived causal support comparable to our normative benchmark of causal support.

Payment. Participants received a *guaranteed reward of $2* plus a bonus of $0.25 for every trial where their estimate of the probability of causal explanation A[2] was within 5 percentage points of the ground truth.

Apparatus. We collected data using a Flask application deployed on Heroku with a Firebase database and visualizations created with D3 [22].[3]

Visualization conditions

Our visualizations show the number of people with and without disease in each cell of a 2x2 contingency table faceted by treatment and gene, with the exception of cross-filter bars which use a different layout. We aimed to test visualizations of count data similar to what an analyst could produce using visual analytics software like Tableau.

Text contingency tables showed the number of people with disease as a fraction of the total number of people in each cell of a faceted table (Fig. 5.3 Ⓐ). Text tables, which have been studied in prior research on causal support, served as a baseline comparison for other

[2]Bonuses in Experiment 2 were based on the probability of explanation D.

[3]E.g., see Experiment 2 interface at `https://bit.ly/3rDcxfn`

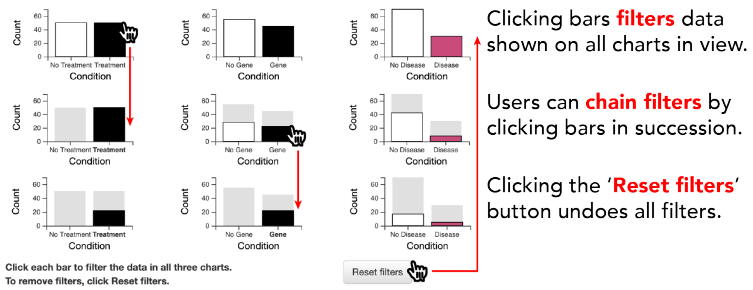


Figure 5.5: Cross-filtering bars mimic coordinated multiple views.

visualizations.

Icon arrays showed counts of people with and without disease as filled and open circles, respectively (Fig. 5.3 Ⓑ). We set the number of dot columns to minimize the aspect ratio on each trial, similar to how analysts might create roughly square icon arrays in Tableau. Icon arrays express both proportion and sample size as natural frequencies, which prior work finds beneficial for statistical reasoning (e.g., [63, 69, 87]).

Bar charts showed counts of people with and without disease using a length/position encoding on a common scale (Fig. 5.3 Ⓒ). On each trial, we set the y-axis scale to the maximum count of the data in view, allowing scales to change from trial to trial as they do when users load a new data set in Tableau. Bar charts are ubiquitous for count data.

Aggregating bars (aggbars) were similar to bar charts, however, users could interactively toggle faceting by treatment or gene by clicking on the table headers (Fig. 5.4). On each trial, we set the y-axis scale to the maximum count of the fully aggregated data. We designed aggbars to roughly mirror Tableau's shelf interactions, where users control faceting by direct manipulation of table headers. Interactive faceting may facilitate causal inferences by enabling users to explore whether "collapsing" [75] over a factor changes patterns in the data.

Cross-filtering bars were three bar charts showing the number of people with and without treatment, the gene, and disease, respectively (Fig. 5.5). We linked these bar charts such that clicking on one bar cross-filtered the rest of the data in view. When users applied a filter (e.g., only show people who received treatment), the corresponding axis label became bold and gray bars persisted in the background, so chart users could compare filtered and unfiltered views without relying on working memory. Users "reset filters" by

clicking a button below the charts. Filterbars emulated coordinated multiple views such as a Tableau dashboard. Interactive cross-filtering might assist in causal inferences because conditioning the data in view based on a specific event is analogous to Pearl's *do operator* (e.g., do(*treatment* = yes)), a notation for reasoning about counterfactuals in a causal networks [160].

Experimental design

We manipulated both the visualizations participants used for the task and the data sets that we showed. We randomly assigned each participant to use one of five visualizations (see Section 5.2.1.2), making comparisions of visualizations between-subjects. We showed each participant a total of 18 trials, which included 16 data conditions (see Section 5.2.1.4) presented within-subjects and two attention checks which we used for exclusion criteria (see Section 5.2.1.6). We randomized trial order for each participant, inserting attention checks on trials 7 and 13.

Stimulus generation

We evaluated causal inferences on realistic data sets, which spanned a range of ground truth causal support. Generating data sets required (1) manipulating data attributes which signaled causal support to participants, and (2) labeling each data set with ground truth causal support.

Our goal was to generate 16 data conditions (i.e., trials in our experiment) that varied delta p and sample size, two data attributes which in turn manipulate ground truth causal support. *Delta p* described the difference in the proportion of people with disease in each data set depending on whether they received treatment. Positive values of delta p indicated evidence that the treatment protected against disease; negative values indicated evidence against treatment effectiveness. *Sample size* was the number of people in each data set we showed participants.

Data conditions. To generate our 16 data conditions, we simulated data from structural models with one parameter per DAG arrow in Figure 5.2. We manipulated both the

probability that treatment prevents disease (4 levels: $\{0, 0.1, 0.2, 0.4\}$) and *sample size* (4 levels: $\{100, 500, 1000, 1500\}$). We controlled the probability of disease due to the gene (0.5), probability of disease due to unobserved causes (0.2), base rate of the gene (0.4), and the proportion of each sample with treatment (0.5). We selected these parameters iteratively by sampling data sets and labeling ground truth until half of the trials had greater than a 50% chance of being generated by causal explanation A.

For each of the resulting 16 data conditions, we simulated many data sets using a binomial random number generator to approximate realistic sampling error. By simulating sampling error, we prevented the count data from appearing contrived. This sampling error resulted in a distribution of ground truth causal support under each data condition, with more variability in the ground truth at smaller sample sizes. To guarantee that each participant saw trials spanning a consistent range of causal support, we selected 16 data sets representing 16 quantiles of the ground truth distribution per data condition, and we counterbalanced the quantile shown for each data condition across participants within each visualization condition using a balanced latin square. For our attention check trials, we selected the two simulated data sets that had the minimum and maximum ground truth causal support.

Labeling ground truth causal support for each data set. We operationalized the ground truth for causal inferences using Griffiths and Tenenbaum's *causal support*, a Bayesian cognition model that estimates the posterior log odds of a target data generating model over a set of alternative data generating models, given a data set. In Experiment 1, we targeted causal support for explanation A over explanation B:

$$\mathfrak{cs}_A = \log\left(\frac{\Pr(C|model_A)}{\Pr(C|model_B)}\right) + \log\left(\frac{\Pr(model_A)}{\Pr(model_B)}\right)$$

where C is the data set we label with ground truth, and models $model_A$ and $model_B$ correspond to causal explanations A and B (Fig. 5.2).

The first term in the formula for $\mathfrak{cs}_A$ is a log likelihood ratio representing the relative compatibility of a given data set with causal explanations A and B. We computed the log likelihood of each data set given $model_A$ and $model_B$ using Monte Carlo simulations

Algorithm 1 Monte Carlo simulation to calculate causal support in Experiment 1. Algorithm for Experiment 2 is similar.

Input: (8, 1) vector of contingency table counts (no disease vs disease, no gene vs gene, no treatment vs treatment) C, Monte Carlo iterations m, set of parameters to fix at zero θ_0 (i.e., parameters representing DAG arrows to omit from the data generating process)

Output: MONTE_CARLO returns log likelihood of the given data generating process *log_lik*; Main returns causal support for the target explanation (Fig. 5.2, Explanation A) $\mathfrak{cs}_A$

```
# Monte Carlo simulation to calculate likelihood
function MONTE_CARLO(C, m, θ_0):
    Parameters corresponding to each DAG arrow in Fig. 5.2:
    θ_P = {        # initialize parameters
        Pr(D),     # p disease due to unknown causes
        Pr(D|G),   # p disease due to gene
        Pr(¬D|T) # p no disease due to treatment
    }
    for parameter θ ∈ θ_P do  # assign parameters
        if θ ∈ θ_0 then Fix parameter at zero: θ = Zeros(m)
        else Uniformly sample probabilities: θ = Random(0, 1, m)
    Calculate probabilities corresponding to contingency table:
    P = [     # p no disease vs p disease given...
        (1 − Pr(D)),     # no treat ¬T, no gene ¬G
        Pr(D),
        (1 − Pr(D)) + Pr(D) · Pr(¬D|T),        # T, ¬G
        Pr(D) · (1 − Pr(¬D|T)),
        (1 − Pr(D | G)) · (1 − Pr(D)),          # ¬T, G
        Pr(D | G) + Pr(D) − Pr(D | G) · Pr(D),
        ( Pr(D | G) + Pr(D) − Pr(D | G) · Pr(D)) # T, G
            · Pr(¬D|T) + (1 − Pr(D | G)) · (1 − Pr(D)),
        ( Pr(D | G) + Pr(D) − Pr(D | G) · Pr(D))
            ·(1 − Pr(¬D|T))
    ]
    return average log likelihood of data:
    log_lik = Σ_{i∈{1,...,m}} Σ (C · log(P)) − log(m)
# Main:  causal support calculation
Calculate likelihood of data given causal explanations A and B:
log_lik_A = MONTE_CARLO(C, 10000, {})
log_lik_B = MONTE_CARLO(C, 10000, {Pr(¬D|T)})
return causal support for explanation A:  # Bayesian update
cs_A = (log_lik_A − log_lik_B) + ( log(0.5) − log(0.5))
```

(Alg. 1, lines 27-29), based on structural models similar to those we used to generate data sets. In practical scenarios, we would not know the true data generating parameters, so we used Monte Carlo simulations of possible parameter values under each model to calculate likelihoods without needing to know the ground truth *a priori*. Under $model_A$ we sampled all three parameters uniformly on the interval $[0, 1]$, representing the assumptions that there is a treatment effect and that both gene and unobserved factors cause disease. Under $model_B$ we sampled $\Pr(\mathcal{D}|G)$ and $\Pr(\mathcal{D})$ uniformly, but we fixed $\Pr(\neg\mathcal{D}|T)$ at zero, representing an assumption of no treatment effect (i.e., omitting the DAG arrow between treatment and gene in Fig. 5.2). In each simulation, we averaged log likelihood of a given data set over $m = 10000$ Monte Carlo iterations (Alg. 1, lines 24-25), marginalizing over sampled parameter values.

The second term in the formula for $\mathfrak{cs}_A$ is a log ratio of the prior probability of explanations A versus B. Following Griffiths and Tenenbaum [77], we assume a uniform prior to be normative, assigning 50% probability to both explanations A and B (Alg. 1, lines 30-31). The prior encodes a bias in belief allocation across a finite set of alternative causal explanations. We assume a uniform prior because we want our benchmark to reflect no *a priori* bias toward causal explanations.[4]

Performance evaluation

We wanted to measure how much participants' causal inferences deviated from our normative benchmark, causal support.

The linear in log odds model & causal support. By choosing to model perceived causal support lrr_A (see Section 5.2.1.1) as a function of ground truth causal support on a log odds scale, we leverage a linear in log odds (LLO) model to extend causal support from a normative cognitive model into a descriptive one. Prior work shows that the LLO model accurately describes natural distortions in mental representations of probability [71, 91, 185, 228]. For example, visualization researchers [109] used the LLO model to measure perceptual distortions in probabilistic judgments about intervention effectiveness. Our normative model

[4]Uniform priors follow a convention of psychometric models that assume guessing responses are informed by the number of response alternatives [121].

of causal support itself (see Section 5.2.1.4) is a sum in log odds units.

Derived measures. Using a LLO model to measure the correspondence between normative and perceived causal support enables us to estimate (1) participants' *sensitivity* to changes in ground truth causal support and (2) *bias* in perceived causal support. From our model, we derive sensitivity and bias per condition as *LLO slopes and intercepts*, respectively. LLO slopes describe sensitivity to ground truth causal support such that a slope of one indicates ideal sensitivity. One can think of slopes as the weight participants assign to changes in the ground truth log likelihood ratio of explanations A versus B. LLO intercepts describe bias in participants' probability allocations when causal support is zero such that an intercept of 50% indicates no bias. One can think of intercepts as the average prior probability that participants allocate to explanation A when there is no signal in the data.

Approach. We used the brms package [29] in R to fit Bayesian hierarchical models on perceived causal support. We adopted a Bayesian workflow called *model expansion* [61], where we started with a simple model and iteratively added predictors to build up to more complex models, running prior predictive checks, model diagnostics, posterior predictive checks, and leave-one-out cross validation for each version of the model. We centered each prior to reflect a null hypothesis of ideal performance and no bias, and we scaled each prior to be weakly informative while providing sufficient regularization for models to converge. We provide more details about our modeling workflow in our preregistrations[5] and Supplemental Materials.[6]

Model specification. We used the following model (Wilkinson-Pinheiro-Bates notation [29, 163, 217]) to evaluate participants' responses:

$$
\begin{aligned}
lrr_A \sim & \text{Normal}(\mu_A, \sigma_A) \\
\mu_A = & \mathfrak{cs}_A * delta\_p * n * vis + \left(\mathfrak{cs}_A * delta\_p + \mathfrak{cs}_A * n \middle| worker\_id\right)
\end{aligned}
$$

where lrr_A was perceived causal support for a treatment effect, $\mathfrak{cs}_A$ was our normative

[5] See preregistrations for Experiment 1 (`https://osf.io/vzmhu`) and for Experiment 2 (`https://osf.io/y46nw`)

[6] `https://github.com/kalealex/causal-support`

benchmark, *delta_p* was the difference in the proportion of people with disease given treatment versus no treatment, n was the sample size as a factor, *vis* was a dummy variable for visualization condition, and *worker_id* was a unique identifier for each participant.

We primarily modeled effects on the mean of perceived causal support μ_A, but our model also learned the residual standard deviation σ_A. Both σ_A and the random effects in the μ_A submodel helped account for the empirical distribution, differentiating between response noise and effects of interest. The term $\mathfrak{cs}_A * delta\_p * n * vis$ enabled our model to learn how the slope on causal support varies as a function of the visual signal on each trial (*delta_p* and n) and visualization condition.

Participants & exclusions

We recruited participants on Amazon Mechanical Turk. Workers had a HIT acceptance rate of at least 97% and were located in the US. We aimed to recruit a total of 400 participants after exclusions using our attention check trials, 80 per visualization condition. We determined this target sample size using a heuristic power analysis based on pilot data and the assumption that the width of confidence intervals would be inversely proportional to $\sqrt{N}$. We recruited a total of 548 participants, and after exclusions we used data from 408 participants in our analysis. We slightly overshot our target sample size because we could not anticipate perfectly how many participants would miss our attention checks (see Section 5.2.1.4). Although we preregistered that we would exclude participants who failed to allocate at least 50% subjective probability to the most likely causal explanation on either attention check, this criterion proved too strict and would have excluded 48% of our sample. Instead, we opted to use only the easier of the two attention checks for exclusions, resulting in the exclusion of 26% of our sample. All participants were paid regardless of exclusions. We compensated the average participant $2.50 for about 9 minutes.

5.2.2 Results

We evaluate chart users' causal inferences using a linear in log odds (LLO) model to assess *sensitivity* to the ground truth and *bias* in probability allocations when each causal

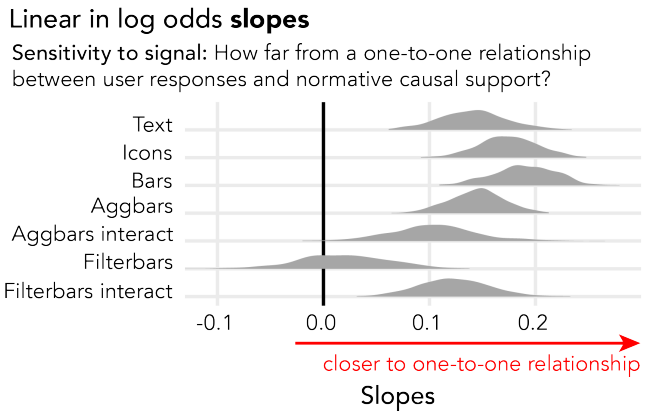


Figure 5.6: Linear in log odds (LLO) slopes per visualization condition.

explanation is equally likely.

Sensitivity. A LLO slope of one indicates one-to-one correspondence between the ground truth and users' probability allocations. In all visualization conditions (Fig. 5.6), we see slopes far below one, indicating that users are much less sensitive than ideal. The only reliable differences between visualization conditions are that filterbars users who do not interact are less sensitive than users in other conditions.

When filterbars users do not interact with the charts, slopes are approximately zero indicating that users are insensitive to signal. Performance improves reliably when users interact with the visualization by applying cross-filters to coordinated multiple views. This is expected because filterbars hide visual signal for the task behind interactions.

Surprisingly, when aggbars users interact with the charts to group by gene or treatment, this leads to lower sensitivity, though this difference is not reliable. To make sense of this result, we analyze interaction log data to see which variables chart users condition on. We find that aggbars users group the data by gene almost as often as treatment. Compare this to filterbars users, who condition on treatment much more often than gene (see Supplemental Material). This suggests that interacting with visualizations only improves sensitivity to causal support when users deliberately generate views of the data which show counterfactual predictions that can distinguish competing causal explanations.

Visual signal effects on sensitivity. We examine sensitivity in each visualization as function of attributes of the visual signal for our task. In Experiment 1, the signal breaks

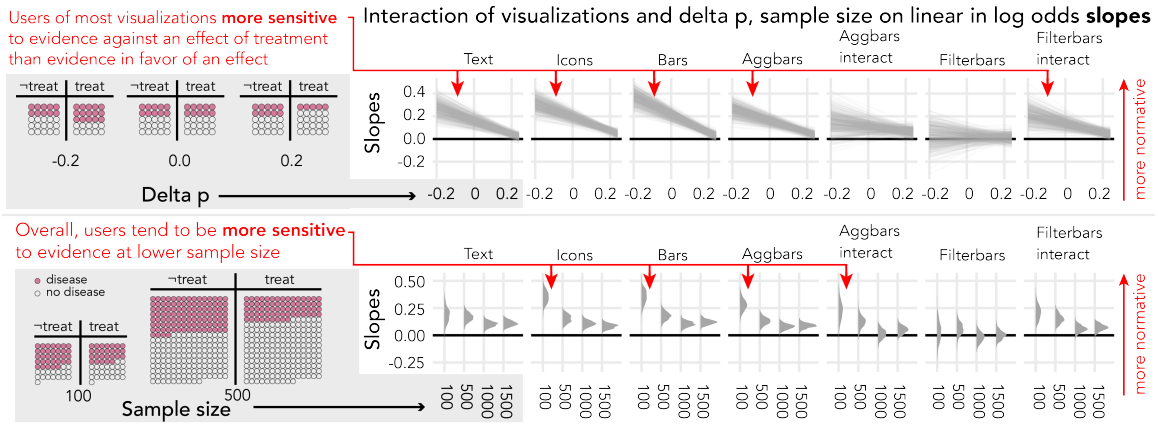


Figure 5.7: Sensitivity (y-axes) conditioned on two attributes of visual signal for treatment effectiveness (rows, x-axes) and visualizations (columns).

down into two data attributes, *delta p* and *sample size* (see Section 5.2.1.4). Normatively, LLO slopes equal one regardless of delta p and sample size, however, our model measures differences in sensitivity depending on these data attributes.

Figure 5.7 shows that in the conditions where slopes are largest—text, icons, bars, aggbars without interaction, and filterbars with interaction—users are more sensitive to causal support at negative values of delta p (e.g., Fig. 5.7, top inset). The average user in these conditions responds more to evidence against treatment (i.e., falsification) than evidence in favor of a treatment effect (i.e., verification). At positive values of delta p (e.g., Fig. 5.7, top inset), LLO slopes are similar across conditions, suggesting that differences in performance between conditions are driven in part by differences in sensitivity to falsifying evidence.

We also see in Figure 5.7 that users of icons, bars, and aggbars are more sensitive to signal when sample size is smaller. This finding is consistent with prior work showing that chart users tend to underestimate sample size when making inferences with data [119], which may be driven by logarithmic perception [204, 228]. Alternatively, we could interpret this result as a cognitive bias where users are unwilling to be certain even when sample size is large enough to support unambiguous inferences, related to non-belief in the law of large numbers [17].

Bias. Intercepts in the LLO model describe bias in users' probability allocations when

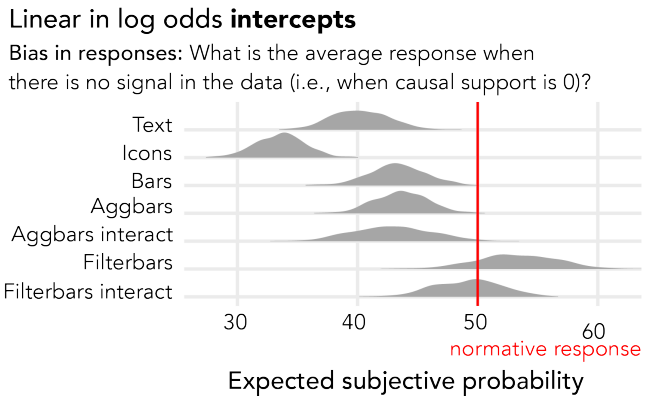


Figure 5.8: LLO intercepts per visualization condition.

ground truth causal support indicates that explanation A (i.e., treatment effect) is just as likely as explanation B (i.e., no treatment effect). Under this condition, a normative observer would allocate equal probability to both causal explanations. We derive expected probability allocated to explanation A based on a logistic transform of LLO intercepts, and compare this to the normative benchmark of 50%.

With all visualizations except for filterbars, probability allocations are far below 50% indicating substantial bias (Fig. 5.8). On average when causal support is zero, users of text tables, icons, bars, and aggbars allocate too little probability to causal explanation A. Users of filterbars, on the other hand, allocate approximately 50% to explanation A. We see the most extreme bias of up to 20% with icons arrays.

Unfortunately, we can only speculate about possible reasons for these biases. We expected that LLO intercepts would indicate average responses near 50% in the absence of signal for all conditions (i.e., a uniform prior), simply because this follows from the structure of the task. Because this pattern of biases across visualizations results from a non-preregistered exploratory comparison, we investigate in Experiment 2 whether these biases replicate for a more complex task.

5.3 Experiment 2

In Experiment 2, we evaluate the same visualization designs on a more difficult task. We asked participants to detect confounding in the presence of a known treatment effect by

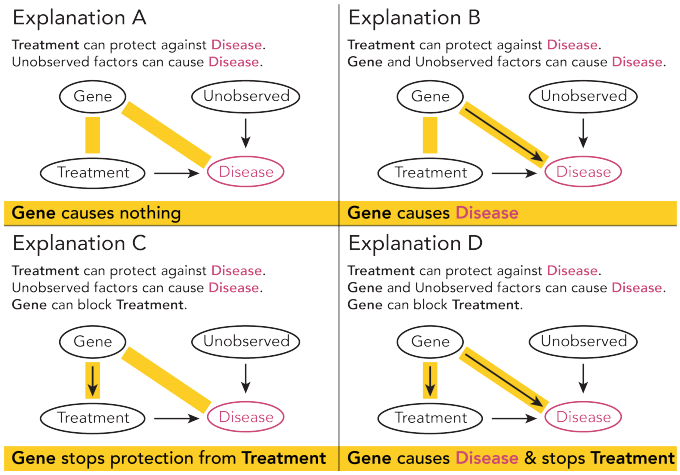


Figure 5.9: DAGs for possible causal explanations in Experiment 2.

allocating probability across four possible "backdoor paths" [160] (Fig. 5.9). We extend causal support to handle more than two alternative causal explanations, demonstrating how causal support can be employed in more complex analyses.

5.3.1 Method

Experiment 2 was the same as Experiment 1 except for the following changes to response elicitation, modeling, and experimental design.

Task scenario & response elicitation

Participants judge the influence of a gene on both disease and treatment effectiveness by allocating probability across the four DAGs in Figure 5.9, separately assessing each DAG arrow in a confounding relationship.

Question & elicitation. We asked participants a similar question as in Experiment 1, where participants allocated 100 votes (i.e., subjective probability) across alternative causal explanations. However, in Experiment 2 we elicited a Dirichlet distribution with four alternatives. Following Chalone et al. [33,150] and extending our interface from Experiment 1, each time participants allocated a number of votes between 0 and 100 to an option, the *remaining votes out of 100 were uniformly distributed across unused response options.* These imputed responses were highlighted along with a prompt to, *"Adjust your responses until all*

the numbers reflect your beliefs." Participants iteratively set and adjusted their probability allocations. We combined these responses into perceived causal support, which we compared to our benchmark.

Perceived causal support. When estimating *perceived causal support* in Experiment 2, we separately evaluated *multiple target explanations*. Primarily, we targeted belief in explanation D (llr_D), confounding:

$$lrr_D = \log\left(\frac{response_D}{\sum_{i=\{A,B,C\}} response_i}\right)$$

where $response_A$, $response_B$, $response_C$, and $response_D$ were participants' probability allocations to causal explanations A through D, respectively, on each trial. We also separately targeted belief in both of the component DAG arrows that constitute a confounding relationship (Fig. 5.9): (llr_{BD}) the effect of gene on disease, which appears in explanations B and D; and (llr_{CD}) the effect of gene on treatment effectiveness, which appears in explanations C and D. We define llr_{BD} as follows,

$$lrr_{BD} = \log\left(\frac{\sum_{i=\{B,D\}} response_i}{\sum_{i=\{A,C\}} response_i}\right)$$

and we define llr_{CD} similarly. We compare log response ratios llr_D, llr_{BD}, and llr_{CD} to corresponding causal support $\mathfrak{cs}_D$, $\mathfrak{cs}_{BD}$ and $\mathfrak{cs}_{CD}$.

Strategy. At the end of the experiment, we asked participants,

> ***How did you use the charts to complete the task?*** *Please tell us what patterns you looked for in the data and what comparisons you made.*

We analyzed these qualitative responses to assess whether participants understood how to use the charts for the confounding detection task.

Experimental design

We manipulated both the visualizations (between-subjects) and the data sets we showed (within-subjects). We showed each participant 19 trials, 18 data conditions (see Section 5.3.1.3) and one attention check used for exclusions (see Section 5.3.1.5). We randomized

trial order for each participant, inserting the attention check on trial 10.

Stimulus generation

We generated data sets that spanned a range of ground truth causal support *for confounding*. Creating these data sets required (1) manipulating data attributes which signaled whether the gene was a confounding factor, and (2) labeling each data set with ground truth causal support.

Our goal was to generate 18 data conditions that varied delta p disease, delta p treatment, and sample size, data attributes which manipulate causal support for confounding. *Delta p disease* described the difference in the proportion of people with disease in each data set depending on whether they had the gene. Negative values of delta p disease indicated evidence that the gene caused disease, whereas values near zero indicated evidence against a gene effect on disease. *Delta p treatment* described the difference in the proportion of people with disease within the treatment group depending on whether they had the gene. Negative values of delta p treatment indicated evidence that the gene stopped the treatment from preventing disease, whereas values near zero indicated evidence against a gene effect on treatment. *Sample size* was the number of people in each data set we showed chart users.

Data conditions. To generate 18 data conditions, we simulated data from structural models with one parameter per DAG arrow in Figure 5.9. We manipulated the *probability that gene causes disease* (3 levels: $\{0, 0.35, 0.7\}$), the *probability that gene prevents treatment from working* (3 levels: $\{0, 0.35, 0.7\}$), and *sample size* (2 levels: $\{100, 1000\}$). We controlled the probability that treatment prevents disease (0.8), probability of disease due to unobserved causes (0.2), base rate of the gene (0.4), and the proportion of each sample with treatment (0.5). We selected these parameters iteratively by sampling data sets and labeling ground truth until half of trials had greater than a 25% chance of having been generated by causal explanation D.

As in Experiment 1, we simulated many data sets for each data condition, and we counterbalanced quantiles of sampling error across participants (see Section 5.2.1.4). For our attention check trial, we selected the simulated data set that maximized causal support

for confounding.

Labeling ground truth causal support. We extended Griffiths and Tenenbaum's model of causal support [77] to account for more than two alternative causal explanations. We primarily targeted causal support for causal explanation D over explanations A, B or, C,

$$\mathfrak{cs}_D = \log\left(\frac{\Pr(C|model_D)}{\sum_{i=\{A,B,C\}}\Pr(C|model_i)}\right) + \log\left(\frac{\Pr(model_D)}{\sum_{i=\{A,B,C\}}\Pr(model_i)}\right)$$

where C is the data set we label with ground truth, and $model_A$, $model_B$, $model_C$, and $model_D$ correspond to causal explanations A through D (Fig. 5.9), respectively. Since we separately targeted belief in both of the component DAG arrows that constitute a confounding relationship (see Section 5.3.1.1, perceived causal support), we needed to calculate ($\mathfrak{cs}_{BD}$) ground truth causal support for explanations B or D over A or C:

$$\mathfrak{cs}_{BD} = \log\left(\frac{\sum_{i=\{B,D\}}\Pr(C|model_i)}{\sum_{i=\{A,C\}}\Pr(C|model_i)}\right) + \log\left(\frac{\sum_{i=\{B,D\}}\Pr(model_i)}{\sum_{i=\{A,C\}}\Pr(model_i)}\right)$$

We similarly calculated ($\mathfrak{cs}_{CD}$) causal support for explanations C or D.

The first terms in the formulae for $\mathfrak{cs}_D$, $\mathfrak{cs}_{BD}$, and $\mathfrak{cs}_{CD}$ are log likelihood ratios representing the relative compatibility of a given data set with causal explanations A, B, C, and D. We calculated the log likelihood of each data set we showed participants given $model_A$, $model_B$, $model_C$, and $model_D$ using Monte Carlo simulations similar to Algorithm 1. In Experiment 2, we introduced one more parameter $\Pr(\neg T|G)$ to our structural models, representing the probability that the gene prevents the treatment effect. We incorporate this parameter into our Monte Carlo simulations (Alg. 1) by making the following substitutions:

$$\begin{aligned}
19: &\ (\Pr(\mathcal{D}|G) + \Pr(\mathcal{D}) - \Pr(\mathcal{D}|G)\cdot\Pr(\mathcal{D}))\cdot(\Pr(\neg\mathcal{D}|T)\\
20: &\quad \cdot(1-\Pr(\neg T|G)) + (1-\Pr(\mathcal{D}|G))\cdot(1-\Pr(\mathcal{D})),\\
21: &\ (\Pr(\mathcal{D}|G) + \Pr(\mathcal{D}) - \Pr(\mathcal{D}|G)\cdot\Pr(\mathcal{D}))\\
22: &\quad \cdot((1-\Pr(\neg\mathcal{D}|T)) + \Pr(\neg\mathcal{D}|T)\cdot\Pr(\neg T|G))
\end{aligned}$$

Under $model_A$ we sampled $\Pr(\mathcal{D})$ and $\Pr(\neg\mathcal{D}|T)$ uniformly on the interval $[0,1]$ and fixed

$\Pr(\mathcal{D}|G)$ and $\Pr(\neg T|G)$ at zero, representing assumptions that the gene impacts neither disease or treatment. Under $model_B$ we sampled $\Pr(\mathcal{D})$, $\Pr(\mathcal{D}|G)$, and $\Pr(\neg\mathcal{D}|T)$ uniformly and fixed $\Pr(\neg T|G)$ at zero, representing the assumption that the gene has no effect on treatment. Under $model_C$ we sampled $\Pr(\mathcal{D})$, $\Pr(\neg T|G)$, and $\Pr(\neg\mathcal{D}|T)$ uniformly and fixed $\Pr(\mathcal{D}|G)$ at zero, representing the assumption that the gene has no effect on disease. Under $model_D$ we sampled all four parameters uniformly to represent confounding.

The second terms in the formulae for $\mathfrak{cs}_D$, $\mathfrak{cs}_{BD}$, and $\mathfrak{cs}_{CD}$ are log ratios of the prior probabilities of the target explanation(s) versus other possible explanations. Again, we assumed a uniform prior to create an unbiased benchmark for our task such that 25% was the normative prior probability for each causal explanation A, B, C, and D, respectively.

Performance evaluation

Again, we used linear in log odds (LLO) models [71, 228] to describe discrepancies between perceived and normative causal support. We also conducted a qualitative analysis of participants' reported strategies.

Model specification. We used three inferential models because we had three dependent variables, representing perceived causal support for confounding (lrr_D) and for the two constituent effects of confounding (lrr_{BD} and lrr_{CD}). Here, we show only the models on lrr_D and lrr_{BD} because the models on lrr_{CD} and lrr_{BD} are identical in form, with $\mathfrak{cs}_{CD}$ and $delta\_p\_t$ replacing $\mathfrak{cs}_{BD}$ and $delta\_p\_d$ as predictors:

$$
\begin{aligned}
lrr_D \sim & \text{Normal}(\mu_D, \sigma_D) \\
\mu_D = & \mathfrak{cs}_D * delta\_p\_d * delta\_p\_t * n * vis \\
& + \left(\mathfrak{cs}_D * delta\_p\_d + \mathfrak{cs}_D * delta\_p\_t + \mathfrak{cs}_D * n \middle| worker\_id\right) \\
lrr_{BD} \sim & \text{Normal}(\mu_{BD}, \sigma_{BD}) \\
\mu_{BD} = & \mathfrak{cs}_{BD} * delta\_p\_d * n * vis + \left(\mathfrak{cs}_{BD} \middle| worker\_id\right)
\end{aligned}
$$

where lrr_D, lrr_{BD}, and lrr_{CD} were perceived causal support for a confounding, the gene effect on disease, and the gene effect on treatment, respectively, $\mathfrak{cs}_D$, $\mathfrak{cs}_{BD}$, and $\mathfrak{cs}_{CD}$ were our normative benchmarks corresponding to each log response ratio, $delta\_p\_d$ was the

difference in the proportion of people with disease given gene versus no gene, *delta_p_t* was the difference in the proportion of people with disease among those in the treatment group given gene versus no gene, n was the sample size as a factor, *vis* was a dummy variable for visualization condition, and *worker_id* was a unique identifier for each participant.

We primarily modeled effects on the mean of perceived causal support μ_D, μ_{BD}, and μ_{CD}, but our models also learned residual standard deviations σ_D, σ_{BD}, and σ_{CD}. The residual standard deviations and random effects in each model helped us separate patterns in responses from noise and individual differences. In the first model, we used the term $\mathfrak{cs}_D * delta\_p\_d * delta\_p\_t * n * vis$ to learn how sensitivity to causal support for confounding varies as a function of sample size n and visualization *vis*. In the second and third models, we used the terms $\mathfrak{cs}_{BD} * delta\_p\_d * n * vis$ and $\mathfrak{cs}_{CD} * delta\_p\_t * n * vis$ to learn how sensitive users in each visualization condition were to the gene effects on disease *delta_p_d* and treatment *delta_p_t*, respectively.

Qualitative analysis. We wondered how well participants would intuit how to perform the confounding detection task, considering it was more difficult than the task in Experiment 1, and we provided no training. To address this we applied a deductive coding scheme. We coded participants' strategy descriptions as *uninformative* if they didn't describe a strategy. Otherwise, we coded whether or not participants described adequate strategies for judging *delta p disease*, *delta p treatment*, or *sample size* (see Section 5.3.1.3), and we coded *confusion* if they stated they were confused or described an incorrect strategy.

Participants & exclusions

We used a similar approach to power analysis as in Experiment 1 to determine a target sample size of 500 participants after exclusions. We recruited a total of 703 participants, and after exclusions we used data from 519 participants in our analysis. Although we preregistered that we would exclude participants who allocated less than 25% probability to confounding on an attention check trial where confounding was very likely (see Section 5.3.1.3), this criterion would have excluded 39% of our sample. We relaxed the cutoff to less than 20% probability of confounding to allow for additional response error, resulting in

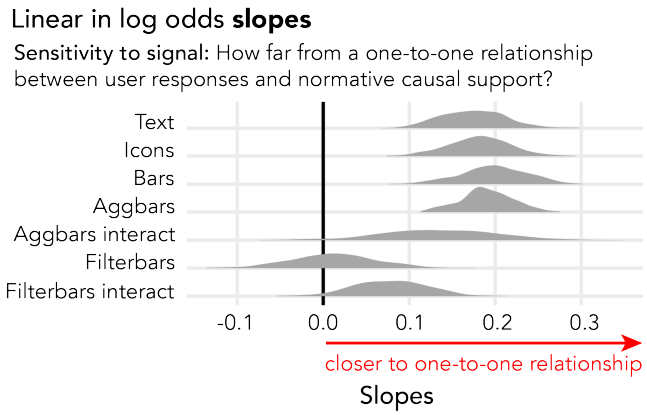


Figure 5.10: Linear in log odds (LLO) slopes per visualization condition.

a 26% exclusion rate. We paid participants $3.04 for 14 minutes on average.

5.3.2 Results

We use a linear in log odds (LLO) model to describe performance in terms of *sensitivity* to ground truth causal support and *bias* in probability allocations when all four causal explanations are equally likely.

Sensitivity. A LLO slope of one indicates ideal sensitivity to the log likelihood of the data given a set of causal explanations Similar to Experiment 1, slopes in all visualization conditions are closer to zero than one (Fig. 5.10), indicating under-sensitivity to the ground truth.

Interacting with filterbars seems to improve sensitivity, while interacting with aggbars seems to decrease sensitivity, although these differences are not reliable. It is surprising to see a similar pattern of results for interactive visualizations in both experiments, since we expected interactive visualizations to be more helpful for detecting confounding than for detecting a treatment effect. Detecting confounding requires users to look for more complex counterfactual patterns in order to distinguish between causal explanations, and manipulating data aggregation and filtering should help users to query visualizations for these patterns. When we analyze interaction logs (see Supplemental Materials), we see that filterbars users interacted with the visualizations more frequently and created more task-relevant views of the data than aggbars users, which may help to explain why interacting

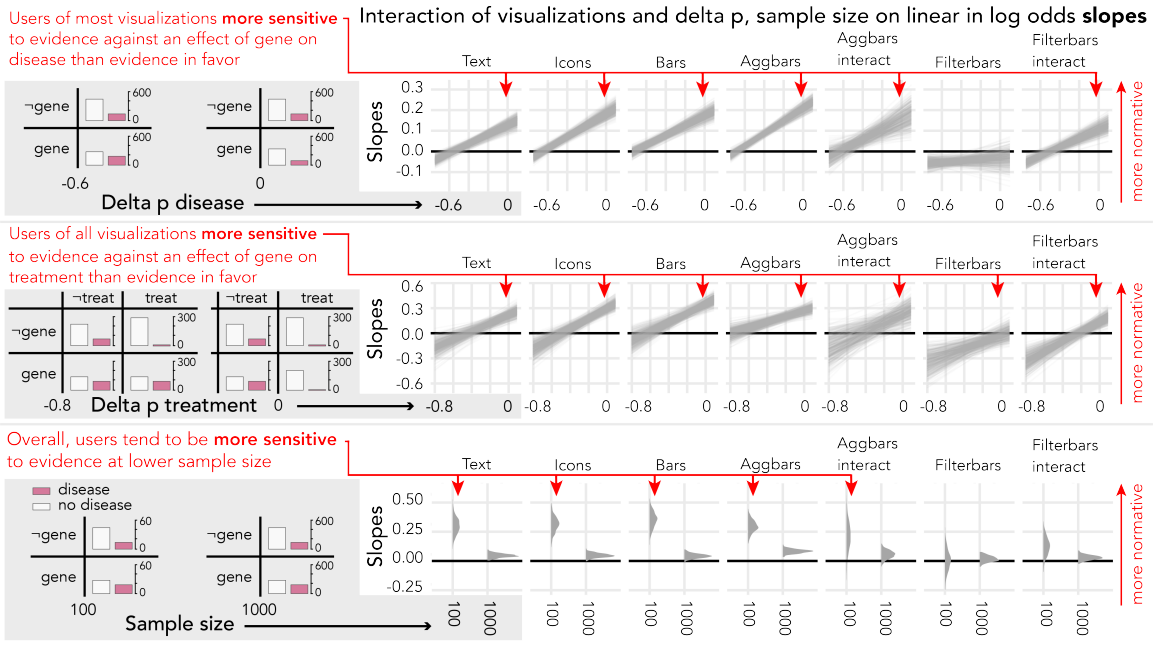


Figure 5.11: Sensitivity (y-axes) conditioned on three attributes of visual signal for confounding (rows, x-axes) and visualization conditions (columns).

with filterbars was somewhat more helpful than interacting with aggbars.

Visual signal effects on sensitivity. We examine sensitivity in each visualization condition to the three visual signals for confounding in our task (delta p disease, delta p treatment, and sample size; see Section 5.3.1.3). Normatively, slopes are one regardless of these visual signals.

Figure 5.11 shows users are more sensitive to causal support at values of delta p disease and delta p treatment near zero, with the exception of filterbars users who don't interact. This pattern is consistent with the findings of Experiment 1 in that chart users respond more to evidence against a given causal effect than evidence in favor of an effect.

In Figure 5.11, we also see that users of every visualization but filterbars are more sensitive when sample size is smaller. This pattern is consistent with prior work [17, 119] and the results of Experiment 1.

Bias. LLO intercepts describe bias in probability allocations when the data are equally likely under each alternative causal explanation. We derive expected probability allocated to explanation D based on LLO intercepts and compare this to the normative benchmark

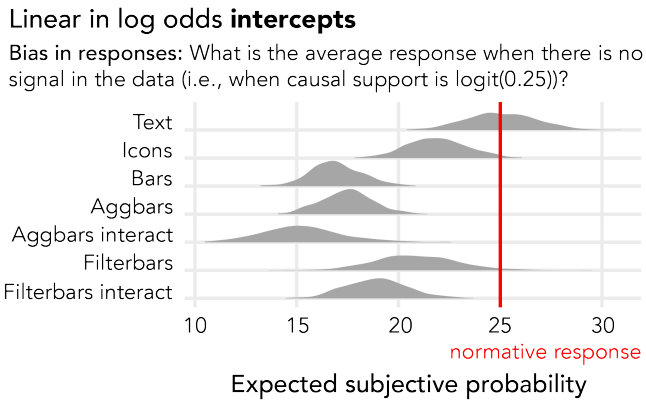


Figure 5.12: LLO intercepts per visualization condition.

of 25%.

Figure 5.12 shows that, with all visualizations but text tables, users underestimate the probability of confounding in the absence of signal. The fact that biases for each visualization condition differ between Experiments 1 and 2 suggests that these results are task-specific. Future work should study reasons for these biases and what visual analytics software can do to help calibrate analysts' probability allocations.

Strategies. We assess users' self-reported strategy descriptions. 235 of 519 (45%) users included in our analysis gave uninformative responses and were excluded from further analysis. 42 of 284 (15%) remaining users either stated they were confused or described an incorrect strategy.

However, many users intuited the important signals in the data:

> *"I relied more on the 'no treatment' cells to consider whether the gene causes the [disease], trying to look at ratio of 'disease' and 'no disease' within those two quadrants... [I] tried to consider the actual counts remembering that small numbers mean loose estimates but this was easy to overlook. Then I compared the two purple bars in the 'gene no' top-half of graph to estimate the treatment effect... and did the same for the two lower purple bars to see if treatment equally effective in those with the gene."*

222 of 284 (79%) described an adequate strategy for inferring the gene effect on disease. 81 of 284 (29%) mentioned sample size information. 168 of 284 (59%) described an adequate strategy for inferring the gene effect on treatment effectiveness. These results suggest that much of our data represent a reasonable understanding of the task, yet participants still appeared to struggle to use the visualizations effectively.

5.4 Discussion

We demonstrate the utility of causal support for evaluating inferences with visualizations, successfully measuring expected patterns in the quality of chart users' causal inferences. For example, filterbars users should not have been able to perform either task without interacting because the visual signals required to perform the tasks were hidden behind interactions. Our method shows that filterbars users were completely insensitive to the signal in data when they did not interact. Similarly, our models corroborate prior work suggesting that chart users underweight sample size when making inferences [17,119]. Findings like these reassure us that causal support can help us understand how users struggle to use visualizations to evaluate causal hypotheses.

Our findings point to unsolved design challenges for supporting causal inferences with visual analytics (VA) tools. Contrary to what we might expect given the emphasis of visualization research on evaluating encodings and interaction techniques, using different encodings for count data doesn't appear to improve sensitivity to evidence for causal inferences beyond text contingency tables. Similarly, common interaction techniques in VA tools, such as manipulating data aggregation or cross-filtering coordinated multiple views, don't seem to improve causal inferences beyond what users can achieve with simpler static visualizations. Interacting with visualizations seems to help or hurt sensitivity depending on how deliberately signal-seeking users are and whether interacting is necessary in order to expose the visual signal in the data. This suggests that VA tools designed to optimize easy exposure of data are not sufficient for supporting causal inferences.

We also find systematic biases in the way that chart users respond to specific visual signals in charts. *Chart users seem ubiquitously more sensitive to falsifying evidence than they are to verifying evidence.* This may reflect a cognitive bias where analysts are more responsive to *discrepancies*, between observed data and the counterfactual patterns expected under a given causal explanation, than they are to *similarities* between observed data and counterfactual patterns. Interestingly, this bias may be somewhat rational to the extent that verifying an inference is probabilistic, whereas the logic of falsification is deductive and thus "more powerful" in that it can definitively rule out an explanation [165].

Insensitivity to sample size remains a major challenge for informal statistical inferences, and it appears not to be sufficiently addressed by common chart types for showing count data. Even icon arrays, which emphasize sample size as the number of equal-sized dots, don't seem to mitigate this problem. Prior work [17, 119] suggests this may be due to perceptual underestimation of sample size and cognitive bias against claiming certainty in inferences. Additionally, our qualitative results suggest chart users may not intuitively pay as much attention to sample size as they do to other signals when making causal inferences.

Consistent with an aversion to believing causal relationships exist, we find that chart users tend to underestimate the probability of a given DAG arrow. In the absence of any signal differentiating between causal explanations, chart users allocate more probability to explanations that posit fewer relationships, rather than allocating probability uniformly across alternatives. Though this tendency interacts with task and visualization in ways that warrant further study, it may reflect an overall cognitive bias toward believing in simpler causal explanations.

5.4.1 Limitations & future work

We set out to run a proof-of-concept study establishing causal support as an evaluation method for VA tools, and our study raises many unanswered questions. A primary limitation of this work is that we recruited participants on Mechanical Turk, who may be less sensitive to causal support than real data analysts to the extent that they may use VA tools less deliberately. However, our qualitative analysis suggests that many participants understood the task and used reasonable strategies. Future work may find causal support helpful in evaluating current practices or novel interfaces with smaller pools of participants, insofar as real data analysts give less noisy responses than crowdworkers. Questions remain about whether our findings generalize for other data types (e.g., continuous [153] and event stream data [152]), for domains outside of medicine, and for analysis scenarios with more complex possible data generating models. Though we suspect our findings will persist in some form across user populations and analysis scenarios, visualizations probably will support some other causal inference tasks better than they support differentiating possible data generating

processes.

5.4.2 Improving visual analytics for causal inference

A theme in visual causal inference is that analysts do not always know what to look for in data [12, 224]. Causal inferences differentiating between possible data generating processes (DGPs) require comparisons between patterns in observed data and counterfactual patterns under a specific DGP [75, 160]. Users of VA software may struggle with causal inferences insofar as they fail to *imagine* counterfactual predictions.

Prior work in statistics and visualization argues for *model checks* that make comparisons between data and model predictions explicit [66, 67, 96]. For example, workflows in Bayesian statistics frequently employ prior and posterior predictive checks [61].Visualizing model predictions alongside empirical data could support causal inference by externalizing discrepancies and similarities between observed and expected patterns.

We envision a VA workflow where analysts cycle between interactively specifying models (e.g., [127]) and generating model checks to gauge model compatibility with their data. This echos calls to make models themselves a primary goal of visual data analysis [8]. Causal support solves an important problem in realizing this vision, defining a "good" model check as one which supports sensitive inferences among a set of candidate DGPs. Though it may be difficult to come up with an exhaustive set of DGPs in many real world applications, we think that this approach would be fruitful even with a relatively simple set of models that a knowledgeable analyst might provisionally entertain. Causal support cannot guard against analysts ignoring possible models, but it can be used to evaluate visualization and interaction designs intended to help analysts collate and compare alternative models.

5.5 Summary

This chapter presents two crowdsourced experiments demonstrating an approach to evaluating causal inferences with visual analytics (VA) tools. No visualization or interaction designs we tested lead to reliably better causal inferences than text contingency tables, suggesting that common VA tools designed for data exposure may not be sufficient for supporting causal inferences. We point to perceptual and cognitive biases which seem to

make visual causal inferences difficult, including tendencies to underweight both evidence verifying a causal relationship and evidence from large samples. We discuss how formal models of causal support can be used to evaluate VA systems that place an emphasis on helping users reason about possible data generating processes.

5.6 Acknowledgements

Alex Kale led this study and worked independently on all of its components. Jessica Hullman and Yifan Wu gave feedback on experimental design, interface design, and writing. The idea to use causal support [77] to evaluate VA tools was initially explored by Yifan Wu in a class project[7], and Jessica suggested following up on the idea in December of 2020. From that point, Alex Kale planned and executed the entirety of the study in 3 months.

We thank the UW IDL and the NU MU Collective for their feedback. We thank NSF (#1930642) for funding this work.

[7] `https://logical-interactions.github.io/causal2020/`

Chapter 6

EVM: EXPLORATORY VISUAL MODELING

In this chapter of my dissertation, I present Exploratory Visual Modeling (EVM), a graphical user interface for exploratory data analysis that enables users to express and check their provisional interpretations of data in the form of statistical models. At the time of this writing, EVM is a prototype implementation that serves as a proof-of-concept for applying design knowledge about cognitive mechanisms, which I have built up through my empirical work presented in Chapters 3–5, to help data analysts reason about uncertainty in data. I present motivation and rationale for the design of EVM along with considerations about how to evaluate the tool with data analysts. Through the design and implementation of EVM, I uncover practical challenges around realizing model checks in visual analytics interfaces, including ways in which model outputs require different treatment that raw data within visual analytics software. This work opens a door to new ways of building analysis tools and presents opportunities for new lines of research addressing fundamental questions about the nature of people's mental models of data and how they should be represented in software.

6.1 The need for model checks in visual analytics

In the previous chapter, I presented evidence that people struggle to reason about causal relationships using visualization and interaction techniques that are common in visual analytics (VA) tools. For example, when asked to gauge the compatibility of a synthetic medical trial dataset (e.g., Fig. 5.3) with the assumption that a treatment influences the rate of a disease, participants made inferences that were mostly insensitive to the displayed evidence for or against this causal explanation. People struggle to make comparisons to an *imagined distribution of counterfactural predictions*—i.e., what the patterns on a chart might look like if a particular account of the data generating process (DGP) were true. One design implication of this work is that tools for sensemaking with data should enable users

to explicitly generate and examine such reference distributions alongside their data.

This design recommendation to make predictive distributions explicit is in line with theories of graphical statistical inference, reviewed recently by Hullman and Gelman [96] and in Chapter 2. In particular, statistical theory suggests explicit comparisons between observed data and *reference distributions*—i.e., predictions drawn from provisional models fit to the data in order to provide a reference for inferences about underlying DGP. These "model checks", as Hullman and Gelman dub them, show up frequently in Bayesian workflows for model building and evaluation [61]. Thinking of visualizations as *model checks*—comparisons between data and model predictions—offers a helpful design strategy to improve visualizations as tools for statistical judgments.

However, model checks seldom enter into graphical user interfaces for exploratory analysis, where they would be helpful for verifying and interrogating discovered patterns. For example, support for modeling in popular commercial VA tools like Tableau is limited to trend lines showing conditional averages. Such superimposed trend lines average over uncertainty information, failing to support analysts in answering questions such as, "Is this pattern real or spurious?" [227], which can be answered more rigorously using simulation or permutation approaches to generate a reference distribution [175]. As a consequence, Hullman and Gelman argue [96], these VA tools leave analysts vulnerable to miscalibrated confidence as they move between exploratory and confirmatory modes of analysis—a distinction which many associate with *viewing versus modeling* data, despite assertions from statistical pioneers like Tukey [197] that inspecting model residuals plays an important role in exploration and discovery.

Here, I set out to realize this vision of model check in graphical user interfaces (GUI) for exploratory data analysis (EDA). I do this by creating Exploratory Visual Modeling (`EVM`), a prototype VA application that enables users to quickly specify reference models and compare predictions from these models to patterns in observed data. `EVM` provides a Tableau-like interface for chart construction and incorporates novel features for authoring provisional statistical models to check against data. Critically, these provisional statistical models are meant to represent the user's mental model of DGP, which updates frequently during EDA in response to discoveries about the data. By providing an explicit way for

users to express and check their mental models of DGP in a VA workflow, I extend the toolkit available to those analysts who typically analyze data in GUIs but tend not to reach for statistical modeling to validate their insights about data.

In this chapter, I provide an exposition of the use cases and principles that guided the design of `EVM`, the basic operations of `EVM`, the implementation of `EVM`, and considerations about how to evaluate whether `EVM` can improve analysts' reasoning about DGP. I share lessons learned from the design and implementation of `EVM`, e.g., the fact that visualizing model predictions introduces a distinct set of design constraints that are difficult to implement by merely composing existing abstractions for chart construction in the browser (i.e., D3 [21], Vega [174], Vega-lite [173]). Future directions for research in this area include advances in interactive visualization techniques for larger-scale or higher-dimensional data, as well as deeper understanding of how people use visualizations and other statistical tools to evaluate claims about data.

6.2 Design rationale

The study team designed and implemented a `EVM` based on a handful of use cases and design principles for model checks. Our aim was to create a proof-of-concept prototype in order to investigate how model checks might be useful in exploratory data analysis (EDA).

6.2.1 Anticipated use cases

Based on previous research on graphical statistical inference (reviewed in Chapter 2), we anticipate a few types of use cases where model checks would add value to EDA workflows. This list is non-exhaustive, and model checks might be helpful for other tasks not mentioned here (e.g., detecting outliers or imputed missing values).

"Is this pattern real or spurious?" Although visual data exploration is open-ended by nature [7, 14], much research on evaluating VA tools focuses on "insights" as an outcome of analysis (e.g., [227]), often analogizing the process of discovering these insights to null hypothesis significance testing (NHST) [28, 175, 216]. Under this NHST framing, we can think of insights about patterns of interest in EDA as *judgments of whether these patterns are likely to have occurred by chance.* For example, imagine a scenario where we want to

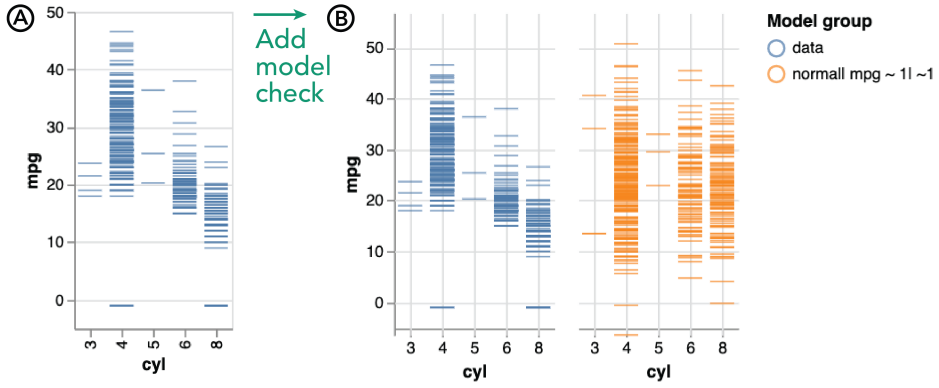


Figure 6.1: Checking a "null model" [28] to assess whether a pattern is spurious: Ⓐ A strip plot showing how miles per gallon varies as a function the number of cylinders in a car's engine. Ⓑ Adding predictions from a reference model to this strip plot using `EVM`. Whereas this is static screenshot, `EVM` animates a new set of model predictions every 500ms as HOPs [98], such that the orange marks seem to dance.

know if the number of cylinders in a car's engine `cyl` influences a car's fuel economy (i.e., average number of miles per gallon `mpg`). In a typical VA workflow, we would assess this relationship by eye (Fig. 6.1 Ⓐ), making a comparison to what we imagine the pattern might look like if it were spurious. With model checks, under the NHST analogy, we might define the likelihood that our pattern of interest occurred by chance relative to a statistical model assuming no influence of `cyl` on `mpg` (Fig. 6.1 Ⓑ). If the likelihood of our data under this "null model" [28] seems very small, we might reject this provisional model as a sufficient explanation for our data generating process (DGP).

Explaining patterns. Upon realizing that a particular explanation for patterns in data is insufficient (i.e., rejecting the "null model"), an analyst's natural inclination is to try to further explain patterns that don't seem spurious [7]. In a typical VA workflow, this might involve exploring different views of the data by adding new variables to the chart specification or by changing encodings. With model checks, *this exploration might be mediated by different choices of reference model*, making increasingly strong assumptions to see whether the resulting predictive distribution is a better or worse match to the patterns in a generated view of the data. For example, we might update our previous "null model" with a new assumption that `cyl` influences `mpg` (Fig. 6.2). This reference model update leads to a noticeable reduction in the discrepancy between patterns in the dataset and our

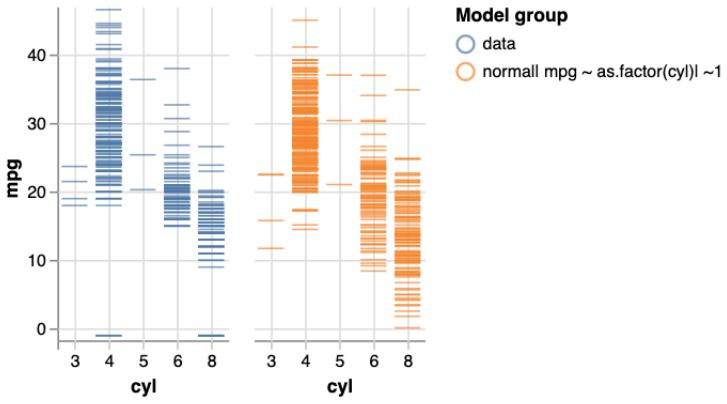


Figure 6.2: An updated model check, following our provisional rejection of the model shown in Figure 6.1 Ⓑ. The updated model assumes that the mean of `mpg` depends on `cyl`.

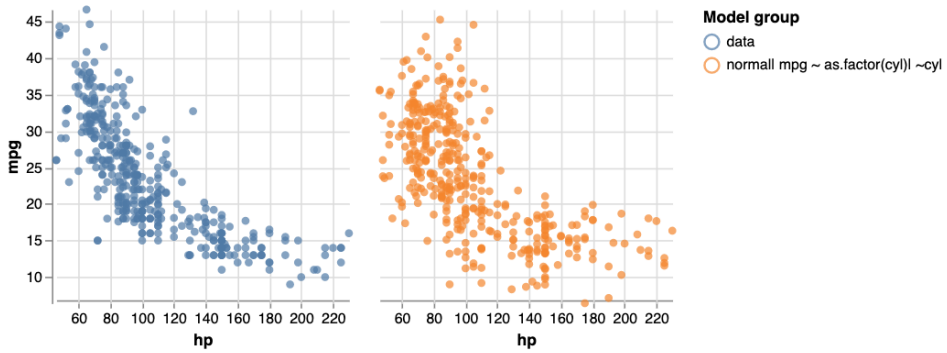


Figure 6.3: A model check showing how `cyl` can help to predict the non-linear trade-off between `hp` and `mpg`. We apply a reference model that we built up by looking at Figures 6.1 and 6.2 in order to explain a related pattern in our dataset.

reference distribution. We might be unable to reject this model under the NHST framing of model checks, although an analyst might still consider these non-rejected explanations to be important or interesting. As we account for more of the apparent structure in our dataset, incremental changes to the model will produce subtler discrepancies between the reference distribution and our dataset. These subtle discrepancies must be judged based on the analyst's domain knowledge and statistical savvy.

Relating disparate patterns. When an analyst feels they have a decent mental model of DGP, they may wish to check how well that model can account for correlations between additional variables in the dataset that they have not yet explicitly modeled. For example,

given the apparent capacity for `cyl` to predict `mpg`, we might want to know if the number of cylinders in a car's engine can also explain the trade-off between `mpg` and variables like horsepower `hp`. In a typical VA workflow, we might try to create complex multidimensional views that give us a sense of the joint distribution of these three variables. With model checks, we can create a simpler scatterplot comparing `hp` and `mpg` and show predictions of what this pattern would look like given our current provisional model (Fig. 6.3). In this way, model checks might facilitate reasoning about correlations in a dataset that remain hidden in simple, low-dimensional visualizations.

In each of these use cases, flexibly expressing and checking provisional mental models of DGP provides an explicit software representation of reasoning that would otherwise be implicit in the mind of the analyst. The remainder of this section outlines a design philosophy for how to support the externalization and interpretation of such mental models.

6.2.2 Design principles for model checks

Model checks require design principles beyond those implicated in visualization and human-computer interaction more broadly. These ideas, derived from informal meetings with statisticians and visualization software builders as well as our study team's expertise, address how `EVM` is intended to support expressing and checking provisional mental models of DGP.

Pattern-seeking, not data mining. Visual analytics tools designed for data exposure wrongly assume that users will know what patterns mean if they see them [12, 111, 224]. We posit that analysts can be easily distracted by small visual details unless they seek these details deliberately based on a mental model of their data. For this reason, `EVM` will avoid dredging up comparisons or views of data that the user has not requested. We may find that lightweight model recommendations are appropriate in future tools like `EVM`, however, given the potential for these recommendations to be misunderstood or poorly integrated with the user's mental model of DGP, we chose to omit automated suggestions and instead require manual model specification in our initial implementation of `EVM`.

Model expansion workflow. Statistical tools should promote a workflow where the user considers and checks changes to their models incrementally [61]. This helps analysts

understand complex models in terms of simpler models that contain a subset of the same assumptions, building this understanding through iterative model checks. `EVM` facilitates this sort of cumulative hypothesis testing about what is and isn't helpful for a model to assume about the data.

Decoupling models from visualizations. Both visualization authoring abstractions—many of them based on Wilkinson's grammar of graphics [218]—and model specification syntax in R [29, 163, 217] share a similar representational form. Both interfaces are composable algebras for flexible declarative specification. At the outside of this project, we considered whether there could be a clear mapping between these algebras, such that chart specifications could be used to derive appropriate model specifications. We approached expert visualization software designers who previously attempted to incorporate modeling into VA software [191],[1] and through this conversation we realized that assuming a correspondence between visualizations and models might be untenable. A given chart entails a set of possible models, and the user must fill in gaps in a tool's ability to automatically infer which assumptions about DGP the user would like to investigate. We posit that tools for model checks require ways for analysts to express their intent to investigate a particular assumption. In `EVM`, we decouple model specification from chart specification, allowing visualizations and models to be "out of sync" such that either a visualization shows more about the data than a model learns or a model learns relationships which are not directly visualized on the chart canvas. This design choice facilitates use cases such as *explaining patterns* and *relating disparate patterns* (see Section 6.2.1).

Smart defaults for model check visualizations. Rather than giving the user control over every aspect of the generated chart's appearance, we attempt to minimize the number of edge cases where the user could create a useless or misleading visualization. Some design choices, such as showing *one mark per data point* by default, follow from empirical findings in visualization research—i.e., that such disaggregated views promote more conservative inferences about the existence of patterns in data [149]. Other design choices serve to facilitate meaningful visual comparisons between patterns in data and model predictions.

[1] Many thanks to Justin Talbot for meeting with us to discuss an old project with Pat Hanrahan aiming to incorporate statistical modeling into Tableau, which remains unpublished.

For example, `EVM` shows observed versus predicted patterns using different colors and in side-by-side facets. This avoids overplotting model predictions on top of the data. `EVM` always facets the position of the model predictions in a way that preserves the ability to *compare observed and predicted outcomes on a common scale*, an important design choice which users might not think of. We realized these smart defaults were necessary after implementing an earlier version of the tool that granted the user more flexibility, and noticing that the number of interactions required to create a "good" visualization were too many.

Discoverable failure modes. For operations that require maximum articulatory flexibility for the user, interfaces should allow the user to discover failure modes and respond to them with agency. The study team sees model specification as one such operation, especially given that `EVM` is a design exploration about how to incorporate model checks into VA tools. Although we provide users with high-quality visualizations by default, when the user chooses to check a nonsense model against that visualization, the resulting visualization becomes ill formed. For example, if the user models a discrete variable using a model that assumes continuous outcomes, the resulting model check visualization will attempt to show continuous predictions on a discrete axis scale. After discovering these failure modes during informal testing, the study team decided that they were an asset, that we should let `EVM` show a bad visualization when the user asks for a model with unreasonable assumptions.

6.3 `EVM` *exposition*

`EVM` is composed of the following modular components linked together to form a single-page web application where users iteratively generate views of data and check models against them. The reader can try out the prototype at `https://kalealex-evm.surge.sh/`.[2]

Chart panel. The chart panel occupies the center of the `EVM` display area (Fig. 6.4 Ⓒ). User-generated views appear in this part of the display and update reactively in response to changes in the state of the application, which are triggered by user interactions with other modular components. The chart panel component contains all of the logic for smart

[2]What I describe in this section is the desired behavior of an interface that is still in development. Although most of the features are implemented and working, I still have a few things to do before handing the tool to potential users.

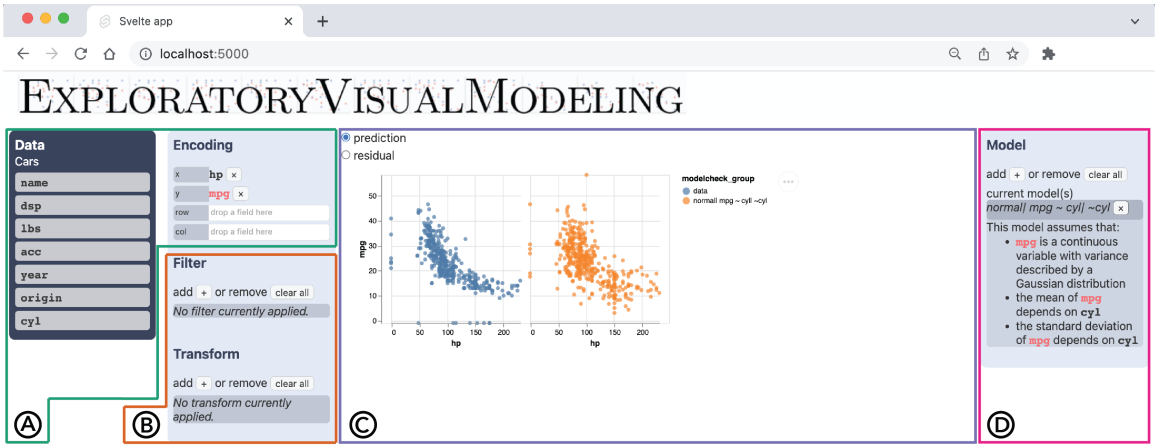


Figure 6.4: A screenshot of `EVM`, annotated to show different components described in Section 6.3: Ⓐ Shelf construction interface; Ⓑ Filters & transformations interface; and Ⓒ Chart panel; and Ⓓ Model bar. The specified model check assumes that the number of miles per gallon a car gets `mpg` is normally distributed with mean and variance both impacted by the number of cylinders in a car's engine `cyl`. By plotting horsepower `hp` against `mpg`, we can check how well the trade-off between `hp` and `mpg` is explained by the predictor in our reference model, `cyl`, which is not shown directly in this view.

defaults such that it determines how the current combination user-specified encodings and models should be displayed.

Shelf construction. The shelf construction interface occupies the far left portion of the display (Fig. 6.4 Ⓐ). It consists of two columns side-by-side: the left column contains variable names from a pre-loaded dataset; the right column contains "shelves" for each possible encoding channel. The user can specify views of data by dragging variables onto a shelf, each representing an *encoding channel*: x-axis, y-axis, row, and column. The x- and y-axis channels define what appears in the main portion of an individual chart. The row and column channels indicate that the user wants a trellis plot [15, 196], consisting of multiple subplots arranged along either the vertical or horizontal span of the chart panel.

Filters & transformations. The interfaces for filtering and transforming data sit beneath the encoding shelves (Fig. 6.4 Ⓑ), and both follow the same design pattern. To add a filter, the user *selects a variable* based on which they will either *include or exclude* values *less than (or equal to), greater than (or equal to), or (not) equal to* a *criterion*. To add a transform, the user *selects a variable* to which they can apply either a *log odds or log* transformation. `EVM` provides warnings when the user applies a transform that will return unusable values,

such as a log transform of values less than or equal to zero. In both the filter and transform interfaces, the working copy of the data is modified to reflect the user's choices, and a backup copy of the data is maintained, which the user can revert to by removing filters and transforms. For simplicity, `EVM` always applies filters before transforms and applies both in the order they are specified. `EVM` refits models after any changes to filters or transforms.

The model bar. The model bar occupies the far right portion of the display (Fig. 6.4 Ⓓ). It follows a similar design pattern as the filter and transform interfaces, in that users can add or remove models from the current set of candidate models. When the user chooses to add a model to the model bar, they *select a distribution family* for their regression model:

- **Normal**: a traditional linear model for continuous outcomes, estimating the conditional mean, with the option to specify separate sub-models for the Gaussian distribution's location and scale.

- **Log normal**: a transformed linear model for continuous outcomes in the domain $(0, +\infty)$, estimating the conditional mean in log units, with the option to specify separate sub-models for the Gaussian distribution's location and scale on a log scale. Predictions of this model are back-transformed into the original outcome scale before being passed from the `modelcheck` R package to `EVM`, so the log link function is encapsulated in the back-end function.

- **Logit normal**: a transformed linear model for continuous outcomes in the domain $(0, 1)$, estimating the conditional mean in log odds units, with the option to specify separate sub-models for the Gaussian distribution's location and scale on a log odds scale. Predictions of this model are back-transformed into the original outcome scale before being passed from the `modelcheck` R package to `EVM`, so the logit link function is encapsulated in the back-end function.

- **Logistic**: a generalized linear model for binary outcomes, estimating the conditional log odds of one vs zero.

- **Poisson**: a generalized linear model for count outcomes, estimating the conditional rate of events.

- **Negative binomial**: a generalized linear model for over-dispersed count outcomes (i.e., a mixture of Poisson distributions with rates drawn from a Gamma distribution), estimating the conditional rate of events, with the option to specify separate sub-models for the location and scale of rates.

Based on the model family the user chooses, `EVM` elicits a *model specification* in Wilkinson-Pinheiro-Bates syntax [163, 217], for some model families providing separate text boxes for location and scale sub-models. `EVM` parses these model specifications into sets of assumptions about DGP and displays natural language descriptions of each reference model's respective assumptions in the model bar. Confirming a new model triggers a post request to the main event handler on the server-side, which makes calls to the `modelcheck` R package.

6.3.1 Implementation

`EVM` is a web application implemented using Svelte. The behavior of `EVM` is mediated by a global state that is shared across modular components, and which is updated and reacted to by each component in real time. We base the interface for `EVM` on PoleStar [220], a research prototype for EDA that uses an interaction model for chart construction based on Tableau, where the user creates charts by dragging and dropping variables onto encoding channels. Our intention is that relying on a popular design pattern for chart construction will facilitate greater ease of use, extensibility, and adoption.

We use Vega [174] to generate visualizations in `EVM`. Chart specifications authored by the user through the shelf construction interface are initially stored using concise Vega-lite [173] syntax. However, some of our design choices around faceting model predictions and showing them in animated HOPs are not supported in Vega-lite. For this reason, we compile the resulting Vega-lite specification to Vega and modify it prior to rendering. This means that `EVM` inherits some defaults and optimizations from Vega [174] and Vega-lite [173], while only trying to support a subset of the visualization design space expressible with these tools.

`EVM` is backed by the `modelcheck` R package, which I built to handle all modeling-related data manipulations. It contains functions to run a specified model in the selected family and add predictions from that model to the input dataframe, calculate residuals for a given model, and merge together outputs from other operations without duplicating entries. I implemented the `modelcheck` package using a combination of base R and `tidyverse` packages [215] for data wrangling, and using fast-fitting `gamlss` [169] models. We expose the `gamlss` model specification syntax (a.k.a., Wilkinson-Pinheiro-Bates notation [163,217]) to the user through the model bar, a design choice which served both as an expedient in the design process (i.e., we didn't have to invent a completely new interaction technique) and as way to make the new design pattern represented by the model bar seem approachable, at least to users familiar with R. The `modelcheck` package provides maximum flexibility for specifying fixed-effects regression models within the scope of a research prototype.

To minimize the required amount of communication between client- and server-side environments, the `modelcheck` package contains a main event handler. When the user specifies a new model to check against data, the current state of the application is passed to the main event handler via a post request. The main event handler determines which of the current set of models have not been run, fits those models, calculates predictions and residuals, and outputs a unified, relatively compact data structure for visualization. This dataflow reduces redundant computation and high-latency interactions.

6.4 Deploying cognitive mechanisms

By design, `EVM` leverages cognitive mechanisms such as *ensemble processing*, *visual heuristics*, and *model-based thinking*—which were the foci of Chapters 3–5 [109–111]—to support more rigorous thinking about uncertainty during data exploration. I posit that cognitive challenges around model-based thinking arise from an unclear mapping between conceptual relationships in a dataset and the low-level visual features that an analyst might notice in a chart. Some design choices in `EVM` represent an attempt to coordinate cognitive mechanisms that mediate processing of such low-level visual cues in order to explicitly connect them to users' mental models of data. In this way, `EVM` is a summative proof-of-concept that brings together the cognitive mechanisms presented in this dissertation.

Visual calibration through ensemble processing. In Chapter 3, I presented evidence that showing animated hypothetical outcome plots [98] (HOPs), representing distributions of possible outcomes under alternative assumptions about DGP, supports greater sensitivity to the true underlying trends in noisy samples of data compared to alternative ways of presenting these reference distributions. The task in this pair of experiments was similar to the judgments `EVM` is intended to facilitate; in both cases, the chart user needs to gauge the compatibility of data with possible DGPs. My colleagues and I argued [110] that HOPs likely support this type of judgment through *ensemble processing*, a cognitive process by which the visual system quickly, automatically, and accurately decodes summary statistics about a set or ensemble of objects in the visual field [3]. Elsewhere [107] and in Chapter 3, I speculate that ensemble processing plays a key role in the calibration of visual inferences. By automatically learning the distributions of visual features conditional on different environments, the visual system provides the biological ingredients for an approximately Bayesian[3] mode of reasoning where incoming signals are weighed against conditional expectations. These learned visual expectations, associated with concepts and contexts through a distributed network of brain activity, drive our sense of what seems normal or surprising [58, 59, 123, 210, 211]. Although the Bayesian nature of the brain is a topic of fierce debate in cognitive science, this framing suggests that we can rely on the ensemble processing mechanism to endow expectations about how the data might look under a given DGP.[4] This is why we use HOPs to show model predictions rather than providing the same information as a static distributional summary (e.g., predictive intervals).

Discrepancy judgments through heuristics. Providing an explicit reference for what data would look like given a particular DGP frames model checks as distributional comparisons. In Chapter 4 [109], I showed that visual reasoning strategies (i.e., heuristics) play a major role in how people compare distributions to make judgments and decisions. Both tasks in that experiment entailed judging the magnitude of a difference between two

[3]Correspondence between empirical evidence (e.g., [18]) and predictions from theory [48] suggests that this calibration process at least resembles Bayesian updating.

[4]The apparent utility of HOPs for helping people compare patterns to reference distributions does not depend on a Bayesian interpretation of this calibration process.

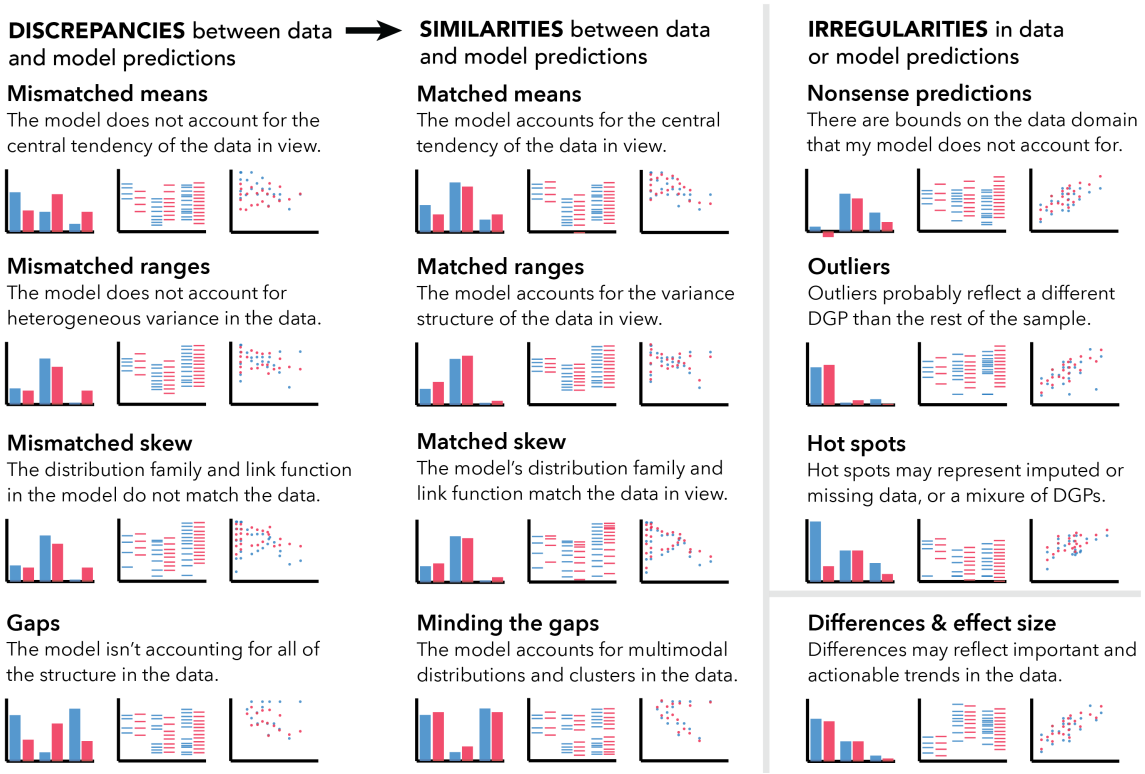


Figure 6.5: A provisional taxonomy of possible discrepancies between data and model predictions that can signal insights about data generating process (DGP) when viewing model checks. The three plots in each panel show mock ups of what each type of discrepancy might look like in a bar chart, a strip plot, and a scatterplot. These are not an exhaustive set of discrepancies. They are based on my experience implementing and using model checks in a statistical workflow. Note that an experienced analyst learns to associate these visual patterns with conceptual claims about the relationship between the data and the assumptions represented in the model.

distributions, which is similar to what EVM users do in some model checks. However, the comparisons between data and reference distributions in EVM tend to be more visually complex, such that the distributions will seldom have an identical shape and there will be multiple visual cues for a mismatch or *discrepancy*. Presumably these multiple visual cues will lead to heterogeneous visual reasoning strategies among users. Some prior work on graphical statistical inference has attempted to isolate the exact heuristics users rely on in order to ensure the statistical rigor of these comparisons [203], and subsequently the authors have argued that this poses a threat to the validity of model checks [202]. I think this view abandons a promising design pattern in search of unrealistically uniform behavior across human perceivers.[5] Instead, I argue that we should accept this heterogeneity of heuristics for the discrepancy between data and model predictions. In my experience with visualizing model outputs, different forms of discrepancy often signal different insights about the specific nature of the mismatch between data and reference distribution (Fig. 6.5). In the design of EVM, we try to promote heuristics that might have more robust correspondence to statistical likelihood of the data given a reference distribution. We do this by showing primarily dissaggredated views of data (i.e., one mark per data point), rather than visually aggregating distributions into intervals or densities which might instead promote heuristics focused on central tendencies or overlaps of distributions. Disaggregated distributions enable users to see precisely where regions of high discrepancy are within the generated view, which can inform iterative improvements to the reference model.

Exploration through model-based thinking. The fundamental goal of EVM is to enable analysts to quickly articulate provisional statistical models that might help to explain their data and to check predictions from these models against observed patterns in the data. Making the reference distribution from a provisional model explicit is a major design implication of the work presented in Chapter 5 [111] (as explained in Section 6.1). The

[5] Signal detection theory [73] (introduced in Chapter 2) suggests that there would be error in individual judgments even if all perceivers relied on the same judgment process or heuristic. This is because the visual system uses inherently noisy signal processing to make judgments that are optimal on the margin given lossy input signals arriving in the retina. Thus, it is unrealistic to ever expect human judgments to be deterministically repeatable in the sense of a traditional statistical test. I see this as a strong reason to reject the analogy between graphical statistical inference procedures like visualization line-ups [28, 216] and null hypothesis significance testing, which is the central project of much research to date on this topic.

model check as a design pattern facilitates iterative refinement of formal models (e.g., in Bayesian workflows [61]). Early statistical theorists like Tukey [197] emphasized that this iterative refinement of understanding through models is well aligned with the objectives of data exploration. Recently, others suggested that models themselves should be considered a primary output of VA tools [8] in part because they promote generative thinking about data. Cognitive neuroscientists implicate this mode of model-based reasoning in creative behaviors like mental imagery [25], which bridge our latent understanding of concepts with our perceptions. By providing a way to quickly articulate and check models, we designed `EVM` make explicit and formal connections between what the analyst thinks about their data and what they see on a chart during data exploration. Critically, this design pattern asks analysts to reframe loose insights [7, 14] about DGP as statistical models with testable predictions, and it externalizes the user's thinking about DGP throughout an analysis session.

6.5 Discussion & future work

We learned a handful of important lessons from the design and implementation of `EVM`.

What makes a "good" model check? One major advance in the study team's thinking about model checks was the concept of a "visualization-adjacent" model, representing a set of assumptions about the variables in the generated view that we are confident in, *but not representing a specific assumption we would like to test.* We propose that such a visualization-adjacent model check should capture what the analyst thinks they know about the data generating process (DGP) up to the current moment in exploratory analysis. Following previous implementations of model checks by statisticians [137, 138], it seems beneficial not to include the relationship the analyst is currently trying to reason about in the reference model because the discrepancy between the data and the resulting reference distribution will indicate whether the relationship or assumption in question might need to be added to the reference model in order to account for the data. This provides a springboard for forward model exploration.

It may be possible for a future version of `EVM` to partially automate the construction of visualization-adjacent models to check by keeping a running record of which assumptions about DGP the user is willing to endorse. This would enable model recommendations to

provide assistance to analysts during data exploration. Prior to designing and building `EVM`, the relationship between a given visualization and appropriate models to check was unclear [191], as described in Section 6.2.2 under "Decoupling models from visualizations".

However, there remain some open questions about potential pitfalls in defining reference models. One such question is whether model checks can be misleading to analysts if they are mis-specified relative to the true DGP. For example, a reference model might wrongly assume homogeneity of variance or might overlook an important confounding variable. In cases like these, model checks could be confusing to interpret if the analyst is overconfident in their reference model. Part of the challenge is that thinking about DGP is cumulative during exploratory analysis, and earlier mis-judgments of DGP might create problems later on. We need to know if a tool like `EVM` should include automations or feedback mechanisms to prevent or caution the user from relying on potentially mis-specified model checks. Such feedback mechanisms might involve using a causal support [77] metric to assess the plausibility of models that are related to the data in view and a set of endorsed assumptions. However, such data mining approaches should still be critically examined through human judgments and domain knowledge. We think of this as searching a *local model space*—"local" in the sense that a tool like `EVM` is not meant for exploring every possible model of the data (e.g., as in multiverse analysis [133, 172, 179, 184]), but rather exploring models that are proximal to the user's current mental model. `EVM` is an initial design exploration about how such tools for traversing model space should be built.

Another question about how to define reference models is more a matter of statistical theory. The typical approach is to frame model checks as comparisons between real data and simulations from a "null model" that makes minimal assumptions about patterns in the data [28, 216]. In contrast, `EVM` fits a model specified by the user to the data and shows predictions from that fitted model. The fitting process by definition minimizes the discrepancy between the patterns in the data and the model predictions, making the visual cues required to judge a model check rather subtle compared to the "null model" approach. Comparing these two approaches to the choice of reference model reveals that the degree to which a model check shows a discrepancy between the data and model predictions mostly depends on *how we implement the assumptions of the model in a sampling procedure*. Theoretically

justifying the choice of implementation will be important for future work on model checks. Our work prototyping `EVM` surfaced this issue as focal to the visual appearance and utility of model checks, whereas previously we viewed this design choice as more of a minor implementation detail.

Toward specialized visualization toolkits for model checks. During the implementation of `EVM` we encountered an issue where, for all but the smallest datasets, superimposing model predictions on top of observed patterns in data led to overplotting such that the reference distribution occluded the data distribution. Originally, we planned to use superimposition because keeping pairs of observed data points and their corresponding model predictions close to each other is ideal for facilitating visual comparisons. However, this was not tractable in practice. What we ended up doing instead was faceting model checks to generate separate subplots for observed data and corresponding model predictions. Although we kept these observation-prediction pairs side-by-side, faceting emphasizes global cues for discrepancies (e.g., differences in shape) rather than local cues for discrepancies (e.g., differences in position on a common scale). Giving users a toggle button to view residuals partially resolves this trade-off by drawing attention to regions of high local discrepancy. However, a more sophisticated implementation of model checks might provide more flexibility to design charts that highlight a particular kind of discrepancy.

Perhaps future tools like `EVM` could provide the user a query interface to declare the kind of discrepancy between data and model predictions that would be meaningful to them in order to assess a given claim or assumption about DGP. We might call this a *declarative intent language for specifying model checks*. Interviews with potential users (see Section 6.5.1) should help us identify a lightweight taxonomy of discrepancies that users tend to look for and how they think of these cues as related to different aspects of DGP.

Implementing model checks in the web browser also revealed that visualizing model predictions introduces design constraints that are difficult to express using the abstractions available in Vega [174] and Vega-lite [173], which are currently the most lightweight way to implement interactive graphics for the web. Some of our design choices—e.g., HOPs, and faceting observed patterns versus model predictions within charts—required bespoke implementations, not so unique that they require the expressive range of an imperative,

low-level specification like D3 [21] but unusual enough not to be supported by a declarative, high-level visualization specification language like Vega-lite [173]. The most appropriate intermediate level of abstraction is Vega [174], which we ended up using. However, Vega best supports modular construction of interaction techniques, whereas the abstractions that we require for a general implementation of model checks resemble commands with constraints—e.g., "Facet observed data versus model checks within each subplot, but align these facets so that the user can compare the outcome variable on a common scale." To work around this difficulty, our implementation contains an edge case for every possible configuration of faceted model checks, rather than providing a more elegant solution.

Future work should aim to create an API designed specifically for creating model checks. Such an API would enable the user to declare constraints or rules about how model predictions in particular need to be treated, provide primitives for sampling from and aggregating distributions in order to generate common uncertainty visualizations, and enable the user to declare annotations to emphasize regions of conceptually-important discrepancy between data and model predictions. Critically, the ability to specify layout constraints and sampling procedures will accommodate use cases more specific to model outputs and uncertainty visualization, and the ability to emphasize specific kinds of discrepancies will incorporate notions of user intent or task into visualization specification grammars.

6.5.1 Future work: Evaluation

In the near future, the study team would like to empirically evaluate whether a tool like `EVM` can help data analysts reason about possible DGPs during exploratory data analysis. We considered running an experiment to test the outputs of `EVM`, different possible model checks. However, it proved difficult to formally define what improved thinking about DGP would look like without presupposing a narrow set of use cases for model checks *a priori*, and without presupposing too much about the user's mental model of data. We opt instead for a more open ended protocol of user testing interviews, which is suitable given that `EVM` is an exploratory prototype.

We plan to conduct think-alouds and semi-structured interviews about `EVM` with real data

analysts, aiming for a typical number of participants in an interview study (i.e., $n \approx 12$). We will recruit primarily from our professional network on Twitter and via email. To be eligible to participate, people must be familiar with VA tools like Tableau and must work with data on at least a monthly basis. Recruiting real data analysts will provide greater ecological validity to any conclusions we might draw about the utility of model checks. The purpose of these interviews will be to elicit expert feedback on the design of `EVM`, to study how analysts might rely on model checks in the context of data exploration, and to investigate what it would take for model checks to become mainstream in VA tools.

Interview protocol. The interview will be 30 minutes of think-aloud and 30 minutes of discussion about `EVM`. We will provide participants with a cleaned dataset to explore in `EVM`. They will have a choice between a handful of popular machine learning datasets, preprocessed to work nicely with `EVM`'s underlying data manipulations.

Participants will be instructed to *think-aloud* as they explore their chosen dataset for 30 minutes. During the think-aloud portion of the interview, the interviewer will engage only to prompt the participant to say what they are thinking, to answer direct questions from participants, and occasionally to ask the participant if they would like to specify a model to check. It's possible that participants, especially those who are unfamiliar with statistical modeling, will need to know certain things about the interface, model notation, or modeling assumptions. The think-aloud portion of the interview will help us identify these stumbling blocks, so we can design for them in future versions of `EVM`. Additionally, it's possible that some participants will not feel comfortable specifying a model check at all without hand-holding. These cases reflect a potential need for automated assistance to help users of `EVM` close the gap between their mental model of the data and a formal statistical model. When this seems to happen, the interviewer will make a note and ask about these instances later in the interview.

The second half of the interview will be a *semi-structured interview*, where the participant and the interviewer discuss `EVM` and model checks in exploratory analysis. The interviewer will follow an interview guide to structure the discussion. The interview guide consists of the following topics or lines of questioning:

1. **Utility of model checks.** In what ways did you use model checks to help you think about the dataset? What specific visual cues on the resulting chart were interesting or helpful to you?

2. **Generative thinking.** Did you find yourself thinking about data generating process? What kinds of assumptions did you make about the dataset? Did using model checks make these assumptions more salient or concrete?

3. **Expressing mental models of data.** Did you have any difficulty using the model bar to express and check provisional assumptions about data? What if anything made it hard to use? What do you think would help a low- or no-code data analyst check their assumptions during data exploration?

In each line of questioning, the interviewer will refer back to specific examples of situations where the participant either created or attempted to specify a model check. The goal of this discussion is to elicit the participant's reflections on what it's like to use `EVM` and potential challenges around adoption.

Qualitative analysis. The entire interview session will be video-audio recorded, and the audio will be transcribed. The interviewer will review each transcript and video, identifying episodes of interest with regard to affordances of model checks in exploratory analysis, potential improvements to the `EVM` interface, as well as user needs and challenges around expressing provisional models. The study team will then analyze these episodes of interest and summarize what participants said in terms of topic themes and design tensions.

6.5.2 Limitations

Large scale and high-dimensional data are open problems in VA tooling that remain unaddressed by `EVM`. For the purpose of creating a proof-of-concept tool, we focused on model checks for relatively small datasets. However, some of our design choices break down at larger scales. For example, HOPs [98] and disaggregated views [149] showing one mark per data point become unwieldy at large sample sizes, due to limitations related to visual crowding [3], human attention, and computer memory. Aggregation is a common approach

to side-stepping these issues of scale, however, aggregated data and model outputs have a fundamentally different meanings as summary statistics than disaggregated data and model predictions. Future work will need to resolve when it is and is not advisable to present aggregated model checks.

Similarly, our choice to limit `EVM` to position encodings (i.e., x-axis, y-axis, row, column) rules out visualization techniques that might be more suitable for higher-dimensional data, where datasets have many variables that can each take on many values. `EVM` only enables viewing a subset of a high-dimensional dataset's features at one time, a limitation of many VA tools. Although model checks carry information about variables that are not currently in view—i.e., a model can make predictions based on a variable that isn't visualized—analysts may still struggle to reason about the complex structure of correlated variables that underlies a particular view. Future work on model checks should more directly investigate whether using model checks to reason about variables that are not in view can solve the high-dimensionality problem, and the ways in which this approach might be error prone.

Although `EVM` provides abstractions for model specification that should be familiar to R users, these abstractions assume too much previous experience with statistical modeling. One goal of this line of work is to help analysts reason about DGP, the presupposition being that any analyst would benefit from generative thinking during data exploration [8], regardless of their familiarity with modeling syntax. Future work should develop novel interfaces for eliciting analysts' provisional beliefs about DGP, ideally providing softer on-ramps to the kind of generative thinking that `EVM` is supposed to facilitate. This will mean addressing fundamental questions about people's mental models of DGP, including how to translate potentially looser or harder-to-articulate hunches into statistical models with testable predictions.

6.6 Summary

I present `EVM` a proof-of-concept tool enabling analysts to express and check statistical models during visual data exploration. `EVM` is an initial design exploration investigating how exploratory data visualization tools can incorporate statistical modeling to facilitate more rigorous and careful thinking about data generating process (DGP). This augments

the typical process of visual discovery with formal procedures for articulating and verifying claims about DGP, elevating visual analytics tools from producing nebulous "insights" to vetting assumptions and providing a more concrete basis for further analysis. `EVM` leverages the cognitive mechanisms presented throughout this dissertation to help analysts reason about discrepancies between data and model predictions and what these discrepancies suggest about conceptual data interpretation. The work I present here sews the seeds of an ambitious research program on tools for visual model exploration and the role cognitive mechanisms in the design of data interfaces for reasoning with uncertainty.

Chapter 7

CONCLUSION

This dissertation demonstrates an approach to visualization research emphasizing the role of cognitive mechanisms in the design and evaluation of tools for reasoning with data. I contribute a series of experiments investigating how visualization design choices promote reliance on various cognitive processes in the mind of the user—whether a layperson or an analyst. My research synthesizes accounts of these mechanisms across literatures in psychology, economics, computer science, and statistics, advances methods for studying these mechanisms, and explores opportunities to deploy these mechanisms through the design of interfaces for reasoning with uncertainty in data.

7.1 Findings & implications

Here, I summarize knowledge gained through the work presented in previous chapters, and I speculate on the far-reaching implications of this work for statistical tools, behavioral science, education, and public discourse.

Calibrating visual inferences. In Chapter 3 [110], I present a pair of experiments investigating the extent to which information formats that promote *ensemble processing* facilitate visual inferences about the underlying trends in ambiguous samples of data. Specifically, we test a design pattern where we show *reference distributions* of what it could look like if there was growth or no growth in jobs added to the economy each month of a year, given an assumed amount of sampling error. We find that showing hypothetical outcome plots or HOPs [98] as a reference for the task leads to better calibrated visual inferences about the underlying trend in samples of jobs data, compared to more common, static uncertainty visualizations such as error bars and line ensembles (Fig. 3.2). I argue [107] that this is because HOPs tap into the visual system's natural mechanisms for automatically calibrating its sensitivity to contextually meaningful signals in the environment, among which mecha-

nisms is ensemble processing or the ability to quickly, automatically, and accurately extract summary statistics from visual scenes [3].

This work establishes HOPs as a preferred visualization technique for helping untrained laypeople reason about what it would look like if a particular trend was present in a dataset. The design pattern of using HOPs to show a reference distribution has a multitude of applications and is consistent with the idea of visual *model checks* proposed by Hullman and Gelman [96]. In Chapter 6, I present `EVM`, a prototype software tool demonstrating how HOPs of reference distributions can be incorporated into tools for data exploration to help analysts check provisional interpretations of data. Another possible application would be helping laypeople assess the claims about data that they encounter online, e.g., by showing HOPs of what it would like if the central claim in a piece of data journalism were true. Both of these applications empower end-users of data visualizations to judge for themselves the potential veracity of claims or assumptions about data, opening up new potential for uncertainty visualization as a tool for healthy skepticism about data.

Designing for satisficing. In Chapter 4 [109], I present a mixed-methods experiment showing that chart users tend to rely on suboptimal *heuristics* [198] when using uncertainty visualizations to make incentivized decisions. Reliance on suboptimal strategies when better ones are available is called *satisficing* [178]. We also see that strategies are often heterogeneous within individuals over time, and that a chart user's reliance on a particular strategy can lead to substantial and predictable biases in decision making.

The facts that satisficing and heterogeneous visual reasoning strategies both seem common with uncertainty visualizations have multiple far-reaching implications. Widespread satisficing with visualization casts doubt on aspirational visualization designs that are optimal in theory, only if the chart user intuits the ideal chart interpretation strategy. This both suggests that applications of visualization may fail to achieve their goals in practice because visualization authors fail to anticipate dominant but suboptimal interpretation strategies, and that explicit education and training on visual reasoning strategies should be developed as a way to promote greater statistical literacy.

Other implications have to do with the way visualization researchers and practitioners have historically conceptualized and measured visualization *"effectiveness"* [140]—the

ability of chart users to extract relevant information from a visualization—and how this has informed the design of software tools for visualization. The visualization community historically measured effectiveness by ranking visualizations based on the average performance of chart users on simple "atomic" tasks in controlled experiments (e.g., [36]). The field has long recognized that this approach is inherently limiting—both because average performance is reductive and because effectiveness ratings are sensitive to the choice of task for evaluation [120]—however, outmoded notions of effectiveness remain the standard metric for popular visualization recommender systems such as the one that drives Tableau. In contrast, my work frames effectiveness as a matter of what visual reasoning strategies a visualization design promotes and which strategies a chart user might rely on, given their background and context. This implies models of effectiveness that have more explanatory power both because they explicitly incorporate notions of strategy and because they suggest new directions for research around what factors influence a chart user's tendency to rely on particular strategies. In reconceptualizing effectiveness, this work opens opportunities for new theories of visualization interpretation and new recommender systems.

The promise of model checks. In Chapter 5 [111], I present a pair of experiments that highlight fundamental cognitive difficulties that seem to make it challenging for people to make causal inferences with visualizations. Recall that by *"casual inference"* I refer to a judgment of the compatibility of a given dataset with a set of plausible causal explanations. Elsewhere in this dissertation, I refer to this behavior as *model-based thinking* or reasoning about the compatibility of data with provisional mental models. The cognitive difficulties that seem to make this judgment challenging are insensitivity to sample size—i.e., people don't become more confident as sample size increases—and insensitivity to evidence that seems to verify an explanation—i.e., people don't know how to weigh evidence matching what they would expect to see if an explanation were true. These remain unsolved problems for visualization, although it is unclear the degree to which poor performance on this task was due to our response elicitation method or the unfamiliarity of our participants (Mechanical Turk workers) with using visualizations to make causal inferences.

One design implication of this work is that it might help to provide an explicit reference of what it would look like if a given causal explanation were true. This is similar to the

design pattern we tested in Chapter 3 [110] and is consistent with Hullman and Gelman's formulation of visualizations as *model checks* [96]. To put these ideas into practice in Chapter 6, I present a prototype visual data exploration tool called `EVM` which enables analysts generate views of a given dataset and to articulate and check provisional assumptions about data generating process (DGP) against those views. Specifically, `EVM` provides HOPs of reference distributions to convey the compatibility of patterns in data with predictions from a *reference model* encoding a set of provisional assumptions. This represents an initial design exploration and proof-of-concept for model checks in visual analytics software.

Model checks hold the promise of a new generation of statistical tools that formalize the design goals of visualizations as assessing a target claim or assumption. Critically, these tools will explicitly represent the model a visualization is intended to help the user check and will render visualizations that tee up valid statistical comparisons in service of this goal. This will improve the rigor of visualization as a tool for asking questions of a dataset and will help chart users guard against erroneous or miscalibrated judgments of patterns in data. Similar to my work on visual reasoning strategies, my work on model checks entails a minor shift in how the visualization community conceptualizes what visualizations do. By making our design goals more formal, we open the door for visualization authoring tools that raise the level of abstraction for specifying a visualization to match the way a person might pose questions about what their data shows.

7.2 Future directions

Through experiments and tool-building efforts, I demonstrate the importance of designing for cognitive mechanisms in data visualization and analysis software. Future work should build on these ideas to create new data science tools which explicitly represent, anticipate, and react to the user's cognitive processes.

Strategy-aware visualization toolkits. The work presented in Chapter 4 [109] highlights the role of visual heuristics in distributional comparisons. One major implication of this work for visualization research is that heuristics and strategies explain substantial variation in performance on inference and decision making tasks. Accounting for these strategies, and what influences a chart user's propensity to rely on them, constitutes a promising new

direction for visualization research towards tools based on more sophisticated predictive models of user behavior.

One way to account for visual reasoning strategies is to explicitly build them into visualization APIs. As grammars for visualization specification (e.g., Vega-lite [173]) raise the level of abstraction from low-level graphical implementation details to higher-level considerations, I propose that these grammars should represent notions of the user's intent, task, and strategy. For example, in Chapter 6, I raise the possibility of a declarative intent grammar for authoring model checks, the core idea being that an API for rendering model checks would benefit from controls indicating what specific visual cues would be meaningful to the user for assessing the compatibility of a dataset with a reference model. For example, one such visual cue would be *gaps* in a data distribution that are not reflected in predictions from a reference model (Fig. 6.5, bottom row), which reveal to us that there may be structure in the data that isn't accounted for by the model. I envision an API where the visualization author could specify that they are looking for gaps, and the resulting visualization would use an algorithm to highlight them. Getting specific about the kinds of visual comparisons we want to promote creates stronger connections between a visualization specification and the chart user's potential cognitive process upon seeing the resulting visualization. As a consequence of this added complexity, such high-level APIs would likely resemble domain-specific languages tailored for a particular type of application (e.g., authoring model checks), rather than general purpose charting toolkits.

An explicit machine representation of possible heuristics and how visualization designs can facilitate them would also help to structure the search space for visualization recommendation systems. Specifically, as proposed in Chapter 4 [109], I envision visualization recommenders could be built on a class cognitive models that represent chart interpretation as a mixture of possible visual reasoning strategies, each implemented as functions operating on visualized information. Such a model would predict the probability of a chart user relying on each of a set of candidate strategies based on relevant characteristics of the visualization design and the user themselves. Creating predictive models of visualization interpretation strategies would require (1) formative work to identify an exhaustive set of candidate strategies for a given visualization design, and (2) a large system of experiments

to learn the weights on candidate strategies under different conditions. The grand challenge of creating such a model would provide a framework in which design knowledge from empirical visualization research is more explicitly cumulative. This model could be responsive to contextual information about the user and their use case scenario in ways that existing models of visualization effectiveness cannot be. The resulting chart recommendations would be highly contextual. Combining the predictions of such a cognitive model with a visualization API capable of tailoring visualizations to support a particular strategy would enable recommender systems that respond to the user's potential goals and cognitive processes.

Interfaces for casual inference. The model check [96] design pattern demonstrated in Chapter 3 [110], and through `EVM` in Chapter 6, could be useful for interrogating possible causal relationships, such as those participants struggled to reason about in the experiments presented in Chapter 5 [111]. However, my research on this topic so far remains somewhat disconnected from the stories an analyst tells themselves about data. For example, in the experiments presented in Chapters 3 [110] and 5 [111], we asked participants to reason about a set of pre-chosen explanations for patterns in data. How would people's data interpretation behavior change if these tasks were more open-ended as they tend to be during data exploration? Our future evaluation of `EVM` (see Section 6.5.1) might address this more directly insofar as participants might pose provisional causal explanations for patterns they find while exploring datasets during a think-aloud protocol. Although these provisional mental models will come from real analysts, our insights into the underlying thought process will remain limited to what we can discover through the analyst's introspection. We need more epistemologically satisfying ways of understanding the nature of people's mental models of data generating process.

One possibility for deepening knowledge of model-based thinking and its interconnections with other cognitive mechanisms (e.g., ensemble processing, heuristics and strategies) revolves around new techniques for eliciting mental models of data. For example, in recent years Hullman, Kim, and colleagues pioneered in approaches to graphical elicitation (e.g., [97,119]), creating novel interfaces that enable people to draw what they believe about data on charts. Perhaps drawing or other interaction techniques involving gesture can serve as a softer on-ramp to externalizing and examining mental models of data, compared to

programming interfaces like the model bar in `EVM` (see Section 6.3). Toward this end, future work should invent design patterns that promote stronger ties between our conceptual understanding of data and the way we expect those concepts to manifest in sensory interfaces. A promising possible direction in this space is to collaborate with researchers studying *human-computer integration* [147], a paradigm in computing that creates interconnections between humans and machines such as by using electrical muscle stimulation to incorporate sketching interfaces with the human body [135, 136]. This might enable new turn-taking interactions between humans and machines that could help to reify connections between conceptual and visceral data experiences through multiple sensory modalities.

Beyond the need for new interaction techniques, we also need to develop clearer guidance on the role of automated diagnostics in data exploration for causal inference. The work presented in Chapter 5 [111] suggests that people may struggle to assess causality with visualizations alone. Even using a design pattern like model checks, analysts may struggle to achieve sufficiently confident or well-calibrated visual inferences for their particular application. When the stakes around an inference are high, such as when resource allocation in an organization depends on the findings of a data science team, more sophisticated modeling and diagnostic checks might be desirable. Statistics used for model comparison such as causal support [77], information criteria [68], or cross-validation [205] should probably play a user-facing role in future tools like `EVM`. However, automated modeling and diagnostics might take too much thinking out of the hands of the user, in line with research on model interpretability suggesting that data scientists are prone to accept generic model explanations at face value [112]. On the other hand, only showing models or diagnostics on the user's request risks that the user will overlook something potentially important. This design tension is at the forefront of much current research in human-computer interaction [83].

All of the preceding ideas operate based on the presupposition that we have a dataset that is collected and cleaned in a way that makes it well suited for causal inference. Most research in data visualization makes these kinds of simplifying assumptions. However, in real world data analysis scenarios, questions about whether a dataset can answer our questions often come up. Research in casual inference frames these kinds of question in terms of whether parameters of interest within a model are estimable or identifiable using a given

dataset [157]. Future work should address how analysts use visualizations to reason about what can and cannot be learned from a dataset. This will entail formalizing the kinds of judgments involved in data collection and curation for causal inference. It will also entail addressing the ways that cognitive mechanisms such as heuristics mediate these judgments, making visualizations both impactful and potentially error prone in applications like data fusion [9], integrating multiple datasets to support causal inference.

7.3 Closing remarks

Visualizations and software mediate most human interactions with data. Central tenets of human-computer interaction tell us that designing these external representations to facilitate interpretations of, and actions in response to, data requires some notion of how these representations engage with human thinking. An approach to visualization research that centers cognitive mechanisms, such as I demonstrate through this dissertation, offers formalisms for reasoning about human behavior that can both guide empirical research and serve as the basis for interfaces that anticipate and respond to patterns of human thinking.

BIBLIOGRAPHY

[1] G Aisch, N Cohn, A Cox, J Katz, A Pearce, and K Quealy. Live Presidential Forecast, 2016.

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

[3] George A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. Trends in Cognitive Sciences, 15(3):122–131, 2011.

[4] John R. Anderson and Ching Fan Sheu. Causal inferences as perceptual judgments. Memory & Cognition, 23(4):510–524, 1995.

[5] Eduardo B. Andrade. Excessive confidence in visually-based estimates. Organizational Behavior and Human Decision Processes, 116(2):252–261, 2011.

[6] Michael Andre, Aliza Aufrichtig, Gray Beltran, Matthew Bloch, Larry Buchanan, Andrew Chavez, Nate Cohn, Matthew Conlen, Annie Daniel, Asmaa Elkeurti, Andrew Fischer, Josh Holder, Will Houp, Jonathan Huang, Josh Katz, Aaron Krolik, Jasmine C. Lee, Rebecca Lieberman, Ilana Marcus, Jaymin Patel, Charlie Smart, Ben Smithgall, Umi Syam, Rumsey Taylor, Miles Watkins, and Isaac White. Election Needles: President, 2020.

[7] N. Andrienko, G. Andrienko, S. Chen, and B. Fisher. Seeking patterns of visual pattern discovery for knowledge building. Computer Graphics Forum, n/a(n/a).

[8] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, and A. Rind. Viewing Visual Analytics as Model Building. Computer Graphics Forum, 37(6):275–299, 2018.

[9] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27):7345–7352, 2016.

[10] J. Baron. Thinking and deciding (4th ed.). Cambridge University Press, 2008.

[11] Nicholas J. Barrowman and Ransom A. Myers. Raindrop plots: A new way to display collections of likelihoods and distributions. American Statistician, 57(4):268–274, 2003.

[12] Carmen Batanero, Antonio Estepa, Juan D. Godino, and David R. Green. Intuitive Strategies and Preconceptions about Association in Contingency Tables. Journal for Research in Mathematics Education, 27(2):151–169, 1996.

[13] L. Battle, R. J. Crouser, A. Nakeshimana, A. Montoly, R. Chang, and M. Stonebraker. The role of latency and task complexity in predicting visual search behavior. IEEE Transactions on Visualization and Computer Graphics, pages 1–1, 2019.

[14] Leilani Battle and Jeffrey Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. Computer Graphics Forum, 38(3):145–159, 2019.

[15] Richard A Becker, William S Cleveland, and Ming-Jen Shyu. The visual design and control of trellis display. Journal of computational and Graphical Statistics, 5(2):123–155, 1996.

[16] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. Psychological methods, 10(4):389, 2005.

[17] Daniel Benjamin, Matthew Rabin, and Collin Raymond. A Model of Nonbelief in the Law of Large Numbers. Journal of the European Economic Association, 14(2):515–544, 2016.

[18] Max Berniker, Martin Voss, and Konrad Kording. Learning Priors for Bayesian Computations in the Nervous System. PLoS ONE, 5(9), 2010.

[19] Steve Blenkinsop, Pete Fisher, Lucy Bastin, and Jo Wood. Evaluating the perception of uncertainty in alternative visualization strategies. Cartographica: The International Journal for Geographic Information and Geovisualization, 37(1):1–14, 2000.

[20] Alexander P. Boone, Peri Gunalp, and Mary Hegarty. Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. Journal of Experimental Psychology: Applied, 24(3):275–295, 2018.

[21] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011.

[22] Mike Bostock. D3.js - data-driven documents, 2012.

[23] Mike Bostock, Shan Carter, Amanda Cox, Jennifer Daniel, Josh Katz, and Kevin Quealy. Who Will Win The Senate? The New York Times, April 2014.

[24] David H Brainard and William T Freeman. Bayesian color constancy. Journal of the Optical Society of America A, 14(7):1393–1411, 1997.

[25] Jesse L. Breedlove, Ghislain St-Yves, Cheryl A. Olman, and Thomas Naselaris. Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. Current Biology, 30(12):2211–2224.e6, 2020.

[26] Willard Cope Brinton. Graphic Presentation. Brinton Associates, 1939.

[27] Quoctrung Bui and Ben Casselman. What the tax bill would look like for 25,000 middle-class families. The New York Times, November 2017.

[28] Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-kyung Lee, Deborah F Swayne, and Hadley Wickham. Statistical Inference for Exploratory Data Analysis and Model Diagnostics. Royal Society Philosophical Transactions, Series A(367):4361–4383, 2009.

[29] Paul-Christian Bürkner. brms: Bayesian Regression Models using 'Stan', 2020.

[30] John Burn-Murdoch. Michigan's covid-19 hospitalizations are rising, 2021.

[31] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. Readings in Information Visualization Using Vision to Think. Morgan Kaufmann Publishers Inc., 1999.

[32] Stephen M Casner and Jill H Larkin. Cognitive efficiency considerations for good graphic design. The Cognitive Science Society, (June):275–282, 1989.

[33] Kathryn Chalone and George T. Duncan. Some properties of the dirichlet-multinominal distribution and its use in prior elicitation. Communications in Statistics - Theory and Methods, 16(2):511–523, 1987.

[34] Beth Chance, Joan Garfield, and Robert DelMas. Developing Simulation Activities to Improve Students ' Statistical Reasoning. In Proceedings of the International Conference on Technology in Mathematics Education, pages 2—-10, 2000.

[35] Patricia W. Cheng. From covariation to causation: A causal power theory. Psychological Review, 104(2):367–405, 1997.

[36] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387):531–554, 1984.

[37] R. Coe. It's the effect size, stupid: What effect size is and why it is important. 2002.

[38] Tom N. Cornsweet. The Staircase-Method in Psychophysics. The American Journal of Psychology, 75(3):485–491, 1962.

[39] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. Visualization and Computer Graphics, IEEE Transactions on, 20(12):2142–2151, Dec 2014.

[40] Michael Correll, Enrico Bertini, and Steven Franconeri. Truncating the Y-Axis: Threat or Menace? In ACM Human Factors in Computing Systems (CHI), 2020.

[41] Paulo Cortez and Aníbal Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. Proceedings of 13th Portugese Conference on Artificial Intelligence, pages 512–523, 2007.

[42] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. 15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008, 2003(2000):5–12, 2008.

[43] Richard Cox. Representation construction, externalised cognition and individual differences. Learning and instruction, 9(4):343–363, 1999.

[44] Geoff Cumming. The New Statistics: Why and How. Psychological Science, 25(1):7–29, 2014.

[45] Geoff Cumming and Sue Finch. Inference by Eye. American Psychologist, 60(2):170–180, 2005.

[46] Peter Cummings. Arguments for and Against Standardized Mean Differences (Effect Sizes). Archives of Pediatrics and Adolescent Medicine, 165(7):592–596, 2011.

[47] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6):1204–1215, 2011.

[48] Michael DeWeese and Anthony Zador. Asymmetric dynamics in optimal variance adaptation. Neural Computation, 10(5):1179–1202, 1998.

[49] Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. Mitigating the Attraction Effect with Visualizations. IEEE Transactions on Visualization and Computer Graphics, 25(1):850–860, 2019.

[50] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. A Task-based Taxonomy of Cognitive Biases for Information Visualization. IEEE Transactions on Visualization and Computer Graphics, 26(2):1, 2020.

[51] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in Non-stationary Environments: A Survey. IEEE Computational Intelligence Magazine, 10(4):12–25, 2015.

[52] Charles R Ehlschlaeger, Ashton M Shortridge, and Michael F Goodchild. Visualizing spatial data uncertainty using animation. Computers & Geosciences, 23(4):387–395, 1997.

[53] Niklas Elmqvist and Philippas Tsigas. Causality visualization using animated Growing Polygons. Proceedings - IEEE Symposium on Information Visualization, INFO VIS, 2003:189–196, 2003.

[54] Beverley J Evans. Dynamic display of spatial data-reliability: Does it benefit the map user? Computers & Geosciences, 23(4):409–422, 1997.

[55] Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R De Ruyter Van Steveninck. Efficiency and ambiguity in an adaptive neural code. Nature, 412(23):787–792, 2001.

[56] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In Conference on Human Factors in Computing Systems - CHI '18, 2018.

[57] Steve Franconeri. personal communication.

[58] Karl Friston. The free-energy principle: A rough guide to the brain? Trends in Cognitive Sciences, 13(7):293–301, 2009.

[59] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B, 364(March):1211–1221, 2009.

[60] Sharma. G. Digital Color Imaging Handbook. CRC Press, 2002.

[61] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(2):389–402, 2019.

[62] Iddo Gal. Adults' Statistical Literacy: Meanings, Components, Responsibilities. International Statistical Review, 70(1):1–25, 2002.

[63] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. Health Psychology, 28(2):210–216, 2009.

[64] Miguel A. García-Pérez. Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. Vision Research, 38(12):1861–1881, 1998.

[65] Joan Garfield. The challenge of developing statistical reasoning. Journal of Statistics Education, 10(3):58–69, 2002.

[66] Andrew Gelman. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. International Statistical Review, 71(2):369–382, 2003.

[67] Andrew Gelman. Exploratory data analysis for complex models. Journal of Computational and Graphical Statistics, 13(4):755–779, 2004.

[68] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24(6):997–1016, 2014.

[69] Gerd Gigerenzer and Ulrich Hoffrage. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. Psychological review, 102(4):684–704, 1995.

[70] Daniel G Goldstein and David Rothschild. Lay understanding of probability distributions. Judgment and Decision Making, 9(1):1, 2014.

[71] Richard Gonzalez and George Wu. On the Shape of the Probability Weighting Function. Cognitive Psychology, 38:129–166, 1999.

[72] Michael F. Goodchild. Geographic information systems and science: today and tomorrow. Procedia Earth and Planetary Science, 1(1):1037–1043, 2009.

[73] D.M. Green and J.A. Swets. Signal Detection Theory and Psychophysics. John Wiley and Sons, 1966.

[74] Sander Greenland and Zad Rafi. To aid statistical inference, emphasize unconditional descriptions of statistics, 09 2019.

[75] Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. Statistical Science, 14(1):29–46, 1999.

[76] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. pages 1–49, 2008.

[77] Thomas L. Griffiths and Joshua B. Tenenbaum. Structure and strength in causal induction. Cognitive Psychology, 51(4):334–384, 2005.

[78] Garrett Grolemund and Hadley Wickham. A cognitive interpretation of data analysis. International Statistical Review, 82(2):184–204, 2014.

[79] S Haroz and D Whitney. How Capacity Limits of Attention Influence Information Visualization Effectiveness. 18(12):2402–2410, 2012.

[80] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. Ranking visualizations of correlation using weber{textquotelefts law. IEEE Transactions on Visualization and Computer Graphics, 20(12):1943–1952, 12 2014.

[81] L Hasher and R T Zacks. Automatic processing of fundamental information: the case of frequency of occurrence. The American psychologist, 39(12):1372–1388, 1984.

[82] Andrew Head, Fred Hohman, Titus Barik, Steven M Drucker, and Robert DeLine. Managing messes in computational notebooks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

[83] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. Proceedings of the National Academy of Sciences, 116(6):1844–1850, 2019.

[84] Mary Hegarty, Mike Stieff, and Bonnie L. Dixon. Cognitive change in mental models with experience in the domain of organic chemistry. Journal of Cognitive Psychology, 25(2):220–228, 2013.

[85] Ralph Hertwig, Greg Barron, Elke U Weber, and Ido Erev. Decisions from experience and the effect of rare events in risky choice. Psychological science, 15(8):534–539, 2004.

[86] Rink Hoekstra, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. Robust misinterpretation of confidence intervals. Psychonomic bulletin & review, 21(5):1157–1164, 2014.

[87] U. Hoffrage and G. Gigerenzer. Using natural frequencies to improve diagnostic inferences. Academic Medicine: Journal of the Association of American Medical Colleges, 73(5):538–540, May 1998.

[88] Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

[89] Robin M. Hogarth and Emre Soyer. Sequentially Simulated Outcomes: Kind Experience Versus Nontransparent Description. Journal of Experimental Psychology: General, 140(3):434–463, 2011.

[90] Fred Hohman, Matthew Conlen, Jeffrey Heer, and Duen Horng Chau. Communicating with interactive articles. Distill, 2020.

[91] J G Hollands and Brian P Dyre. Bias in Proportion Judgments: The Cyclical Power Model. Psychological Review, 107(3):500–524, 2000.

[92] Bjorn Hubert-Wallander and Geoffrey M Boynton. Not all summary statistics are made equal: Evidence from extracting summaries across time. Journal of Vision, 15(2015):1–12, 2015.

[93] Darrell Huff. How to Lie with Statistics. WW Norton & Company, 1993.

[94] Jessica Hullman. Why Authors Don't Visualize Uncertainty. In IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 2020.

[95] Jessica Hullman, Eytan Adar, and Priti Shah. Benefitting infovis with visual difficulties. Visualization and Computer Graphics, IEEE Transactions on, 17(12):2213–2222, 2011.

[96] Jessica Hullman and Andrew Gelman. Interactive Analysis Needs Theories of Inference (Working Paper). 2020.

[97] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall estimation of effect uncertainty. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2018.

[98] Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. PloS one, 10(11):e0142444, 2015.

[99] N Irwin and K Quealy. How Not to Be Misled by the Jobs Report, 2014.

[100] Nicole Jardine, Brian D. Ondov, Niklas Elmqvist, and Steven Franconeri. The Perceptual Proxies of Visual Comparison. IEEE Transactions on Visualization and Computer Graphics, 26(1):1012–1021, 2020.

[101] Zhuochen Jin, Shunan Guo, Nan Chen, Daniel Weiskopf, David Gotz, and Nan Cao. Visual Causality Analysis of Event Sequence Data. IEEE Transactions on Visualization and Computer Graphics, (1):1–1, 2020.

[102] Susan Joslyn and Jared LeClerc. Decisions With Uncertainty: The Glass Half Full. Current Directions in Psychological Science, 22(4):308–315, 2013.

[103] Susan L. Joslyn and Jared E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. Journal of Experimental Psychology: Applied, 18(1):126–140, 2012.

[104] Nivedita R. Kadaba, Pourang P. Irani, and Jason Leboe. Visualizing causal semantics using animations. IEEE Transactions on Visualization and Computer Graphics, 13(6):1254–1261, 2007.

[105] Daniel Kahneman. Thinking, fast and slow. Macmillan, 2011.

[106] Daniel Kahneman, Amos Tversky, B Y Daniel Kahneman, and Amos Tversky. Prospect Theory : An Analysis of Decision under Risk. Econometrica, 47(2):263–292, 1979.

[107] Alex Kale and Jessica Hullman. Adaptation and learning priors in visual inference. In VisXVision Workshop at IEEE VIS 2019, Vancouver, BC, Canada, Oct 2019.

[108] Alex Kale and Jessica Hullman. Adaptation and learning priors in visual inference. In IEEE VIS 2019, VisxVision Workshop Paper, 2019.

[109] Alex Kale, Matthew Kay, and Jessica Hullman. Visual reasoning strategies for effect size judgments and decisions. IEEE Trans. Vis. Comput. Graph., 27(2):272–282, 2021.

[110] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2019.

[111] Alex Kale, Yifan Wu, and Jessica Hullman. Causal support: Modeling causal inferences with visualizations. IEEE Trans. Vis. Comput. Graph., 28, 2022.

[112] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Conference on Human Factors in Computing Systems - Proceedings, pages 1–14, 2020.

[113] Matthew Kay. tidybayes: Tidy Data and Geoms for Bayesian Models, 2022. R package version 3.0.2.

[114] Matthew Kay and Jeffrey Heer. Beyond weber's law: A second look at ranking visualizations of correlation. IEEE Transactions on Visualization and Computer Graphics, 2015.

[115] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In Proceedings of the 2016 ACM annual conference on Human Factors in Computing Systems, 2016.

[116] Mary Beth Kery and Brad A. Myers. Exploring exploratory programming. In 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pages 25–29. IEEE, 2017.

[117] Mary Beth Kery and Brad A Myers. Interactions for untangling messy history in a computational notebook. In 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pages 147–155. IEEE, 2018.

[118] Mel Win Khaw, Ziang Li, and Michael Woodford. Risk Aversion as a Perceptual Bias. 2017.

[119] Yea-Seul Kim, Logan A Walls, Peter Krafft, and Jessica Hullman. A Bayesian Cognition Approach to Improve Data Visualization. In ACM Human Factors in Computing Systems (CHI), 2019.

[120] Younghoon Kim and Jeffrey Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. Computer Graphics Forum (Proc. EuroVis), 2018.

[121] Frederick A.A. Kingdom and Nicolaas Prins. Psychophysics: A Practical Introduction. Elsevier Ltd., first edit edition, 2010.

[122] Donald E. Knuth. Literate programming. Comput. J., 27(2):97–111, May 1984.

[123] Adam Kohn. Visual Adaptation: Physiology ,Mechanisms, and Functional Benefits. Journal of Neurophysiology, 97:3155–3164, 2007.

[124] Elysse Kompaniez, Craig K Abbey, John M Boone, and Michael A Webster. Adaptation Aftereffects in the Perception of Radiological Images. PloS One, 8(10):1–12, 2013.

[125] Elysse Kompaniez-dunigan, Craig K Abbey, John M Boone, and M. A. Webster. Adaptation and visual search in mammographic images. Attention, Perception, {&} Psychophysics, 77(4):1081–1087, 2015.

[126] Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When Does Model-Based Control Pay Off? PLoS Computational Biology, 12(8):1–34, 2016.

[127] Tim Kraska. Northstar: An interactive data science system. Proceedings of the VLDB Endowment, 11(12):2150–2164, 2018.

[128] Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. Cognitive science, 11(1):65–100, 1987.

[129] Allison Yamanashi Leib, Anna Kosovicheva, and David Whitney. Fast ensemble representations for abstract visual impressions. Nature Communications, 7:1–10, 2016.

[130] Sharon Lin, Julie Fortuna, Chinmay Kulkarni, Maureen Stone, and Jeffrey Heer. Selecting semantically-resonant colors for data visualization. Computer Graphics Forum (Proc. EuroVis), 2013.

[131] Raanan Lipshitz and Orna Strauss. Coping with Uncertainty: A Naturalistic Decision-Making Analysis. Organizational Behavior and Human Decision Processes, 69(2):149–163, 1997.

[132] Le Liu, Alexander Boone, Ian Ruginski, Lace Padilla, Mary Hegarty, Sarah H. Creem-Regehr, William B. Thompson, Cem Yuksel, and Donald H. House. Uncertainty Visualization by Representative Sampling from Prediction Ensembles. IEEE Transactions on Visualization and Computer Graphics, 23(9):2165 – 2178, 2016.

[133] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses. IEEE Trans. Visualization & Comp. Graphics (Proc. VAST), 2021.

[134] Zhicheng Liu and John Stasko. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. IEEE Transactions on Visualization and Computer Graphics, 16(6):999–1008, 2010.

[135] Pedro Lopes and Patrick Baudisch. Interactive systems based on electrical muscle stimulation. Computer, 50(10):28–35, jan 2017.

[136] Pedro Lopes, Doăa Yüksel, François Guimbretière, and Patrick Baudisch. Muscleplotter: An interactive system based on electrical muscle stimulation that produces spatial output. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16, page 207–217, New York, NY, USA, 2016. Association for Computing Machinery.

[137] Adam Loy, Lendie Follett, and Heike Hofmann. Variations of Q–Q Plots: The Power of Our Eyes! American Statistician, 70(2):202–214, 2016.

[138] Adam Loy, Heike Hofmann, and Dianne Cook. Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners. Journal of Computational and Graphical Statistics, 26(3):478–492, 2017.

[139] Claes Lundström, Patric Ljung, Anders Persson, and Anders Ynnerman. Uncertainty visualization in medical volume rendering using probabilistic animation. IEEE Transactions on Visualization and Computer Graphics, 13(6):1648–1655, 2007.

[140] Jock Mackinlay. Automating the design of graphical presentations of relational information. ACM Trans. Graph., 5(2):110–141, apr 1986.

[141] Charles F Manski. Communicating uncertainty in policy analysis. Proceedings of the National Academy of Sciences of the United States of America, 2018:201722389, 2018.

[142] Charles F Manski. The lure of incredible certitude. Economics & Philosophy, pages 1–30, 2019.

[143] Richard McElreath. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press, 2016.

[144] K. O. McGraw and S. Wong. A common language effect size statistic. Psychological bulletin, 111(2), 1992.

[145] L. Micallef, P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing to cite this version. IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 18(12):2536–2545, 2012.

[146] Dominik Moritz, Bill Howe, and Jeffrey Heer. Falcon: Balancing interactive latency and resolution sensitivity for scalable linked visualizations. In ACM Human Factors in Computing Systems (CHI), 2019.

[147] Florian Floyd Mueller, Pedro Lopes, Paul Strohmeier, Wendy Ju, Caitlyn Seim, Martin Weigel, Suranga Nanayakkara, Marianna Obrist, Zhuying Li, Joseph Delfa, Jun Nishida, Elizabeth M. Gerber, Dag Svanaes, Jonathan Grudin, Stefan Greuter, Kai Kunze, Thomas Erickson, Steven Greenspan, Masahiko Inami, Joe Marshall, Harald Reiterer, Katrin Wolf, Jochen Meyer, Thecla Schiphorst, Dakuo Wang, and Pattie Maes. Next Steps for Human-Computer Integration, page 1–15. Association for Computing Machinery, New York, NY, USA, 2020.

[148] George E Newman and Brian J Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. Psychonomic bulletin & review, 19(4):601–607, 2012.

[149] F. Nguyen, X. Qiao, J. Heer, and J. Hullman. Exploring the Effects of Aggregation Choices on Untrained Visualization Users' Generalizations From Data. Computer Graphics Forum, 39(6):33–48, 2020.

[150] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons, 2006.

[151] Alvitta Ottley, Evan M. Peck, Lane T. Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A. Taylor, Paul K J Han, and Remco Chang. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. IEEE Transactions on Visualization and Computer Graphics, 22(1):529–538, 2016.

[152] Michael Pacer and Thomas Griffiths. Upsetting the contingency table : Causal induction over sequences of point events. Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci'15), 2015.

[153] Michael D. Pacer and Thomas L. Griffiths. A rational model of causal induction with continuous causes. Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011, pages 1–9, 2011.

[154] Lace M. Padilla, Sarah H. Creem-Regehr, Mary Hegarty, and Jeanine K. Stefanucci. Decision making with visualizations: a cognitive framework across disciplines. Cognitive Research: Principles and Implications, 3(1):29, 2018.

[155] Lace M Padilla, Grace Hansen, Ian T Ruginski, Heidi S Kramer, William B Thompson, and Sarah H Creem-Regehr. The influence of different graphical displays on nonexpert decision making under uncertainty. Journal of Experimental Psychology: Applied, 21(1):37–46, 2015.

[156] Lace M. Padilla, Ian T. Ruginski, and Sarah H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. Cognitive Research: Principles and Implications, 2(1):40, 2017.

[157] Judea Pearl. Causal inference in statistics: An overview. Statistics Surveys, 3(0):96–146, 2009.

[158] Judea Pearl. Generalizing Experimental Findings. 3(July):259–266, 2015.

[159] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science, 29(4):579–595, 2014.

[160] Judea Pearl and Dana MacKenzie. The book of why: the new science of cause and effect. Basic Books, New York, 2018.

[161] K. Pearson. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 186:343—414, 1895.

[162] Ole Peters. The ergodicity problem in economics. Nature Physics, 15(12):1216–1221, 2019.

[163] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, EISPACK authors, Siem Heisterkamp, Bert Van Willigen, and R-core. nlme: Linear and Nonlinear Mixed Effects Models, 2020.

[164] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of International Conference on Intelligence Analysis, 2005(May 2014):2–4, 2005.

[165] Karl Popper. The problem of induction. In The logic of discovery, pages 27–33. Hutchinson, London, 1959.

[166] Kristin Potter, Paul Rosen, and Chris R Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In Andrew M. Dienstfrey and Ronald F. Boisvert, editors, Uncertainty Quantification in Scientific Computing, pages 226–249. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[167] Kevin Quealy and Amanda Cox. The first g.o.p. debate: Who's in, who's out and the role of chance. The New York Times, July 2015.

[168] R. Rensink and G. Baldridge. The visual perception of correlation in scatterplots. Journal of Vision, 10(7), 2010.

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

[170] Daniel M Russell, Mark J Stefii, Peter Pirolli, and Stuart K Card. The Cost Structure of Sensemaking. CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, pages 24–29, 1993.

[171] Joshua I Sanders, Balázs Hangya, and Adam Kepecs. Signatures of a Statistical Computation in the Human Sense of Confidence. Neuron, 90:499–506, 2016.

[172] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.5.0). OSF Preprints, 2021.

[173] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2017.

[174] Arvind Satyanarayan, Ryan Russell, Jane Hoffswell, and Jeffrey Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2016.

[175] Rafael Savvides, Andreas Henelius, Emilia Oikarinen, and Kai Puolamaki. Visual data exploration as a statistical testing procedure: Within-view and between-view multiple comparisons. IEEE Transactions on Visualization and Computer Graphics, pages 1–12, 2022.

[176] Thomas Schäfer and Marcus A. Schwarz. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in Psychology, 10(APR):1–13, 2019.

[177] Yi Shen and Virginia M. Richards. A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. The Journal of the Acoustical Society of America, 132(2):957–967, 2012.

[178] Herbert A Simon. Rational Choice and the Structure of the Environment. Psychological Review, 63(2):129–138, 1956.

[179] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. SSRN, Nov 2015.

[180] Michael E. Sobel. The Analysis of Contingency Tables. In Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel, editors, Handbook of Statistical Modeling for the Social and Behavioral Sciences, chapter Chaper 5 T, pages 251–. Plenum Press, New York, 1995.

[181] I. M. Sobol. Uniformly distributed sequences with an additional uniform property. U.S.S.R. Comput. Maths. Math. Phys., 16:236–242, 1976.

[182] Emre Soyer and Robin M. Hogarth. The illusion of predictability: How regression statistics mislead experts. International Journal of Forecasting, 28(3):712–714, 2012.

[183] David J. Spiegelhalter. Surgical Audit: Statistical Lessons from Nightingale and Codman. Journal of the Royal Statistical Society. Series A (Statistics in Society), 162(1):45–58, 1999.

[184] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science, 11(5):702–712, 2016.

[185] S. S. Stevens. On the psychophysical law. Psychological Review, 64:153–181, 1957.

[186] Mike Stieff, Stephanie Werner, Dane DeSutter, Steve Franconeri, and Mary Hegarty. Visual chunking as a strategy for spatial thinking in STEM. Cognitive Research: Principles and Implications, 5(18), 2020.

[187] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. IEEE Transactions on Visualization and Computer Graphics, 8(1):52–65, 2002.

[188] Elizabeth Stoycheff. Please participate in Part 2: Maximizing response rates in longitudinal MTurk designs. Methodological Innovations, 9:1–5, 2016.

[189] Danielle Albers Szafir. Modeling Color Difference for Visualization Design. IEEE Transactions on Visualization and Computer Graphics, 24(1):392–401, 2017.

[190] Danielle Albers Szafir, Steve Haroz, Michael Gleicher, and Steven Franconeri. Four types of ensemble coding in data visualizations. Journal of Vision, 16(5), 2016.

[191] Justin Talbot. personal communication.

[192] Barry N. Taylor and Chris E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, 1994.

[193] J. Gregory Trafton, Sandra Marshall, Farilee Mintz, and Susan B. Trickett. Extracting Explicit and Implicit Information from Complex Visualizations. 2002.

[194] Konstantinos Tsetsos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. Economic irrationality is optimal during noisy decision making. Proceedings of the National Academy of Sciences, 113(11):3102–3107, 2016.

[195] Edward Tufte. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Conn., 1983.

[196] J. W. Tukey. Exploratory data analysis. Addison-Wesley Pub, Reading, Mass, 1977.

[197] John W. Tukey and Martin B. Wilk. Data analysis and statistics: An expository overview. In Proceedings of the November 7-10, 1966, Fall Joint Computer Conference, pages 695–709, 1966.

[198] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157):1124–1131, 1974.

[199] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. In Utility, probability, and human decision making, pages 141–162. Springer, 1975.

[200] Amos Tversky and Daniel Kahneman. The Framing Of Decisions And The Psychology Of Choice. Science, 211(30):453–458, 1981.

[201] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. Animation: Can it facilitate? Int. J. Human-Computer Studies, 57:247–262, 2002.

[202] Susan Vanderplas. Designing Graphics Requires Useful Experimental Testing Frameworks and Graphics Derived from Empirical Results. Harvard Data Science Review, pages 1–8, 2021.

[203] Susan VanderPlas and Heike Hofmann. Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics. Journal of Computational and Graphical Statistics, 26(2):231–242, 2017.

[204] Lav R Varshney and John Z Sun. Why do we perceive logarithmically? Significance, February:28–31, 2013.

[205] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. J Stat Comput, 27(5):1413–1432, 2017.

[206] John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition). Princeton University Press, 1944.

[207] Jun Wang and Klaus Mueller. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. IEEE Transactions on Visualization and Computer Graphics, 22(1):230–239, 2016.

[208] Jun Wang and Klaus Mueller. Visual Causality Analysis Made Practical. 2017 IEEE Conference on Visual Analytics Science and Technology, VAST 2017 - Proceedings, (October):151–161, 2018.

[209] Elke U Weber. From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events. Psychological Bulletin, 115(2):228–242, 1994.

[210] Michael A Webster. Adaptation and visual coding. Journal of Vision, 11(5), 2011.

[211] Michael A Webster. Visual Adaptation. Annual Review of Vision Science, Nov 1(1):547–567, 2015.

[212] Michael A Webster and Deanne Leonard. Adaptation and perceptual norms in color vision. Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 25(11):2817–2825, 2008.

[213] Sean Westwood, Solomon Messing, and Yphtach Lelkes. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. Journal of Politics, pages 1–38, 2019.

[214] Felix A Wichmann and N Jeremy Hill. The psychometric function : I . Fitting , sampling , and goodness of fit. 63(8):1293–1313, 2001.

[215] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. Journal of Open Source Software, 4(43):1686, 2019.

[216] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. Graphical inference for infovis. IEEE Transactions on Visualization and Computer Graphics, 16(6):973–979, 2010.

[217] G. N. Wilkinson and C. E. Rogers. Symbolic Description of Factorial Models for Analysis of Variance. Journal of the Royal Statistical Society. Series C (Applied Statistics), 22(3):392–399, 1973.

[218] Leland Wilkinson. The Grammar of Graphics. 2005.

[219] Jessica K. Witt. Introducing hat graphs. Cognitive Research: Principles and Implications, 4(1), 2019.

[220] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In ACM Human Factors in Computing Systems (CHI), 2017.

[221] Jo Wood, Alexander Kachkaev, and Jason Dykes. Design Exposition with Literate Visualization. IEEE Transactions on Visualization and Computer Graphics, 25(1), 2018.

[222] Xiao Xie, Fan Du, and Yingcai Wu. A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications. 09 2020.

[223] Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. Illusion of causality in visualized data. arXiv, 2019.

[224] Chi Hsien Eric Yen, Aditya Parameswaran, and Wai Tat Fu. An exploratory user study of visual causality analysis. Computer Graphics Forum, 38(3):173–184, 2019.

[225] Lei Yuan, Steve Haroz, and Steven Franconeri. Perceptual proxies for extracting averages in data visualizations. Psychonomic Bulletin and Review, 26(2):669–676, 2019.

[226] J Zacks and B Tversky. Bars and lines: a study of graphic communication. Memory & cognition, 27(6):1073–1079, 1999.

[227] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. 2018.

[228] Hang Zhang, Laurence T Maloney, and De Lange. Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. Frontiers in Neuroscience, 6(January):1–14, 2012.