Designing Belief-driven Interactions with Data

Yea Seul Kim

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

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Jessica R. Hullman, Chair

Amy J. Ko, Chair

Jeffrey M. Heer

Jevin D. West

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Designing Belief-driven Interactions with Data

Yea Seul Kim

Co-Chairs of the Supervisory Committee:
Professor Jessica R. Hullman
Computer Science, Northwestern University

Professor Amy J. Ko
Information School, University of Washington

One's beliefs play a critical role in interpreting new information and making decisions. However, conventional visualization frameworks rarely consider users' beliefs in designing and evaluating visualizations. The main goal of this dissertation is to formalize visualization interaction in light of beliefs to inform visualization design and evaluation. As a first step, I introduce belief elicitation techniques and evaluate them on whether the elicitation act can impact how much users process data and reason with uncertainty. The multiple controlled studies reveal that the belief elicitation act has positive impacts on people's ability to reason with data and its uncertainty. With the understanding on the effect of elicitation, this dissertation presents a Bayesian modeling approach to understand people's belief updating process during visualization interaction. The analysis demonstrates that people's belief updating process slightly deviates from the Bayesian standard when they examine small data and severely deviates when they examine large data. To mitigate the deviation, I also introduce and evaluate personalized data presentations formulated using one's prior beliefs. By working toward formalizing visualization interaction in light of beliefs, this dissertation sheds light on how designers and researchers can take into account one's beliefs in understanding visualization interactions.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Jessica Hullman. She has helped me grow in many ways over the past five years. She has taught me not only how to define a problem, how to approach it, how to solve it, but also the importance of creativity and persistence in doing research. She eliminated many physical and emotional obstacles so that I could focus on my research, and she was always there for me whenever I need her. I can't express enough how grateful I am to have her as my advisor. She is my true role model.

I would also want to thank the mentors I met at UW, and during my internships. I want to thank Amy Ko for generously sharing insights and ideas about research and teaching from years and years of her experience. Thanks to Jeff Heer for consistently asking essential questions about my work. Thanks to Jevin West for being a continuous source of encouragement. Thanks to Katharina Reinecke for providing warm support and being available for chat and coffee. Thanks to Eytan Adar for being an amazing academic grandfather. Thanks to Mira Dontcheva for sharing her optimism and enthusiasm for research. Thanks to Nathalie Riche for convincing me to believe that if I pursue what I love to do, it will eventually pay off. Thanks to Bongshin Lee for inspiring me to be brave like her. Thanks to Jake Hoffman and Dan Goldstein for letting me experience what the ultimately optimal collaboration would look like!

Many thanks to my lab mates and cohorts, who I shared a numerous amount of time in the last few years. I would like to thank Alex Kale, Lavi Aulck, Annuska Zolyomi, and Annie Yan for taking the time to sit down with me, listening to me, having coffee with me, and being happy and frustrated with me.

Thanks to my parents, brother, and extended family for encouraging me to pursue whatever I dream of. I wouldn't get through this journey without them.

Last but not least, thanks to Pedro, my ukulele buddy, who makes me fearless in exploring the unknown. I will be forever grateful for your unconditional trust in me.

DEDICATION

To my parents, for their boundless love and support

Chapter 1

INTRODUCTION

As society becomes increasingly data-driven, people encounter estimates and model predictions related to various aspects of their life on a daily basis. For example, people regularly encounter an unemployment report (Fig. 1.1 left) that signals the pace of job growth or poll results to predict an election outcome (Fig. 1.1 right).

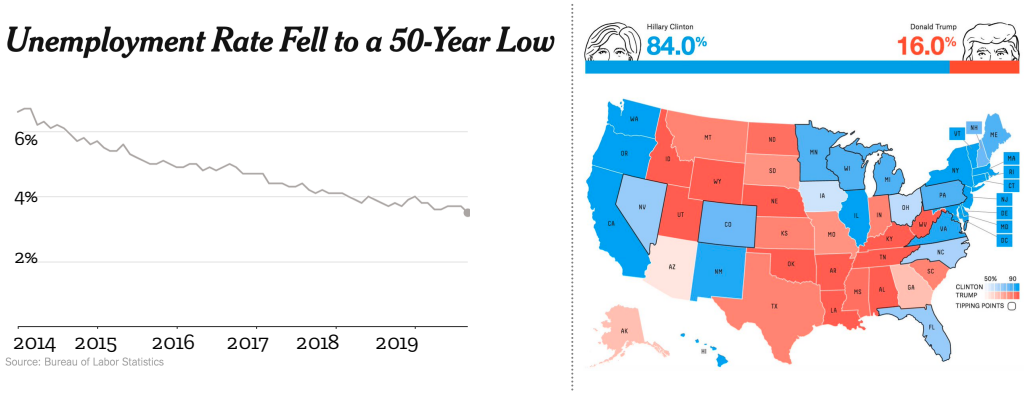


Figure 1.1: Example visualizations that people consume to make sense of various socially relevant phenomena. Left: The visualization describes the unemployment rate over time [34], right: The visualization depicts a poll result from the 2016 presidential election [128].

In addition to understanding socially relevant phenomena, people consume personal data to understand their own status and make better decisions. For example, people make sense of their credit standing by looking at a score calculated from a model (Fig. 1.2 left), and some people make health-related decisions based on predicted menstrual cycles (Fig. 1.2 right).

These estimates are often visualized to inform viewers by engaging with them in ways that text alone cannot. For example, visualizations aid the sense-making pro-

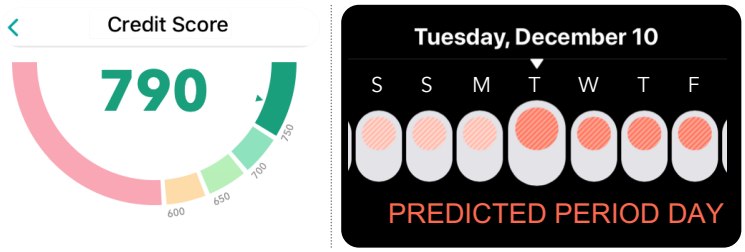


Figure 1.2: Example visualizations that people consume to make sense of their personal status. Left: The visualization shows a credit score (From Mint iOS application), right: The visualization depicts a predicted menstrual cycle for a person (From Health iOS application).

cess by freeing up cognitive resources [25, 93], improve comprehension and retention of information [27, 41, 80, 151], or providing common ground and reasoning artifacts facilitating communication and collaboration [140, 147]. In addition to static views aiding people's cognition, interactivity including filtering, transformation, and view navigation enable user-driven querying of the data.

Visualization research has contributed in many ways to the prevalent use of visualizations in communication and analysis settings. Constructing visualization principles through empirical studies and building tools that operationalize these principles have been a core part of visualization research. For example, people might be familiar with the discouragement of using the rainbow color scheme since it is not effective in visualizing quantitative data. Researchers demonstrate that people do not perceive hue as naturally ordered; thus, the rainbow color scheme leads to error when people decode data from visualizations or compare two extracted values in a visualization [20]. As an example of more comprehensive guidelines in constructing visualizations, A Presentation Tool (APT [100]), which is created by synthesizing results from several studies, provides a recommendation for which visual encodings (e.g., area, hue) to use when a user specifies input data type (e.g., ordinal, quantitative). The tool uses two criteria (expressiveness and effectiveness) to rank different visual encodings for each data type. Effectiveness criteria describe to what

extent a visualization express all aspects of the data, but only the aspects of the data. Effectiveness criteria describe to what extent a visualization support effective perception (e.g., fast/accurate data decoding) of the visualization.

Reflecting on what visualization research has achieved, the predominant paradigm of designing and evaluating visualizations mostly focus on how to present the data accurately by minimizing cognitive and perceptual errors [100, 24, 32, 145]. Perhaps as a result of this emphasis, the evaluation metrics of research focus on how much a visualization supports accurate and fast reading of data, how well people can judge the magnitude between two visual components, and how a visualization can support a specific pre-defined task. The implicit assumption of this paradigm is that factors other than its presentation, such as the user's beliefs or context, don't significantly influence how they interpret the visualizations.

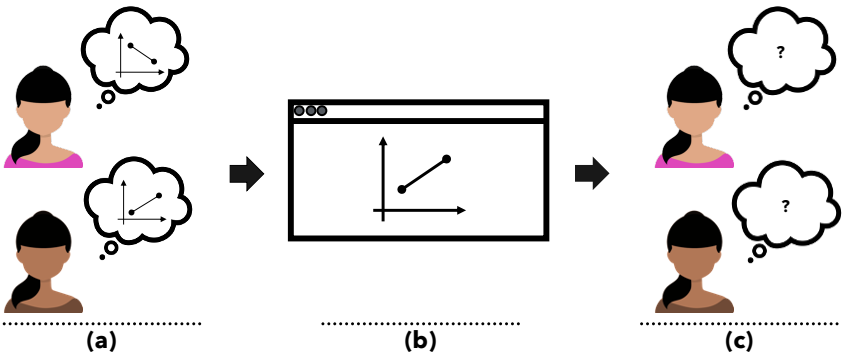


Figure 1.3: A depiction of how I envision to model visualization interpretation process as a belief updating process. (a) People will bring their prior beliefs, (b) in interpreting visualizations. Then (c) they will update their beliefs based on the interpretation.

However, evidence from cognitive science and behavioral economics indicates that how one interprets new information is influenced by one's beliefs, not just how it is presented. To contextualize how we could think of the influence of beliefs in visualization interpretation, imagine a visualization user believes that the median house price in a city will go down in the next two years. Imagine there is another

user who believes that the median house price will go up (Fig. 1.3a). Then they encounter the visualized model's prediction, which indicates the median house price will go up (Fig. 1.3b). Wouldn't we expect their prior beliefs to influence how they interpret visualizations? For example, the user whose prior beliefs conflict with the data might be more uncertain after seeing the data. Their final beliefs would appear to believe a wide range of values (i.e., the percentage changes in house prices) as the user becomes more uncertain about the house price. Whereas the user whose prior beliefs resembled the data might become more confident with their beliefs. Then their final beliefs would appear to believe a relatively small range of values as they become more certain about the phenomenon. It is also expected that their final beliefs are different from each other, and potentially different from the data (Fig 1.3c).

A conventional framework that focuses on accurate delivery of the underlying data would not be sufficient to observe the actual impact that the visualization has on the users. While the idea of exposing interactive features to support task-driven queries is fundamental to interactive visualizations, there have been few attempts to directly incorporate a user's beliefs into their interactions. In the dissertation, I propose an alternative framework to understand visualization interactions by taking into account users' beliefs.

1.1 Approach: Re-thinking Visualization Interaction as a Belief Updating Process

People consume visualizations to understand the underlying phenomenon and form beliefs about the true state of the world, which data proxies for. Therefore designing and evaluating visualizations in light of how visualization impacts people's beliefs will be critical to better support people's visualization interaction.

To build frameworks that systematically takes into account visualization users' beliefs, I address a few relevant research questions.

1.1.1 *RQ1: What are the requirements to design belief elicitation techniques to integrate ones' beliefs into visualization interaction?*

As a first step in considering one's beliefs in the context of visualization interaction, I need to elicit people's beliefs. As people often struggle with reason with data and the associated uncertainty, designing elicitation techniques without confusing people can be challenging. My collaborators and I aim to design a technique that supports easy and natural interaction to lower the bar for users to articulate their beliefs. We also aim to develop techniques that the elicited beliefs are compatible with the underlying visualized data (e.g., numerical beliefs) so that the beliefs can be directly shown with the visualized data or used as a prior when we mathematically model their interaction.

We start by designing a technique that elicits a point-estimate of users' beliefs using graphical and interactive ways. We then further develop interfaces to realistically represent users' beliefs, including how uncertain they are with their beliefs. We devise a technique that allows us to elicit one's beliefs and its uncertainty without prompting them to think about probability.

With carefully designed techniques, we would like to evaluate the effect of belief elicitation in visualization interaction, model their belief updating process, and learn about how they reason with uncertainty.

Through designing techniques, the dissertation contributes to *interactive interfaces* to elicit people's beliefs that allow visualization designers to incorporate people's beliefs into visualization interactions.

1.1.2 *RQ2: What is the effect of eliciting prior beliefs on how much people engage with data?*

Intervening people's interaction by prompting them to articulate their beliefs may have some unexpected effect. Thus we evaluate the effect of the designed elicitation techniques through controlled study in Amazon Mechanical Turk. While some form

of belief elicitation, such as the New York Times's You Draw It series, has been used in practice to help engage users, the effect has not been empirically evaluated. We conduct controlled experiments to demonstrate that eliciting prior beliefs before seeing data have positive impacts on users' ability to recall data.

We also test whether the elicitation act itself can improve uncertainty comprehension, as the technique may help users to be more attentive to data and the uncertainty associated with it. We show the elicitation process can benefit users to guide them to update their beliefs rationally (i.e., match with Bayesian standard) potentially by comprehending the uncertainty of data better.

By conducting these experiments, the dissertation contributes to *empirical findings* that provides contexts when and how elicitation beliefs benefit users to further inform visualization interaction design.

1.1.3 RQ3: What models should we use to understand visualization interpretation in light of a user's prior beliefs, data, and their final beliefs?

While high-level theories in graph comprehension and visualization acknowledge the influence of prior beliefs in visualization interpretations [116, 138], they often stay in the theoretical space. We propose Bayesian frameworks that accommodate one's beliefs in designing and evaluating visualizations in practice. We apply the frameworks to model participants' data collected from experiments and report the observations regarding people's belief updating. We propose quantifications using the Bayesian frameworks that provide insights on how people interpret data.

The proposed modeling approach provides a promising step to understanding the influence of visualization on people's belief changes. By applying a simple Bayesian model to visualization scenarios, the dissertation contributes to *methodology* to observe how people update their beliefs and to define and quantify bias in visualization interpretation.

1.1.4 RQ4: Can we intervene in people's belief updating process to improve users' data comprehension?

Once we understand how people bring their prior beliefs in interpreting data, personalizing visualizations can help improve people's reasoning by mitigating bias in belief updating. We propose and evaluate Bayesian data personalizations aiming at improving uncertainty comprehension and rational belief updating. Through a controlled experiment, the dissertation contributes to *empirical findings* that demonstrates the effectiveness of those personalizations when data are varied in sample sizes and controversiality.

1.2 Dissertation Outline and Findings

This dissertation introduces three research projects that address the above questions.

Chapter 2 presents related work around the role of beliefs in interpreting data, judgment and decision-making under uncertainty, and Bayesian modeling and cognition.

Chapter 3 focuses on designing and evaluating techniques that prompt a user to reflect on their prior beliefs while interacting with visualizations. Inspired by theories in Cognitive and Education Psychology, I introduce a graphically-based technique for eliciting and incorporating a user's prior beliefs about data into visualization interaction. I present the result from a controlled study to evaluate the efficacy of this technique in recalling the data, a common measure of comprehension in Education Psychology. The study shows that participants who are prompted to reflect on their prior beliefs by predicting and self-explaining data outperform a control group in the ability to recall the data. These effects persist when participants have moderate or little prior beliefs on the datasets. I discuss how the effects differ based on text versus visual presentations of data. I conclude the chapter by characterizing the design space of graphical elicitation and feedback techniques and describe design recommendations.

Chapter 4 demonstrates how formalizing visualization interpretation as a belief updating process could inform the design and evaluation of visualizations. The chapter demonstrates a Bayesian cognitive model for understanding how people interpret visualizations in light of prior beliefs and show how this model provides a guide for improving visualization evaluation. The first study shows how applying a Bayesian cognition model to a simple visualization scenario indicates that people's judgments are consistent with a hypothesis that they are making approximate Bayesian inference. The second study evaluates how sensitive our observations of Bayesian behavior are to different techniques for eliciting people's subjective distributions, and to different datasets. The studies show that people don't behave consistently with Bayesian predictions for large sample size datasets, and this difference cannot be explained by elicitation techniques. The final study shows how normative Bayesian inference can be used as an evaluation framework for visualizations, including uncertainty.

Chapter 5 demonstrates the use of a Bayesian framework to guide more rational belief updating by personalizing the visual presentation of observed data using information contained in a user's prior beliefs. I introduce the design of a personalized uncertainty analogy that numerically relates uncertainty in observed data to the user's subjective uncertainty, and a personalized posterior visualization that prescribes how a user should update their beliefs given their prior beliefs and the observed data. The study shows that when a newly observed data sample is relatively small, both personalizations reliably improve people's Bayesian updating on average compared to the current best practice of visualizing uncertainty in the observed data. For large data samples, where people's updated beliefs tend to deviate more strongly from the prescriptions of a Bayesian model, the study shows evidence that the effectiveness of Bayesian personalization may depend on people's proclivity toward trusting the source of the data. I introduce the discussion of how the results provide insight into individual processes of belief updating and subjective

uncertainty, and how understanding these aspects of interpretation paves the way for more sophisticated interactive visualizations for analysis and communication.

Chapter 6 summarizes the contribution and findings of this dissertation and introduces future directions related to belief-driven data interaction.

Chapter 2

RELATED WORK

We aim to build a framework that take into account users' beliefs in designing and evaluating visualization. We review prior work in psychology, education, behavioral economics and other relevant fields to develop the framework for modeling and understanding belief updating process of visualization users, and for constructing design implications.

2.1 Modeling Visualization Interaction

Most well-known studies of information visualization have focused on how comprehension arises from a "bottom-up" perception: spontaneous visual processing of features involving boundary detection, feature integration, color vision, and Gestalt perception, among other mechanisms [32, 145]. On the other hand, cognitive psychologists proposed early models of visualization interpretation, implying that "top down" factors relating to a user's information needs, prior knowledge, and graph literacy affect how visualized data is interpreted, for example, by guiding attention [94, 127]. Studies in graph comprehension provide evidence of such top-down effects [23, 116, 154, 115]. Inspired by the evidence of top-down effects, the dissertation investigates the possibility of incorporating beliefs in understanding visualization interpretation.

Several models are proposed to specifically explain how a person's state of knowledge/beliefs changes over time as they interact with visualizations. A conceptual model created by Van Wijk [139] describes that the user gains knowledge by examining the image (e.g., visualization), and the amount of knowledge that the user gains is influenced by the user's perceptual and cognitive characteristics. Van Wijk

notes that the knowledge gained from a visualization will depend on the prior state of knowledge that a user brings in [139]. Similarly, many others propose models to apply in specific contexts, such as visual analytic scenario [126, 144, 50, 51, 125], or specific user groups [97]. For example, Sacha et al. [126] propose a knowledge generation model for visual analytics where an analyst goes through exploration, verification, and knowledge generation loops. Similarly, Federico et al. suggest a model describing the process of knowledge generation, conversion, and exploitation with the emphasis on the role of prior knowledge in the quality analytical outcome [51]. However, these models often lie in theoretical space. The dissertation aims to formalize the model of visualization interaction by collecting and analyzing empirical data to provide practical implications for visualization design.

2.2 Role of Internal Representations

Several works demonstrate that internal representations of relevant knowledge that one already possesses often play a critical role in reasoning with an external static or interactive visualization [69, 96, 99, 134]. For example, static visualization of processes, which require the use of internal representations to interpret, often outperforms animations [69]. Mayer et al. [103] compared well-designed animations to well-designed texts and diagrams for teaching, finding that performance on retention and transfer tests was better among the static media group that relied on internal representations to understand how concepts related.

Similarly, Hegarty et al. [71] found that viewers who initially engaged in mental animation of a set of static views, then used an external animated visualization, understood the content better than those who used only the animated visualization. The authors propose that the mental animation exercise induced the viewers to articulate their intuitions about the process. As a result, when a participant later viewed the external visualization, they could compare the two representations and note parts of the process where they had lacked sufficient prior knowledge. The

implication of these findings is that interactive information visualizations could be more effective if informed by internal representations for a given data domain citehegarty2004. However, internal representations are rarely explicitly considered in visualization research and development. While Liu et al. provide theoretical explanations of the interactions between internal representations and visualizations, specific to information visualization context [99], the visualization community lacks practical methods to incorporate one's internal representations into visualization interaction to gain the benefit of using it.

Other studies suggest that *externalizing* one's internal representations leads to better understanding of visualized information [40, 72, 111, 130]. For example, Stern et al. [130] found that individuals who had to construct a line graph of presented stock data were more accurate on transfer questions involving a new problem with a similar structure than participants who passively viewed a chart of the data. Constructing external representations is believed to help individuals to translate information between different representations, resulting in a more nuanced understanding of the concepts [42, 130].

An interesting theory in education named "Self-explanation" evaluates the effect of explicit externalizations in learning. Self-explaining information to oneself is a constructive "meta-cognitive" learning activity in which a person actively reflects on the mechanism behind a given phenomenon [30]. Self-explaining has been elicited by having learners explain the meaning of sentences [29, 30, 31], or diagrams [4], as they study a target domain. When quality self-explanations are generated, such as statements that link concepts from the text using tacit knowledge or attempt to fill in gaps through inferences [29], comprehension tends to be better than without self-explanation [29, 30, 31]. A mental model repair hypothesis is proposed to describe the benefits of self-explaining: by generating inferences to fill in missing information, integrating new information with prior knowledge, and monitoring and repairing faulty knowledge, learners who self-explain develop more accurate internal

representations of a concept [29]. In this hypothesis, it is assumed that learners engage more with the process if they identify the discrepancy between their mental model and the presented information [29]. While students who spontaneously generate self-explanations perform better [30, 52, 117, 121, 122], most learners tend not to naturally self-explain [16, 124]. However, even simple prompts [31, 16, 124] have been found to be effective at triggering the self-explanation process [31]. Other studies have shown that immediate feedback on the accuracy of self-explanations can prevent wrong inferences in the explanation process [37, 7, 28].

2.3 Judgment and Decision-making under Uncertainty

2.3.1 Understanding Bias in Judgment under Uncertainty

It is a well-documented challenge that people often struggle to incorporate uncertain information to make judgments. To learn how belief updating process is influenced by people's ability to process uncertain information, we survey research in judgment and decision-making, specifically, those demonstrate how human judgments under uncertainty can diverge from statistical accounts. For example, belief in the law of small numbers describes how many people are too confident in the representativeness of small samples [137], while non-belief in the law of large numbers refers to how many people are not as confident as they should be in estimates based on large samples [15]. When asked to provide their subjective probability of the truth of statements or the values of a measurable quantity, people often display overconfidence [55, 137] or interpret uncertain information in a way that is consistent with their existing beliefs [46].

2.3.2 Supporting Judgment under Uncertainty with Visualizations

To present the uncertainty for decision making, visualization research around uncertainty communication has proposed many techniques for visually representing quantified uncertainty to improve judgments or decisions [14, 43, 48, 118, 148, 149].

Researchers have proposed and evaluated visualizations intended to be accessible to non-expert viewers, including frequency representations of probability visualizations [53, 79, 85, 86, 38, 39]. Some techniques, such as visualizations of probability density functions (e.g., violin or gradient plots [38]) and frequency-based representations of probability like hypothetical outcome plots and quantile dot plots [53, 73, 79, 85, 86] have been found to improve uncertainty comprehension over more conventional displays like error bars among lay audiences. The dissertation offers evidence on how these uncertainty visualizations can help people's belief updating, and further proposes personalized interventions to improve belief updating. Until recently, however, research on the role of visualizations in promoting Bayesian reasoning was limited to studying how visualizations of classic discrete Bayesian reasoning problems like the mammography problem [109, 114, 58, 136, 113, 60, 35, 36].

Rather than accounting for quantifiable uncertainty, recent work by McCurdy et al. [104] suggests that implicit errors represent how users "mentally adjust" data-driven estimates in interpretation, yet cannot be mathematically accounted for given the subjective nature of these beliefs. To accommodate factors that implicitly impact people's beliefs, we vary the underlying characteristics of datasets when we test our modeling approach and personalized presentations to observe the influence of the factors on belief updating. For example, we vary the controversiality of the topic and report findings on how the factor influence people's belief updating.

2.4 Bayesian Cognition

In cognitive science, Bayesian statistics has proven to be a powerful tool for modeling human cognition [64, 133]. In a Bayesian framework, individual cognition is modeled as Bayesian inference: an individual is said to have implicit beliefs about the world ("priors"); when the individual observes new data, their prior is "updated" to produce a new set of beliefs which account for the observed data (this new set of beliefs is referred to as the "posterior"). The prior is formalized as a probability dis-

tribution and Bayes' rule is used to obtain the posterior from the prior distribution and the likelihood function that the observed data is derived from.

This approach has been used to model many aspects of human cognition at various levels of complexity, such as object perception [87], causal reasoning [131], and knowledge generalization [132].

Griffiths and Tenenbaum [66] compared people's predictions for a number of everyday quantities to the predictions made by a model that used the empirical distribution as a prior (e.g., for human lifespans they used a model with a prior calculated from historical human lifespan data). The study found that in aggregate, people's judgments closely resembled the normative Bayesian posterior.

Mathematical psychologists have shown how Bayesian models of cognition help explain a range of perceptual and cognitive phenomena, such as inferring causal relationships or learning categories (e.g., [64, 133, 87, 131, 132]). For example, Griffiths and Tenenbaum [66] demonstrate that the aggregate posterior belief distribution across people approximates the normative Bayesian posterior over various "everyday quantities" such as cake baking times and human lifespans.

2.4.1 Approximate Inference and Sampling Behavior

While Bayesian models of cognition have seen wide applications, the idea that human cognition is accurately described as Bayesian inference is inconsistent with previous influential findings in cognitive psychology from authors such as Tversky and Kahneman [2]. Tversky and Kahneman found evidence that humans often use simple heuristics in their decisions, and that these heuristics lead to sub-optimal judgments. More recent research indicates that heuristics are adaptive and often lead to accurate judgments (e.g., [61]). A recently proposed explanation, which reconciles the seemingly opposing findings between Bayesian models of cognition and the idea that heuristics lead to non-optimal judgments, is motivated by Bayesian cognition [65]: what if human cognition is not *exact* Bayesian inference, but instead

is *approximate* Bayesian inference? One such approach proposes that while people have a prior probability distribution that encodes their beliefs, they do not form judgments using the entire distribution at once [143]. Instead, they take a small number of samples from the distribution, and reason with these samples instead of the full distribution (we which refer to as *sample-based Bayesian*). Being a sample-based Bayesian can lead to sub-optimal individual inferences, but in aggregate, it can produce results very similar to exact Bayesian inference. If an individual's full internal distribution is not readily available to articulate, elicitation techniques that we design should accommodate it. In this dissertation, we present and evaluate a sample-based elicitation technique to reduce elicitation noise, inspired by the sample-based Bayesian hypothesis.

2.4.2 Application to Data Visualization

Recent work by Wu et al. [152] explored the application of the Bayesian framework to examine how people update their beliefs when viewing visualized data. However, Wu et al. prompted participants to internalize a provided prior, show them the observed data, and then ask for their posterior beliefs. Using a fixed prior is not ideal in cases where participants' pre-existing beliefs about a phenomenon will impact their ultimate beliefs. This dissertation demonstrates how to elicit and model participants' personalized prior beliefs for a more realistic application of Bayesian inference, including proposing and evaluating multiple elicitation techniques.

Chapter 3

EXPLAINING THE GAP: VISUALIZING ONE'S PREDICTIONS IMPROVES RECALL AND COMPREHENSION OF DATA

To model people's beliefs, interfaces that can collect people's beliefs are needed. Before we dive into modeling visualization cognition, this chapter introduces elicitation techniques that can be integrated into visualization interaction (Research Question 1) and evaluate potential effects of elicitation while people are interacting with visualizations (Research Question 2).

3.1 Incorporating Prior Beliefs into Visualization Interaction

Most visualizations do not provide ways for users to explicitly incorporate their internal representations. The recent New York Times interactive visualization "You Draw It" [5] is a rare exception that attempts to prompt reflection by enabling a user to explicitly incorporate their prior beliefs. The interactive visualization asks viewers to draw their expectation of the data before presenting the observed data alongside their predictions. To prompt reflection on the difference, the interface provides feedback based on the accuracy of a user's expectation. However, it remains unknown whether the reflection on prior beliefs induced by such visualizations can positively impact recall and comprehension of data, or how to design such visualizations for maximal benefit.

We expand on prior work with five contributions: We contribute a set of novel elicitation techniques for eliciting users' prior beliefs in visualization interaction. These include a graphical prediction technique for eliciting users' predictions of data, a feedback technique for presenting personalized feedback on the gap between predictions and observed data, and a self-explanation prompt to explicitly ask par-

ticipants for self-explanations, which have been found to improve learning from texts and diagrams [4, 30, 31, 29], among others.

We contribute a controlled experiment to test the effect of these techniques on recall and comprehension of data. We find that prompting participants to first predict data, or to self-explain presented data, or to do both, improves data recall and comprehension.

By further contributing replications of our controlled study with datasets that differ in familiarity, we find that these techniques improve recall for datasets for which participants have moderate or little prior beliefs.

We also evaluate how the impact of the techniques differs based on whether the information is presented using text or information visualization. We find that the visualization conditions benefit from predicting the data and viewing the gap between their prediction and the observed data whereas the text conditions do not.

Finally, we contribute a characterization of the design space of data prediction and feedback techniques for information visualization and provide practical design recommendations.

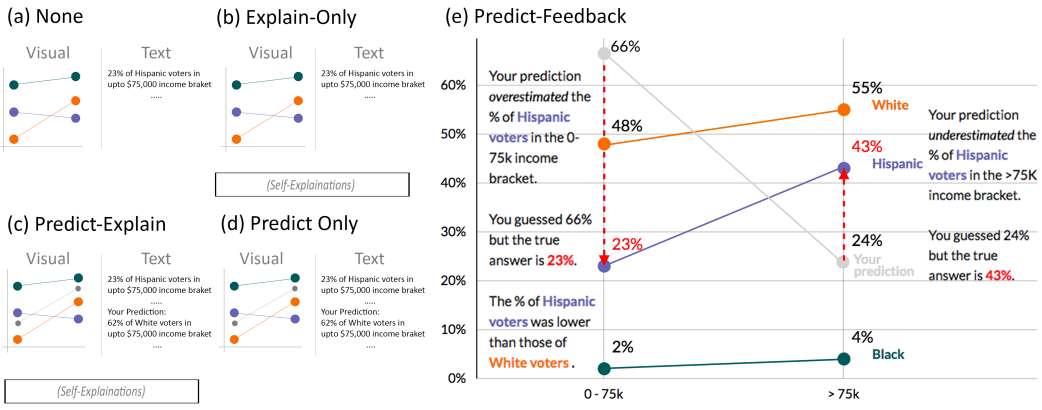


Figure 3.1: The study interface for experimental visual conditions.

3.1.1 Formulating Study Conditions and Hypotheses

Based on the previous research in psychology and education, we devise and evaluate several elicitation techniques for information visualization with associated hypotheses.

Study Conditions: Elicitation Techniques

Our techniques are based on three non-mutually exclusive mechanisms for eliciting reflection on prior beliefs and its relationship to presented data in a visualization.

1. Prompting a user to generate self-explanations of the observed data: In a digital setting, prompting a user to type in sentences explaining the data is an *explicit* way to elicit self-explanations [8].

2. Prompting a user to predict the data before seeing it: Asking a user to predict the data has two advantages for prompting reflection on prior beliefs: (1) Prior work showed that asking a user to actively construct an external representation of their prior beliefs about data results in a deeper understanding of the meaning of a dataset and its visual representation [42, 130]. Predicting may also trigger self-explanation [42].

(2) By asking the user to provide predictions of the data, an interface can then *visualize the gap* between the user's expectations and the observed data. Reviewing the gap may *implicitly* prompt self-regulated learning, in which the user becomes motivated to generate inferences to repair their beliefs [108].

3. Providing the user with feedback on their prediction: Providing direct feedback on the gap between the user's prior beliefs and the observed data may increase the likelihood that a user will recognize the gap and generate inferences to repair their beliefs [37, 7, 28].

Using these three mechanisms, we designed four elicitation techniques and one baseline condition. The observed data is presented using visualization in the visual conditions, and using text in the text condition.:

- **None (baseline)**: The user is prompted simply to examine the observed data (Fig. 3.1(a)).

- **Explain-Only**: The user is prompted to type self-explanations in a text box as she views the observed data (Fig. 3.1(b)).

- **Predict-Explain**: The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After their prediction, the user is shown the observed data against their prediction. She is prompted to type in self-explanations about the gap between their prediction and the observed data in a text box (Fig. 3.1(c)).

- **Predict-Only**:The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After their prediction, the user is shown the observed data against their prediction (Fig. 3.1(d)).

- **Predict-Feedback**: The user is shown the observed data, but with some of the data omitted. The user is prompted to predict the omitted data. After their prediction, the user is shown the observed data against their prediction. Textual and visual feedback are presented to annotate the difference between their prediction and the observed data to draw their attention to the gap (Fig. 3.1(e)).

Hypotheses

Since explicitly prompting self-explanations improves comprehension of information in texts and diagrams [4, 31, 30, 29], we expect a similar beneficial effect for data in visualizations:

H1: Participants in the Explain-Only conditions will recall data more accurately than participants in the None condition for visual and text modalities.

Based on the implicit prompting toward reflection by predicting, we expect that:

H2: Participants in the predict conditions (Predict-Explain, Predict-Feedback, Predict-Only) will recall data more accurately than participants in the None condition for visual and text modalities.

While reviewing their predictions and the observed data, participants in the text conditions must actively *seek and infer* the gap as opposed to the visual conditions where the gap is visually available [4]. We therefore expect that effects of predicting will be less pronounced in the text conditions compared to the visualization conditions:

H3: The effects of predicting using text (Predict-Explain-Text, Predict-Only-Text) on recall will be smaller than the effects of predicting using visualizations.

3.2 Preliminary Survey: Choice of Datasets

To select a dataset for our main study, we conducted a preliminary survey on Amazon's Mechanical Turk (AMT). We sought a dataset with properties amenable to our elicitation techniques. If a user is extremely familiar with a dataset, prediction or explanation may not offer benefits over their prior beliefs. If a dataset is too unfamiliar, it may be too difficult for a user to make predictions about it. Additionally, our interest in comparing techniques across visualization and text modalities is best supported by a dataset that includes both a higher level relational structure (i.e., trends, which visualizations are often best at depicting) as well as individual data points, which can be remembered with greater numerical accuracy from text [81]. By quantitatively measuring the familiarity of multiple datasets, the preliminary survey also enables us to later test the robustness of our results across datasets of varying familiarity.

We selected datasets with a range of topics from the results of a scientific experiment to the average smart phone price from different manufacturers (Fig. 3.3).

The datasets had the same format, consisting of two categorical dimensions and one continuous measure, resulting in six total data points (Fig. 3.2). This format

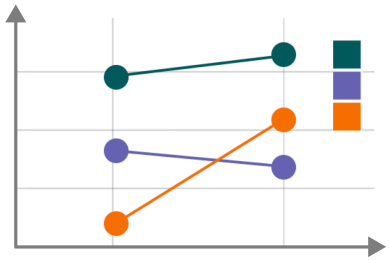


Figure 3.2: Multivariate data format to allow value and trend estimation.

is commonly used in the social and physical sciences [26] and allows us to evaluate how well people observe and recall the higher level patterns (e.g., trend lines) as well as how well people observe and recall individual data points. We formulated three measures that approximate prior familiarity with each dataset:

Perceived familiarity: How familiar participants perceive themselves to be with the data after seeing a short description and a visualization with labeled axes but without data points.

Value familiarity: The absolute error of participants' predictions for each data point. We calculated the absolute difference between the participant's prediction of each data point and the observed data. We normalized the values by dividing by the maximum value on the *y*-axis to allow for comparison of the value familiarity across the datasets.

Trend familiarity:

The difference between participants' predicted slopes and the true slopes for each line in the visualization. We calculated the absolute difference between a participant's slope and the true slope for each of the three groups in each dataset.

3.2.1 Procedure

The survey consisted of two parts. In the first part, we examined *perceived familiarity* by presenting participants with 15 visualization-pairs (resulting from the pairwise

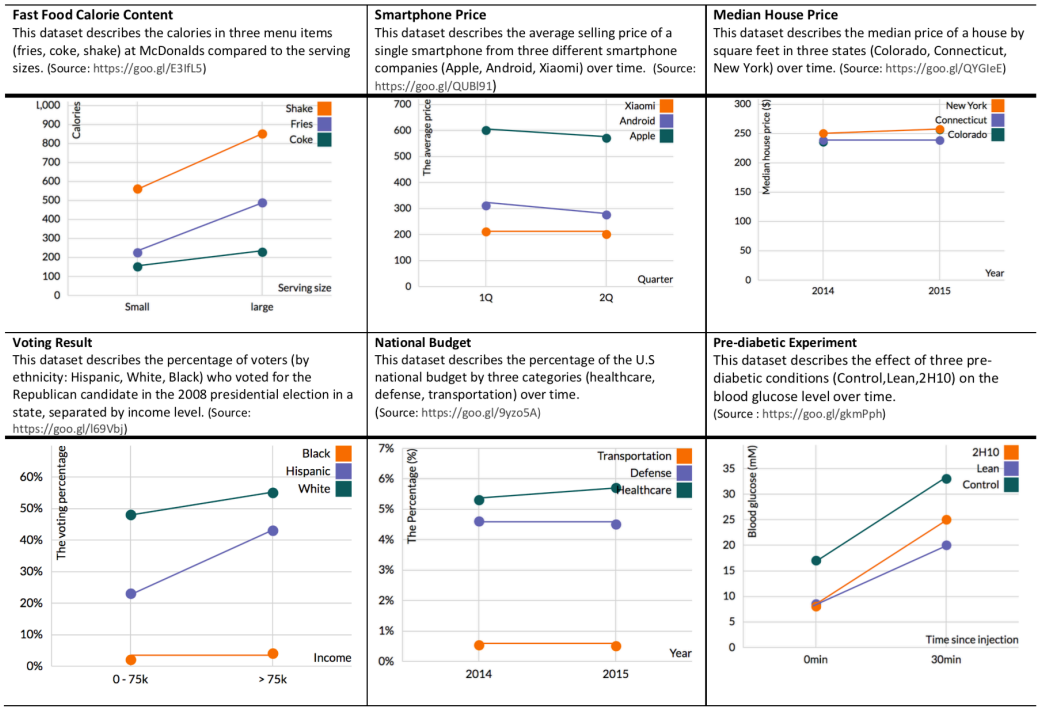


Figure 3.3: The original datasets used in the preliminary survey.

combinations of the six datasets), one at a time in a randomized order (Fig. 3.4).

The visualizations did not reveal any data; instead, participants only saw labeled axes with a short description of each dataset. Participants were asked to select the dataset that they were more familiar with using a radio button. After watching a short tutorial video on how to enter a data point in a chart, participants were asked to predict the values of all six data points for each of the six datasets (Fig. 3.5). Each dataset appeared on a separate screen in randomized order. Each screen presented an empty chart area with labels. Three buttons appeared to the right of the chart labeled with each of the three groups for that dataset (e.g., french fries, coke, shake).

The range of the *y*-axis of the visualizations was set to [0, 1.2*max_data_value]. To input a prediction, participants selected a group by clicking the button for that group, then clicked the chart area to set the position of the two data points for that

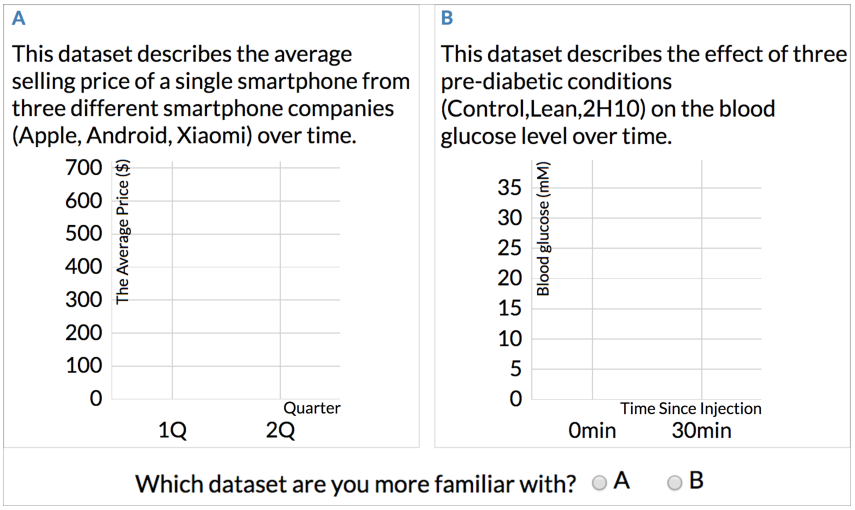


Figure 3.4: A pair of visualizations with the data omitted. Participants were asked to choose which dataset is more familiar based on labels.

group. Participants could drag the circles to adjust their prediction for each group.

3.2.2 Results

We recruited 100 workers from AMT. To calculate the perceived familiarity of each dataset, we summed the number of votes per dataset from the first part of the study.

Table 3.1 shows the ranking of the datasets by these three measures. While perceived familiarity and value familiarity are highly correlated (Spearman's $\rho = .84$, $p < .001$), perceived familiarity and trend familiarity are more weakly correlated (Spearman's $\rho = .42$, $p < .001$). We see evidence of this in the dataset on median house prices, where participants' perceived familiarity is relatively low (rank 5), but they do well on the prediction tasks (rank 1 for predicting trend). We suspect that various factors might contribute to a difference between perceived familiarity and prediction accuracy. In addition to having prior beliefs specific to the dataset, heuristics may allow participants to guess reasonably accurately because they have some beliefs of the domain. In particular, domain specific beliefs (e.g., the average

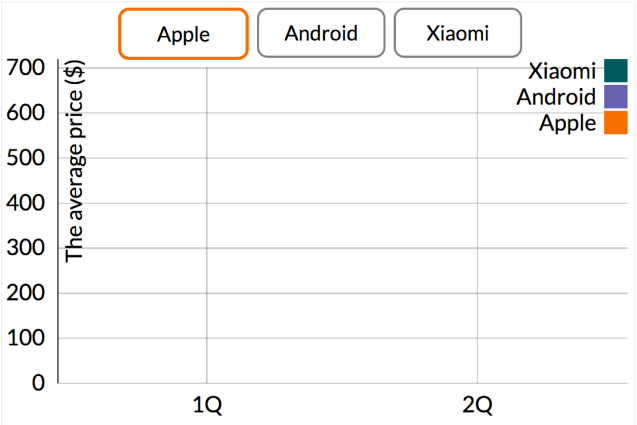


Figure 3.5: Example of a prediction interface.

price of houses is cheaper in Colorado than those in New York), and domain general beliefs (e.g., prices tend to go up over time) may allow participants to make reasonable guesses even when they feel they have little expertise on a topic.

We aggregated the three familiarity rankings and sorted the datasets by the aggregated familiarity (the order of the Table 3.1).

For our main study, we chose the Colorado voting results data, since this dataset was neither clearly familiar nor unfamiliar to participants. To ensure that recalling the data is sufficiently challenging in our main study, we included two visualizations of voting results, for Connecticut and Colorado. We used the more familiar fast food calorie content dataset and the more unfamiliar pre-diabetic experiment dataset to later check the robustness of results from our main study in partial replications.

3.3 Study Design

3.3.1 Study Objectives and Experimental Conditions

To understand how different techniques for eliciting prior beliefs with visualization impact data recall and comprehension, we designed a between-subjects factorial study. Participants were assigned to a baseline condition (**None**) or one

Table 3.1: Three familiarity measures for the six datasets. The order of rows in the table is by the mean rank across the three measures. The number indicates the ranking of the dataset with the actual measure in parentheses. Perceived familiarity votes are out of 500 (100 workers * 5 maximum votes per dataset).

	Perceived Familiarity	**Value Familiarity**	**Trend Familiarity**
Fast Food Calorie Content	1 (348)	1 (0.14)	2 (0.11)
Smartphone Price	2 (312)	2 (0.15)	3 (0.13)
Median House Price	5 (188)	3 (0.18)	1 (0.10)
Voting Result	3 (297)	4 (0.21)	4 (0.20)
National Budget	4 (283)	5 (0.26)	5 (0.25)
Pre-diabetic Experiment	6 (72)	6 (0.28)	6 (0.36)

of four elicitation techniques: **Explain-Only, Predict-Explain, Predict-Only, Predict-Feedback**.

Additionally, to better understand the effects of the elicitation techniques with a visualization, we varied whether a participant interacts with (and makes predictions about) the data using text or visualization (**Text** or **Vis**).

We crossed the elicitation techniques with modality, with the exception of Predict-Feedback, resulting in 9 possible conditions: None-Vis, Explanation-Only-Vis, Predict-Explain-Vis, Predict-Only-Vis, Predict-Feedback-Vis, None-Text, Explanation-Only-Text, Predict-Only-Text, Predict-Explain-Text. We excluded a Predict-Feedback text modality treatment due to the difficulty of generating personalized feedback based on freeform text predictions.

3.3.2 Participants

A prospective power analysis was performed for sample size determination based on the effect size and standard error of each technique and modality in pilots using a mixed effects model. We achieved 0.8 power under $\alpha = 0.05$ with 42 participants per condition. We then recruited 378 participants (42 per condition) from AMT, rewarding their participation with $1.50.

3.3.3 Procedure

Fig. 3.6 shows an overview of the study procedure. The study started with an *introduction* (Fig. 3.6(1)), in which we explained the data domain as the percentage of voters of different ethnicities (Hispanic, White, Black) who voted Republican in the 2008 presidential election, for several income brackets and states. To eliminate possible difficulties with the interactive nature of data entry, participants in the visual conditions watched a tutorial video to learn how to set and adjust a value in an related line chart. On the next page, participants in the prediction conditions (Predict-Explain, Predict-Feedback, and Predict-Only) were asked to *predict* the voting percentages for one randomly selected ethnic group across two income levels for each state (Colorado, Connecticut) in randomized order. To ensure that participants interacted to a similar degree across treatments, participants who were not prompted to predict were asked to *retype* a paragraph about elections in the U.S. (Explain-Only and None condition) (Fig. 3.6(2)).

On the next page, all participants *examined* the observed data (Fig. 3.6(3)), with prompts and feedback varying by condition. Participants in the None conditions were asked to examine the observed data several times (Fig. 3.1(a)).

Participants in the Explain-Only conditions were asked to generate and type in a few sentences of explanations to help themselves understand the data (Fig. 3.1(b)).

Participants in the Predict conditions saw their predictions in a lighter color against the observed data in the visual conditions. In the text conditions, the

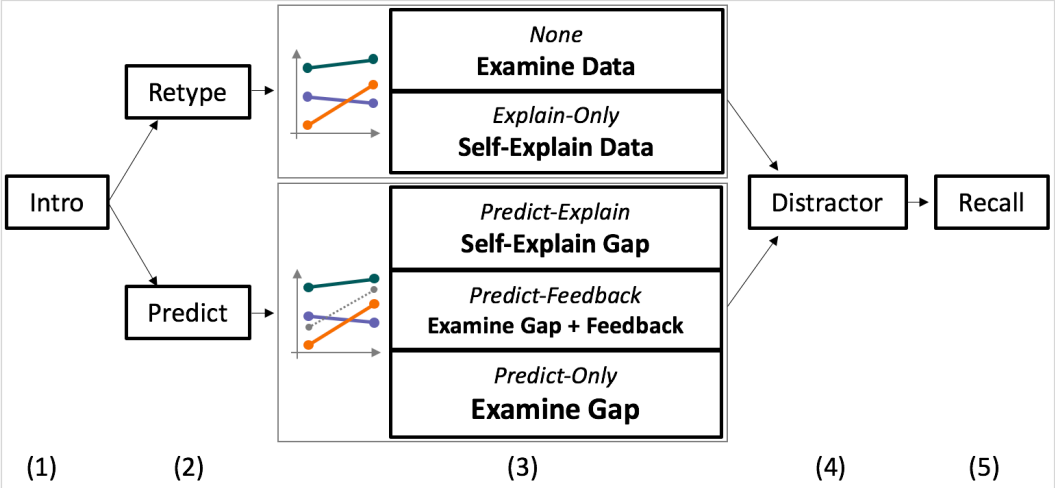


Figure 3.6: Overview of the study procedure. If participants were not asked to predict, they were asked to retype a general text on elections. If participants were not ask to generate self-explanations, they examined either the data or feedback depending on the condition.

textual predictions that the participant made were shown with the observed data presented in text.

Participants in the Predict-Explain conditions were asked to self-explain the difference between their prediction and the observed data (Fig. 3.1(c)).

Those in the Predict-Feedback condition saw accuracy feedback based on their predicted values (Fig. 3.1(e)). Feedback contained 1) the directionality of the participant's error (e.g., "Your prediction *overestimated* the percentage of Hispanic voters."), 2) verbal statements that reiterated the participant's prediction and the observed data (e.g., "You guessed 66%, but the true answer is 23%"), and 3) comparative information that indicated high-level patterns (e.g., Higher portions of white voters vote for John McCain then hispanic voters.), if the participant violated the pattern (e.g., "The percentage of Hispanic voters was lower than those of White voters").

Participants in the Predict-Only condition were asked to examine their predictions and the observed data several times (Fig. 3.1(d)).

As a distractor task, all participants then completed as many questions on a 10 question digital paper folding test as they could in three minutes [49]. The task also served to gather information on participants spatial visualization abilities, which have been shown to correlate with effective use of internal representations [71, 69, 70], (Fig. 3.6(4)).

After completing the paper folding test, participants in all conditions were asked to recall the percentage of voters of different ethnicities (Fig. 3.6(5)). Recall interfaces for each state were provided on separate pages and presented in reverse order from that in which the data was examined. Participants used an interface that matched the modality by which they viewed the data (text or visualization).

Participants were asked to respond to demographic questions, including age, education level, gender, and ethnicity, and were asked about their experience with visualizations.

3.4 Results

3.4.1 Data Preliminaries

The average time to complete the experiment was 19.4 minutes (SD=8.4), with no differences in response time across the conditions ($F(4) = 1.073, p = .37$). There were no significant differences between participants' demographic responses or relevant experience across the conditions. We excluded 3 participants that did not specify predictions in the text conditions, and 2 participants who participated multiple times.

3.4.2 Analysis Approach

We used two mixed effects models implemented in R's lme4 package to evaluate H1, H2, and H3. We used the normal approximation to calculate p-values of fixed effects using t-scores produced by lme4.

Dependent Variables

We considered two types of error indicating how well participants could recall the observed data. The accuracy in recalling individual data points was measured using the *absolute error*, i.e., the absolute difference between the recalled value and the observed value. To measure accuracy in recalling the higher level structure of datasets (e.g., trends within each group), we calculated the *trend error*, i.e., the absolute difference between the recalled and the actual slope of each line (set of values for an ethnicity) in the visualization.

Model Specification

In each mixed effects model, we included the four elicitation techniques (Explain-Only, Predict-Explain, Predict-Only, and Predict-Feedback), modality, and the interaction terms between modality and the techniques as fixed effects, with the control/baseline condition as the omitted reference condition. We included the participant id and the ethnicity group (e.g., Hispanic, etc.) as random effects. The spatial ability score, calculated as the number of correct answers out of 10 on the paper folding task, was included as a fixed effect. We centered the spatial ability score by its mean so that fixed effects describe a participant of average spatial ability. For easier interpretation, we report the intercepts for each elicitation technique separately for visual versus text with 95% confidence intervals. Coefficients are expressed in terms of the actual units used in the datasets (i.e., percentage).

3.4.3 Core Results

Visual Conditions

Absolute Error: Participants who used one of the four elicitation techniques recalled individual data points more accurately than those in the None-Vis condition (Fig. 3.7(a)). Proceeding by magnitude of effect, the Predict-Explain-Vis condition

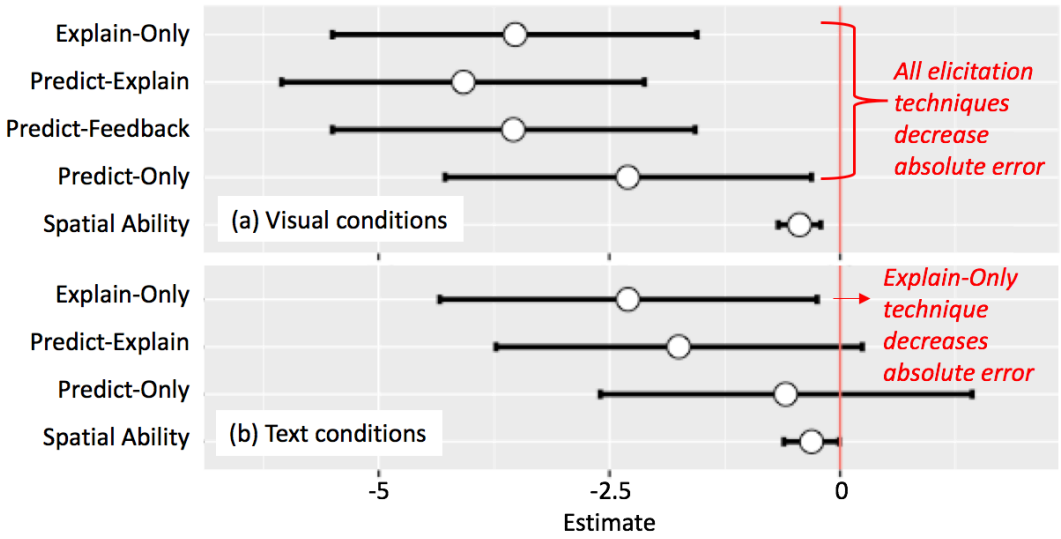


Figure 3.7: Estimated fixed effect coefficients from analyzing absolute errors for (a) visual and (b) text conditions for the voting result dataset. The error bars indicate 95% confidence intervals. Intervals that do not include include zero imply that we can be reasonably sure that some effect exists.

had the lowest absolute error relative to the None-Vis condition by -4.08 (i.e., participants in the Predict-Explain-Vis condition were, on any given recalled point out of the 12 total, more accurate at recalling the voting percentage by 4.08% out of 100% compared to the those in the None-Vis condition: $t = -4.02, p < .0001$) The Predict-Feedback-Vis condition had the next lowest effect (-3.54; $t = -3.49, p < .001$), followed by the Explain-Only-Vis condition (-3.52; $t = -3.47, p < .0001$), and the Predict-Only-Vis condition (-2.30; $t = -2.25, p < .05$). Hence, interactive elicitation appears to be reliable way to improve absolute recall of data presented through visualizations, even with variations in how prior beliefs is elicited.

Participants' scores on the spatial ability test also predicted a lower absolute error, as we would predict from prior work indicating the relationship between spatial visualization ability and visualization comprehension [71, 69, 70]. With each additional correct answer in the paper folding task, participants' expected absolute

recall error decreased by 0.44 ($t = -3.63, p < .001$).

We observed no difference in absolute error between the four elicitation techniques.

Trend Error: Only participants in the Predict-Explain-Vis and the Predict-Feedback-Vis conditions had a lower trend error compared to the None-Vis condition. Specifically, being in the Predict-Explain-Vis condition lowered errors by -2.06 relative to the None-Vis condition ($t = -2.07, p < .05$), while being in the Predict-Feedback-Vis condition lowered errors by -2.79 relative to the None-Vis condition ($t = -2.80, p < .05$). We observed no effect of the spatial ability score on participants' trend errors ($t = -0.53, p = .596$).

Comparing Effects of Techniques: Visual vs Text

Absolute Error: Overall, we observed 3.76 percentage points (36%) more errors on average per recalled value among visual conditions compared to text. This aligns with prior research indicating that text is better for exact value retention than visualization [81].

In comparing the absolute error between the text conditions, the Explain-Only-Text condition was the only condition that led to lower absolute errors than the None-Text condition, by an average of -2.00 ($t = -2.07, p < .05$). We observed no effect from the Predict-Explain and the Predict-Only techniques in the text modality conditions when we compared them to the None-Text condition (Fig. 3.7(b)). Text presentations may not provide the same type of natural support for prompting implicit reflection and correction of one's prior beliefs compared to visualizations.

Participants with higher spatial ability scores again had a lower absolute error on recall by 0.31 ($t = -2.04, p < .05$).

Trend Error: Overall, we did not observe differences in average trend errors between visual conditions and text conditions. In comparing the trend error between the text conditions, we found no effect of any of the elicitation techniques (i.e.,

Explain-Only-Text, Predict-Explain-Text, or Predict-Only-Text) compared to the None-Text condition. We saw no apparent decrease in trend error from spatial ability ($t = -1.30, p = .193$).

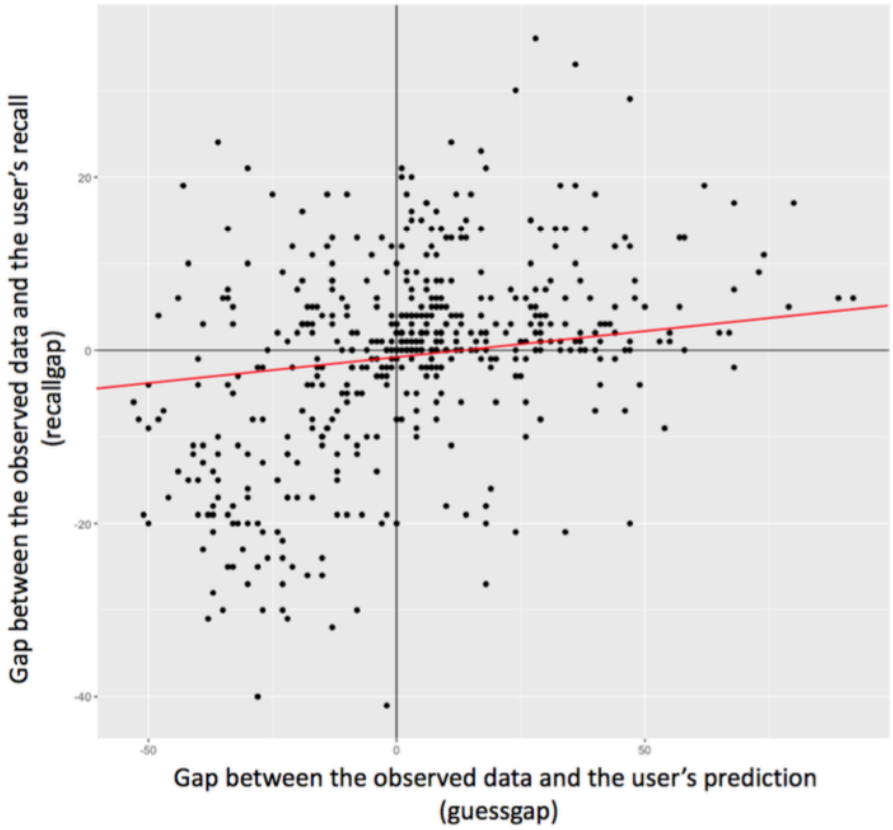
The Effect of Prior: Anchoring in the Prediction Conditions

Participants in the Predict-Only-Vis, Predict-Explain-Vis, and Predict-Feedback-Vis conditions may have a tendency to recall aspects of their own prediction due to the deliberate attention required to generate the prediction. We observed a weak positive correlation between the values that participants predicted and the values that they recalled (Fig 3.2, $R^2 = 0.176$, intercept$= -1.28$, $slope = 0.18$) Additionally, we found that for 75.9% of the data points from participants who were asked to predict, their recalled value showed a bias in the same direction as their predicted value: if they underestimated the value in the prediction phase, they tended to underestimate the value in the recall phrase. The same pattern could be observed when they overestimated. Hence, a slight anchoring effect appears to be present.

Quantity and Quality of Self-Explanations

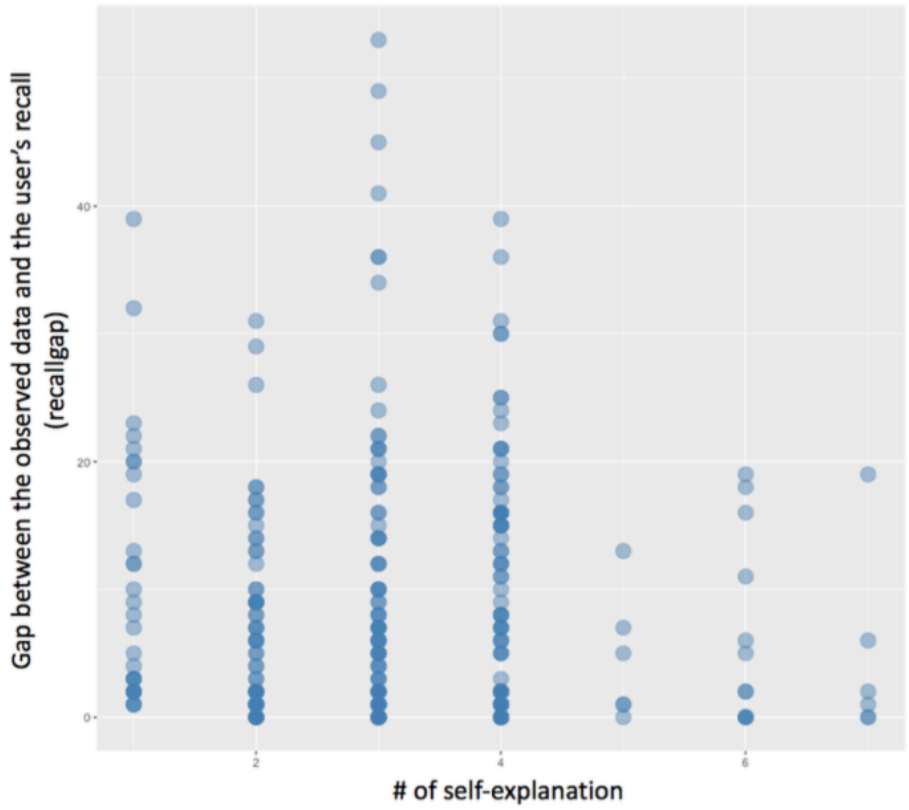
The quantity and quality of self-explanations have been shown to affect comprehension in prior work [31, 123]. We analyzed the correlation between self-explanation quantity and quality and recall performance for participants in the Explain-Only condition. As a proxy of quantity we counted each sentence as a self-explanation (mean:3.2, range: 1-7). We tallied the total number of self-explanations generated by a participant and regressed the average recall error made by the participant on the sum. We observed no effect of the number of explanations on recall error (Fig 3.3, $R^2 = 0.001, F = 0.39, p < .533$).

To measure the quality of each self-explanation, we devised criteria informed by Chi's approach to distinguish high and low quality explanations [29]. We differentiated between two factors that characterize the quality of self-explanations: the level



Model: lm(recallgap ∼ guessgap)				
Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-34.35	-5.40	0.43	5.10	32.15
	Coefficients	**Estimate Std. Error**	**t value**	**Pr(>|t|)**
Intercept	-1.28	0.47	-2.75	0.00609 **
guessgap	0.18	0.02	10.35	<2e-16 ***
Multiple R-squared: 0.1758, Adjusted R-squared: 0.1742				
F-statistic: 107.1 on 1 and 502 DF, p-value: <2.2e-16				

Table 3.2: The visualization and the result of the anchor effect analysis. The result shows a weak anchoring effect.



Model: lm(recallGap ~# of Self-Explanation)				
Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-6.83	-3.67	-1.17	1,46	16.90
	Coefficients	**Estimate Std. Error**	**t value**	**Pr(>|t|)**
Intercept	10.47	2.49	4.20	0.00014 ***
# of Self-Explanation	-0.23	0.76	-0.31	0.761
Multiple R-squared: 0.002339, Adjusted R-squared: -0.0226				
F-statistic: 0.09377 on 1 and 40 DF, p-value: 0.761				

Table 3.3: The correlation analysis between the quantity and the quality of self-explanations. The model shows no effect.

of prior beliefs involved in inference (inference with prior beliefs, inference with no prior beliefs, no inference), and the level of detail of the inference (high, low).

Inference without prior beliefs, High detail: "Generally, people with incomes over 75k were less likely to vote for John McCain in 2008. Blacks who made over 75k were slightly more likely to vote for him, but it was a very small increase to 3% meaning not many blacks voted for McCain in any income category."

Inference without prior beliefs, Low detail: "Majority of people no matter ethnicity voted for the Democrats and not Republicans."

Inference with prior beliefs, High detail: "There was a slight increase in the White voting population with the higher income bracket, I could assume that this is due to McCain's policies which benefit the wealthier. "

Inference with prior beliefs, Low detail:"People are more conservative in Colorado."

No inference: "Each colored line is a different race. Each point is a different income bracket."

Two researchers coded the set of 42 explanations (*Cohen*ś $kappa = 1$). We conducted a two way factor analysis on the average absolute error and the trend error, but observed no effect of either quantity or quality on factors. It may be that the difference between participants' self-explanation quality was smaller overall than in educational studies of spontaneous self-explanation, perhaps due to the incentives to work quickly on AMT.

3.4.4 Replication on Low and High Familiarity Datasets

We conducted two additional partial replications of our study to evaluate the effect of elicitation techniques on datasets that our preliminary survey identified as more and less familiar on average. We replicated all visual conditions.

Low Familiarity Dataset: Scientific Experiment Results

We created two visualizations depicting results from a scientific experiment on the blood glucose level of various groups of mice after antibody injection [67]. Each visualization differentiated two amounts of time since injection (0 and 30mins, and 60 and 120mins). Each visualization included lines for three groups (Lean, 2H10, and control) similar to the three groups of Hispanics, White, and Black voters in the main study. Hence, the two visualizations replicated the structure of the voting results data across two states.

Absolute Error: We observed a similar pattern of effects of the techniques on decreasing absolute error as for the voting result data, with the exception of the Predict-Only-Vis conditions (Fig. 3.8(a)). The Explain-Only-Vis condition had lower errors by 1.78 ($t = -3.75, p < .001$), and the Predict-Explain-Vis conditions had lower errors than the None condition by 1.27 ($t = -2.62, p < .01$). The Predict-Feedback-Vis condition had lower errors than the None condition by 1.31 ($t = -2.69, p < .01$). Predict-Only-Vis condition had no effect compared to the None condition ($t = -1.43, p = .152$).

We observed no difference in recall performance between the Explain-Only-Vis, the Predict-Explain-Vis, and the Predict-Feedback-Vis condition.

Trend Error: We also observed no effects of the elicitation techniques on decreasing trend error compared to the None condition.

High Familiarity Dataset: Calorie Content of Fast Food

We created two visualizations using data on the calorie content of three fast foods (Milkshake, Coke, French fries) for two serving sizes (small, large) at McDonald's and Burger King.

Absolute and Trend Error: Results of the mixed effect model for absolute error (Fig. 3.8(b)) and trend error indicate no differences between any conditions for the fast food dataset.

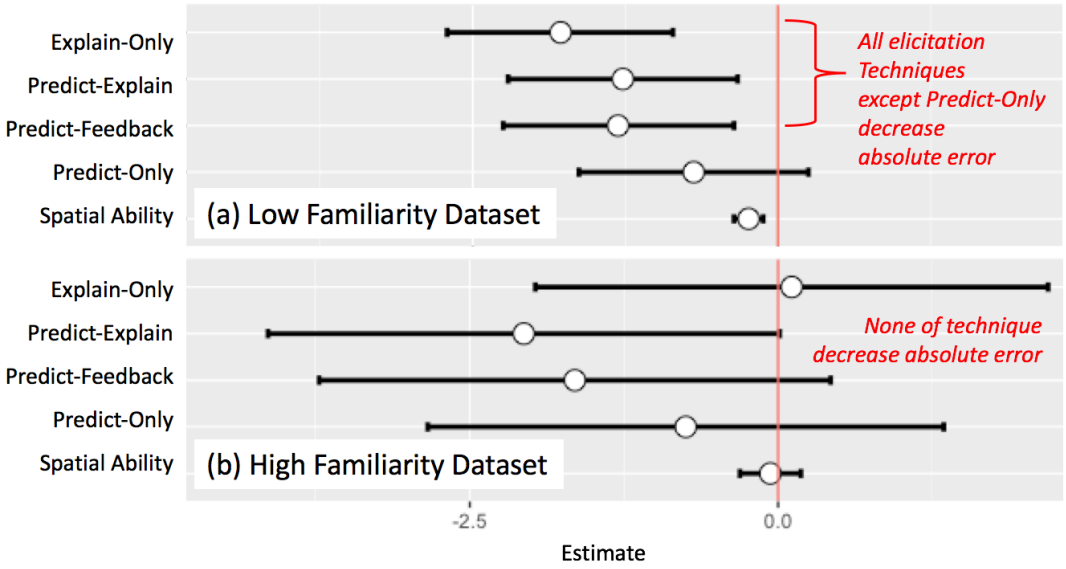


Figure 3.8: Estimated fixed effect coefficients from analyzing absolute errors for visual conditions (a) for the scientific experiment dataset, and (b) the fast food calories dataset.

We conclude from these additional partial replications that prediction, when combined with additional mechanisms like explicit self-explanation or feedback, can lead to lower absolute error even for very unfamiliar data where prediction might be difficult. Perhaps because the data was already too familiar, we saw no replication of effects for any elicitation technique on the high familiarity dataset.

3.5 Discussion

3.5.1 The effect of Reflecting on Prior Beliefs

Our results suggest the promise of incorporating mechanisms for eliciting prior beliefs in visualization.

First, our work extends prior work on self-explanation by showing that prompting users to self-explain information visualizations can improve their ability to recall specific data points later.

Additionally, our work is the first to show that incorporating prediction tasks, as

in our Predict-Explain-Vis, Predict-Only-Vis, Predict-Feedback-Vis also improves users' ability to recall specific data. We hypothesize that predicting focuses a user's attention on their prior beliefs, making them more likely to attend to the gap between their prior beliefs and the observed data when it appears. We expected that predicting would be less effective in the text modality than in the visual modality. In fact, we did not observe an effect of any of the prediction techniques on improving recall in the text conditions. This may be because participants in the visual conditions are shown the gap in a visual form, reducing the deliberative effort required to process the gap in a text format. Using a visualization of the gap is likely to free up participants' cognitive resources in contrast to text, so that they can focus on more meaningful activities [129] such as updating and repairing their mental model.

Except for the Predict-Only-Vis conditions, we were able to replicate the effect of the elicitation techniques for decreasing absolute error with a less familiar dataset, the scientific experiment results. These results suggest that prior beliefs about a dataset is not necessarily required for a user to benefit from prediction, explanation, and feedback techniques. The fact that we were not able to replicate the effects in the Predict-Only condition may indicate that when familiarity is lower, a user needs additional reinforcement to recognize the gap, such as being prompted to explain or being given visual feedback on the gap.

We could not replicate the effects of the elicitation techniques on the high prior familiarity dataset, the calorie content of fast food. One possible reason that we did not see an effect here is that participants were already relatively accurate in estimating the values, such that the initial 'gap' is too small to see much positive impact from the elicitation techniques.

We observed that the Predict-Feedback-Vis and Predict-Explain-Vis techniques decreased the trend error in recalling the visualized trends for the voting results data. However, we did not see similar improvements of trend errors with either of the other datasets. Varying graphical complexity may be one reason for this discrepancy: in

the voting results data, two of the lines intersect (thus adding complexity), whereas in the other two datasets all three lines have similar, non-intersecting trends.

We believe that prompting predictions and providing feedback, which we found to reduce both the absolute and relative recall error, would work well when presenting visualizations in practice. Compared to explicitly prompting a user to provide self-explanations, first prompting a user to make predictions is likely to engage a user's curiosity. Once the user has "invested" attention by predicting, they may be more open to feedback that can further direct their attention to needed adjustments in their prior beliefs.

Though we asked participants in the predict conditions to predict only select data points, we observed a decrease in recall error across all data points. In the Predict-Explain-Vis conditions, we see evidence that participants are generating explanations associated with all three ethnic groups despite being prompted to explain only the difference between the predicted group and observed data. For example, participants wrote in the comments:

> "I really overestimated the numbers of black people. I expected more of them to vote than the Hispanics."

> "I may have underestimated the population of Hispanics (and Blacks) in Colorado. I assumed both were minorities who held liberal viewpoints, since traditionally minorities tend to prefer Democratic candidates."

This suggests that a designer need not require a user to predict every data point in the visualization to engage users with the entire dataset.

3.5.2 Design Space for Graphical Prediction and Feedback

Our study shows the benefits of eliciting users' prior beliefs, such as their expectations of data, and prompting them to reflect on how their beliefs relates to the

Task	Detail Task	Manipulation Component	Encoding	Possible Interaction
Predict Continuous (Quantitative) Variable	Predict Data Value	Mark (bar)	Bar chart	Drag up a bar to set height
		Mark (line)	Line chart	Draw a line
		Mark Attribute (color of areas)	Map (choropleth)	Brush on color over an area
Predict Categorical Variable	Predict Categorical Membership	Mark Attribute (color of areas)	Area chart	Brush on color over an area
			Pie chart	Brush on color over a sector
Predict Data Structure and Model	Predict Correlation	Mark (line)	Scatter plot	Draw a regression line
			Line chart	Draw a line
	Predict Cluster		Scatter plot	Draw a contour
			Dendrogram	Drag and drop element to the cluster
	Predict Connectivity		Node-link diagram	Draw an edge
	Predict Confidence Interval		Box plot	Draw a line to mark confidence interval

Table 3.4: Possible tasks.

data. However, the design space for applying graphical prediction and feedback techniques to information visualization remains relatively unexplored. In the following, we characterize key considerations in applying these techniques to visualizations.

We informed our elaboration of the design space through several forms of evidence: observations from our studies; examples in the media, primarily from the New York Times [5]; and our own development of prototypes applying the techniques to visualizations like bar charts and line charts.

Based on these experiences, we differentiate three considerations that influence the effectiveness of graphical prediction and feedback applications: the *prediction*

task and graphical elicitation technique (for what tasks and in what ways can the user express their prior beliefs?), the *contextualization mechanism* (how does the interface provide clues to constrain the user's prediction?), and the *feedback technique* (how does the interface draw the user's attention to the gap between their predictions and the observed data?).

Prediction Task and Elicitation Technique

A first question in designing a visualization that elicits predictions is "What should the user predict?". A visualization can elicit value predictions for quantitative or nominal (categorical) variables, or the outcome of a model or analysis.

Direct manipulation is a natural way to implement the first one, *value prediction*. For example, a user might click to add a mark, or drag a mark from an axis to set or update its position in a 2D plot like a scatter plot. Other marks require different interactions: a user might drag a bar to set its height or click to position the top of the bar in a bar chart, and use a smooth dragging operations to position a line in a line chart.

Data encoded in the visual attributes of marks, such as color or shape may also be predicted. For example, nominal data (categories) might be encoded by the color hue of marks in a scatter plot. *Predicting categorical membership* can be instrumented with interactions like brushing. For example, some points may remain uncolored in a scatter plot where color encodes the categorical membership of data. The designer can have a user select a color from an interactive legend, and drag across points to assign that category.

In The New York Times' elicitation of users' predictions for the 2014 senate election, a user was able to cycle through different binned probability levels for the voting percentage for each party by repeatedly clicking on a state in the map [82].

Similar to categorical membership, the designer of a visualization may ask a user to *predict clusters* in a scatterplot or network diagram. For example, given

an interactive network diagram, the user can draw a contour around the nodes or use brushing interactions similar to those described for categorical membership to designate clusters of related nodes. *Predicting connectivity* could also be applied to support edge prediction in a network diagram.

Alternatively, the designer can ask the user to visualize their expectations for the outcome of more complex analyses applied to raw data. For example, graphical prediction techniques could be used to elicit predictions on multivariate correlations, uncertainty, or other results attained through statistical modeling. For example, the New York Times 'You Draw It' interactive prompts the user to *predict a regression line* representing the relationship between parents' income percentile and percent of children who attended college [5]. To facilitate understanding of uncertainty in data, a user might be prompted to *predict a confidence interval or region* given a 2D presentation of bars, points, or lines denoting sample statistics.

Regardless of what is predicted, *the directness of the prediction interaction and degrees of freedom* provided to the user by the interface are important design considerations. Freeform interactions can be used to allow the user greater flexibility in drawing, such as providing a high degree of resolution (i.e., space of possible fits) to users drawing regression lines in 2D visualizations. More constrained forms of interaction can be realized through snapping functions (e.g., snapping a predicted regression line to the nearest grid point). Similarly, a designer may choose to only allow the user to manipulate certain parameters of components (e.g., drag a curve or slider to change line curvature, drag the edge of a circle to increase size while maintaining shape). While more constrained interactions may serve to reduce error and focus user's attention on key parameters (e.g., slope or magnitude alone), they are likely to add abstraction. Our own experimentation with interactive prototypes suggest that many users enjoy the novelty of using an interactive visualization interface to draw with few constraints.

Contextualization Mechanism

Contextualization mechanisms can be used to guide the user's prediction as they form a guess. For example, the amount of effort required for the user to make a guess can differ based on the number of *reference marks* (e.g., dots, bars, lines) that the visualization initially presents. In our study, we presented two of the three ethnic groups by default, which provided cues to guide users' predictions of the remaining group. How much data to reveal through reference marks can be decided based on how familiar users are expected to be with the dataset (less familiar=more reference marks). Or, the reference marks can be selected dynamically through personalization. For example, for datasets that depict regional data, identifying and initially presenting marks depicting the user's region based on their IP address may increase engagement while providing useful context for the user.

Prediction hints provide more direct guidance, either through text or visual annotation, on where a user's prediction should be made, helping educate users about the meaning of the encodings.

As the New York Times' visualization "You Draw it" [5] demonstrates, one or more data points can be presented as a hint that the user's prediction line should pass through.

Designers should consider the scale of the x and y-axes in 2D charts. In piloting our preliminary study for choosing a dataset, we observed that users' predictions were quite sensitive to the axis range. When we presented the full 0-100% percentage range for percentage variables (e.g., the percentage of the U.S national budget of health care), users' estimates showed a bias toward the center in the plotting range. This effect was lessened when we trimmed the axis range based on the maximum value of the dataset, suggesting that users implicitly view the axis range as a clue to the data scale.

Feedback Technique

After a user draws their prediction, feedback on how the user's prediction compares to the observed data can help prompt reflection on prior beliefs.

For example, personalized feedback can provide information on the *accuracy of a prediction*, as we provided in our study. Feedback may take the form of aggregated, quantified accuracy information (e.g., "Overall, you were 80% right in guessing the amount of CO_2 emission in U.S") or information on the directionality of biases (e.g., "You over-estimated the overall trend."). Feedback may also occur at a more granular level, encouraging the user to adjust their expectations of individual data points: "Your guess on year 2001 was 6 points off; a little higher and you would be correct."

As users' predictions are collected, *social feedback* may serve to further engage users to think about the data and their own expectations. However, social feedback may also overshadow user's own interpretations; hence social feedback might be withheld until after the user has provided their own prediction [76] The interface can prompt social comparisons by visualizing other users' predictions alongside the user's.

General design considerations affecting feedback include how specific and in what modality feedback is presented (e.g., visual, text, etc.). Based on our study finding that textual presentations are less effective for drawing attention to the gap, we except visual feedback or a combination of visual and text feedback to be more powerful. Animating feedback, such as by dynamically moving marks added by the user to their true positions, or adding textual feedback to prediction errors point by point, may be particularly effective for drawing a user's attention to the gap.

3.6 Summary

This chapter introduces the interfaces to prompt people's reflection on their prior beliefs, the evaluation of the effects of the interfaces on data recall, and the inves-

tigated mechanisms behind the effects. The investigation provides knowledge on how visualization designers and researchers could elicit people's beliefs without having negative impacts on people when they interact with visualizations. In the next chapter, I will introduce how to extend the approach to elicit probabilistic beliefs and model people's beliefs with a Bayesian framework using elicited beliefs.

Chapter 4

A BAYESIAN COGNITION APPROACH TO IMPROVE DATA VISUALIZATION

Chapter 3 demonstrates how people's beliefs could be elicited in a natural way from users to integrate into visualization interaction. The studies show that the act of the externalization benefits users to remember data better, possibly by seeing the gap between their prior beliefs and the actual data.

With that understanding, this chapter introduces ways to elicit people's probabilistic beliefs to model their visualization interaction (Research Question 1). As oppose to eliciting only the "location" of the beliefs as modeled in chapter 3, an individual's uncertainty of their beliefs is elicited this time. I introduce a demonstration of how we could apply Bayesian models using the elicited beliefs to understand visualization interpretations (Research Question 3). I present what we learn about users' data interpretation process by modeling visualization interpretation as Bayesian inference, and how the approach can shape how we design and evaluate visualizations. This chapter also provides an analysis that shows the effect of prior elicitation (Research Question 2).

4.1 Visualizations as Media to Inform Belief Change

Opposing a "data-only" view of visualization, models of graphical comprehension from psychologists have described how top-down influences, including prior beliefs and expertise, influence what a person attends to [26]. Studies demonstrate how prior knowledge can lead to other "downstream" effects on visualization related outcomes, such as how effective an interactive visualization is for different users [69]. The previous chapter demonstrates how eliciting people's beliefs about data directly

through the interface can positively impact data recall and prompt critical thinking about data. However, research in data visualization has yet to develop descriptive or normative cognitive models for predicting and evaluating how people update the prior beliefs they bring upon viewing data.

Outside of visualization research, psychologists have developed these types of models of how people update their beliefs or opinions about data or a proposition, given information about their prior beliefs [64, 133]. Bayesian models of cognition compare human cognition, which is assumed to draw on prior beliefs, to a normative standard for rational induction from noisy evidence [66]. By combining key components of Bayesian statistics—including a likelihood function describing the probability of the data given some assumed distribution, a description of the prior probability of different values, and laws of conditional probability—Bayesian cognitive modeling can *describe* how people update their beliefs given data. Bayesian models have provided explanatory accounts of how people make various real-world perceptual judgments, higher cognitive inferences, and learn and reason inductively [132, 64, 98, 133]. Bayesian cognitive models can also *prescribe* what updated beliefs are most consistent with one's prior beliefs and the data, providing a normative framework for evaluating interactions with data presentations.

We make several contributions in this chapter. First, we demonstrate a Bayesian cognitive model for assessing how people interpret data presentations like simple visualizations. In contrast to other frameworks for studying visualization use, a Bayesian cognitive model can be used to examine how people *change their beliefs* in response to presented data.

Deploying the model we develop, we characterize the extent to which the belief updating of users of a simple visualization of survey results resembles Bayesian inference (Pilot). We find evidence that on average people update their beliefs rationally, but individuals often deviate from expectations of rational belief updating. These findings hold across multiple datasets and prior elicitation methods (Study

1). We find that people deviate considerably more from the predictions of Bayesian inference even in aggregate when presented with datasets of a very large sample size.

Finally, we demonstrate how a Bayesian cognitive model can be used to evaluate data presentations (Study 2). We show how Hypothetical Outcome Plots (HOPs), animated plots that show uncertainty via draws from a distribution, improve deviation from normative Bayesian responses relative to not presenting error information.

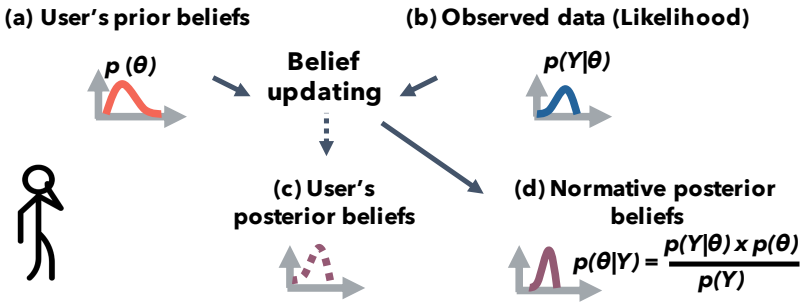


Figure 4.1: Distributions of residuals (observed - predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of the normative posteriors.

4.2 Visualization Interpretation as Bayesian Inference

Before introducing our studies, we will walk you through an example scenario that illustrates how we envision Bayesian framework can be applied in visualization interpretation. Imagine a user will be presented with a visualized estimate of a parameter θ. As an example, imagine that the parameter is the proportion of residents of U.S. assisted living centers who have Alzheimer's. Before the user views *observed data*, they articulate their prior beliefs. The user specifies a prior by assigning probability over plausible values of θ using a graphical or text-based interface (Fig 4.1a).

In Bayesian inference, beliefs take the form of a probability distribution. For a proportion parameter θ, a Beta distribution is the appropriate distribution to capture beliefs. Two parameters sufficiently define a unique Beta distribution:

$Beta(\alpha, \beta)$. We can think of $\alpha - 1$ as the number of successful events (e.g., the number of residents in assisted living centers who have Alzheimer's), and $\beta - 1$ refers to the number of unsuccessful events (e.g., the number of residents in assisted living centers who don't have Alzheimer's).

Imagine a user who guesses that approximately 10% residents in assisted living centers have Alzheimer's, but with relatively high uncertainty. Assume that the information their beliefs imply is equivalent to having observed a sample of 10 assisted living center residents, one of which had dementia. Their prior beliefs are captured by the distribution $Beta(2, 10)$. The sum the successful events and the failure events (i.e., 10) represents the amount of information (or conversely uncertainty) contained in the user's prior distribution.

Imagine that the user is next presented with a visualization of an estimate captured by observed data (Fig 4.1 (b)), such as the proportion of assisted living center residents with dementia according to records for a chain of centers with locations across the country. Out of 1,000 residents of these chains, 420 have dementia. We model the data generating process as a binomial process in which any individual independently has the disease with a certain (identical) probability θ. We represent the observed data as a *likelihood function* capturing the probability of different values of θ given the observed data. The likelihood encodes the relative number of ways that different values of θ could produce the observed proportion given our assumptions about the data generating process and the size of the observed sample. The likelihood function for a reported proportion, 42%, from a total sample of 1,000 residents can be represented by $Binomial(1000, 0.42)$, implying an expected 420 successful events and 580 failure events but with some uncertainty due to sampling error.

$$\begin{aligned} \#of successes_{posterior} &= \#of successes_{prior} + \#of successes_{data} \\ \#of failures_{posterior} &= \#of failures_{prior} + \#of failures_{data} \end{aligned} \tag{4.1}$$

In a Bayesian framework, "rational" belief updating is prescribed by the *normative posterior distribution* (Fig 4.1d). This distribution is calculated by using Bayes rule to update the probability of θ in the prior with the information about θ implied by the likelihood function. Equation 4.1 results from using Bayes rule to estimate the number of successful events and the failure events in the posterior beliefs as a function of the estimates implied by the observed data and prior. The number of successful and failure events in the posterior beliefs is equivalent to a Beta distribution: $Beta(422, 590)$. Intuitively, under Bayesian inference the user's belief distribution after encountering the observed data shifts proportionally to the amount of information contained in the two distributions.

4.3 Developing Research Questions and Goals

Prior beliefs clearly play a role in data interpretation. However, belief updating is rarely formally modeled in research related to data presentation and visualization. Studies of Bayesian cognition suggest that Bayesian inference can be used to characterize many aspects of learning and cognition. We apply a Bayesian cognitive modeling approach to a simple data interpretation task to understand where people align with, and deviate from, normative Bayesian inference individually and in aggregate. While the computational complexity of Bayesian inference makes it doubtful that cognition uses exact inference [84], in the context of interpreting presented data in everyday settings (such as in data journalism) we would expect under Bayesian assumptions to see that people (1) are capable of providing *priors* describing the uncertainty in their beliefs about a parameter, and (2) update these beliefs to incorporate observed data. Our work sheds light on the degree to which these assumptions hold for a simple data interpretation task.

In contrast to prior work in Bayesian cognition that avoids obtaining priors directly from people [66, 152], we design and apply a paradigm in which we *elicit people's prior and posterior beliefs about the probability that a parameter takes var-*

Figure 4.2: Distributions of residuals (observed - predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of the normative posteriors.

ious values (Pilot, Studies 1, 2). Though it is commonly argued that people have difficulties reasoning about probability, the notion that people are capable of maintaining subjective probabilities is well-established in decision theory, congruent with canonical work in judgment and decision making like that of Tversky and Kahneman [2], and supported by a body of work in economics on subjective probability elicitation, including from laypeople (see [101] for a review. Having obtained prior beliefs, we fit a distribution to them then use Bayes' rule to compute the *normative posterior distribution* for each person, the posterior distribution that is expected if the person is a perfect Bayesian agent given the observed data and their prior distribution.

In a first study, we compare the distribution fit to the posterior beliefs we elicited from each person to the normative posterior beliefs computed using that person's prior Bayesian solutions (Fig. 4.2 top row). We also compare people's *aggregate posterior distribution* (i.e. the posterior distributions representing the aggregate of all people's posterior distributions) to the *normative aggregate posterior distribution* (i.e., the normative posterior distribution calculated using a prior distribution representing the aggregate of all people's prior distributions) (Fig. 4.2 bottom row).

Alignment between people's responses and the normative Bayesian solution at this aggregate level may suggest that people are "sample-based Bayesians" [143] performing approximately inference.

Prior work in visualization and judgment and decision making suggests that different subjective probability elicitation techniques can produce varying results, perhaps because some techniques (such as frequency framings) better align with people's internal representations of uncertainty [62, 78, 112]. In a second study, we assess *how sensitive people's responses are to different elicitation methods*, which vary in the input format for beliefs they use (i.e., continuous probability versus discrete samples).

In a second study, we show how a Bayesian cognitive model can be used to assess the effectiveness of design changes. One aspect of visualization design that is likely to be relevant to how people update beliefs is the presentation of uncertainty. If people see the observed data as more certain than it is (e.g., reflecting belief in the law of small numbers [137]), their posterior judgments may reflect overweighting of the observed data and underweighting of their prior. On the other hand, if people see the data as less certain that it is (e.g., non-belief in the law of large numbers [15]), their posterior judgments may reflect underweighting of the observed data and overweighting of their prior. To demonstrate how a Bayesian cognitive model can support visualization design decisions, we *compare the results of Bayesian modeling across a default static visualization typical of those found in the media and an animated hypothetical outcome plot (HOP [79]) uncertainty visualization.*

4.4 Pilot: Developing a Bayesian Model of Data Interpretation

We demonstrate a Bayesian model of cognition for assessing visualization interpretation. We evaluate the extent to which individuals' judgments are consistent with "fully" Bayesian inference by assessing how closely their individual posterior distributions align with the normative posterior distribution calculated given their

prior. Secondly, we also consider whether people's judgments might instead be consistent with what has been termed "sample-based" Bayesian inference (a form of approximate Bayesian inference) by evaluating how closely the aggregate posterior distribution aligns with the normative aggregate posterior distribution.

4.4.1 Study Design

We recruited 50 participants with 95% or above approval rating from Amazon Mechanical Turk, rewarding their participation with $1.0. The average completion time was 7.3 minutes (SD=5.2).

Dataset and Presentation

For our studies we sought a simple dataset that would nonetheless be representative of those shown in the media or public facing reports. We selected a dataset with a single variable which represents a proportion. The dataset describes survey results intended to measure attitudes towards mental health in the tech workplace (N=747) [1]. We chose one question from the survey "how often do you feel that mental health affects your work?" to formulate our proportion parameter: "the proportion of women in the tech industry who feel that mental health affects their work often." To present the observed proportion to participants in our study, we created an "info-graphic" style visualization (Fig. 4.7 (a)) which shows this proportion using a grid format commonly used in the media to present proportions (e.g., [6, 83, 102]).

Prior and Posterior Elicitation

To elicit participants' prior and posterior distributions, we used a technique that asks participants about two properties of their internal distribution: the most probable value of the parameter (*mode (m)*) and their *subjective probability* (Fig. 4.6(b)) that the parameter falls into the interval around the mode ($[m-0.25m, m+0.25m]$). Prior research in probability elicitation for proportions indicates that this technique

is less sensitive to imprecision that may arise when one externalizes a subjective distribution compared to a percentile approach and alternative location plus interval implementations [153]. A benefit of this approach is that estimates of Beta distribution parameters can be analytically computed from participants' answers [56].

4.4.2 Results

Fitting Individual Responses

We first converted participants' elicited prior and posterior beliefs to Beta distributions using an optimization approach suggested in previous work [112]. The approach finds an optimal Beta distribution parameterized by α and β which minimizes the sum of two terms: (1) the square difference between the participants' mode and the estimated mode of the Beta distribution and (2) the square difference between the probability that each participant associated with the interval and the estimated probability of the interval in the distribution.

Fitting Aggregate Responses

To obtain parameters for the aggregated prior/posterior distributions (α_{agg} and β_{agg}), we averaged participants' αs and βs respectively from the individual prior/posterior distributions: $\alpha_{\text{agg}} = (\alpha_1 + ... + \alpha_{\text{N}})/N, \beta_{\text{agg}} = (\beta_1 + ... + \beta_{\text{N}})/N$ ($N = \#$ of participants).

Calculating Normative Posteriors

We can calculate a participant's normative posterior by using α and β estimates from their prior distribution combined with the number of successes (e.g., the number of women who said their mental health affects their work often) and failures (e.g., the number of women who said their mental health affects their work not often) in the observed data (Eq. 4.2). The α and β for the aggregated normative posterior are

calculated in the same manner using the aggregated prior α and β estimates.

$$\begin{aligned} \alpha_{\text{normative posterior}} &= \#successes + \alpha_{\text{prior}} \\ \beta_{\text{normative posterior}} &= \#failures + \beta_{\text{prior}} \end{aligned} \tag{4.2}$$

We evaluate the degree to which individual and aggregate posterior distributions resemble the normative Bayesian posterior distributions by plotting residuals (*observed* - *predicted*) when predicting the means and standard deviations of participants' posterior distributions using normative Bayesian inference (Fig. 4.3). A distribution of residuals that is loosely centered around zero suggests "noisy" Bayesian inference, where each individual may deviate from the normative posterior due to approximate inference but in aggregate, the observed posterior resembles the normative posterior. Residuals for means are roughly centered around zero, with 95% of the values falling between -0.16 and 0.58). A small number of participants provided posterior distributions with means that were considerably greater than predicted (i.e., believed that the true proportion of women in tech who feel that mental health affects their work often was much larger than predicted from the prior and the observed data).

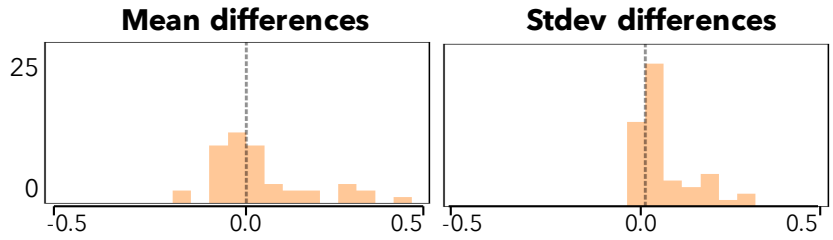


Figure 4.3: Distributions of residuals (observed - predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of the normative posteriors.

Residuals for standard deviation are also roughly distributed around zero, but show that participants were biased on average to produce posterior distributions

with greater variance than the normative posterior. This suggests a tendency among participants to provide posterior beliefs indicating more uncertainty than is rational given the observed data and the information contained in their prior.

Following this observation, we analyzed where each participant's posterior distribution was located relative to the normative posterior distribution (Fig. 4.4). We found that 44% of participants (22 out of 50) overweighted the mode of the observed data (i.e., their posterior distributions are closer to the observed data than they should be), while 34% of participants (17 out of 50) overweighted the mode of their prior distribution, and 18% of the participants (9 out of 50) provided posterior beliefs that moved further than the prior from the observed data. Only two participants (4%) were within ±1% range of the mode of their normative posteriors.

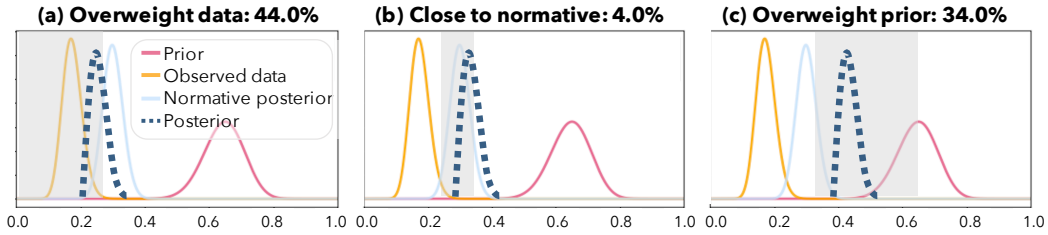


Figure 4.4: Example illustrations of three different types of the update. Proportions of participants whose posterior distributions (dotted line) imply overweighting of the mode of the observed data, reasonable alignment with the normative posterior, and overweighting of the mode of the prior distributions. An additional 18% of participants (not shown) provided posterior beliefs that were further than the prior from the observed data.

Per our pre-registration we report log KL divergence (KLD) [95] between normative and observed posteriors. KLD is an information theoretic measure of the difference between two probability distributions. Examining log KLD at the individual and aggregate levels aligned with our observation from the residual plots: few individuals act "fully Bayesian", but in aggregate the responses are close to normative predictions. The mean log KLD for a participant at the individual level was 0.52 (SD=1.18; 3.31 in non-log terms).

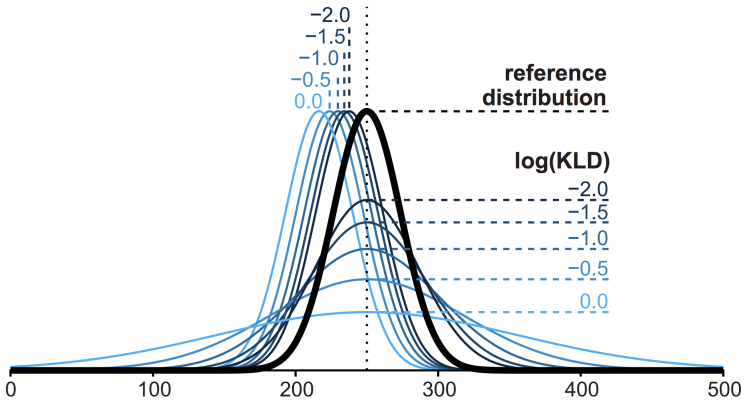


Figure 4.5: A depiction of how log KLD varied by different means and standard deviation (adapted from [77]).

Normative behavior is represented by a smaller log KLD and non-log KLD close to 0. The aggregate log KLD was -2.18 (non-log KLD=0.11), which aligns with previous work that demonstrates people's collective reasoning is more consistent with Bayesian optimal behaviors even when individuals do not necessarily act as a fully Bayesian agent [66].

4.5 S1: Elicitation Techniques and Dataset

Our pilot study used an elicitation technique from the literature which was designed for fitting Beta distributions to participants' responses using a numerical solution [56]. While the technique has been shown to be more robust to imprecision in the elicitation process than several other techniques [153], it is possible that the evidence for approximate or "sample-based" Bayesian inference that we observed was an artifact of the elicitation technique. For instance, by asking for a mode value, it is possible that the technique prompted people to consider only a single sample. We are interested in evaluating how robust our result in our pilot study is to changes in the dataset that is presented. In a pre-registered study,[1] we therefore

[1] http://aspredicted.org/blind.php?x=4bf9ci

evaluate three additional elicitation techniques and introduce a new dataset. The elicitation techniques vary in the degree to which they ask a participant to provide a full distribution versus a small set of samples. By manipulating both representation of uncertainty and the dataset, we aim to gain a better sense of how robust our observation of approximate Bayesian inference is.

4.5.1 Developing Elicitation Techniques and Conditions

We are interested in comparing a set of interfaces which vary in the format they use to elicit participants' responses. We describe two sample-based techniques of our own design, as well as two elicitation techniques from the literature. While our data interpretation task requires eliciting a Beta distribution specifically, we expect that the techniques we evaluate will generalize between Beta distributions, for example to symmetric distributions like Gaussians.

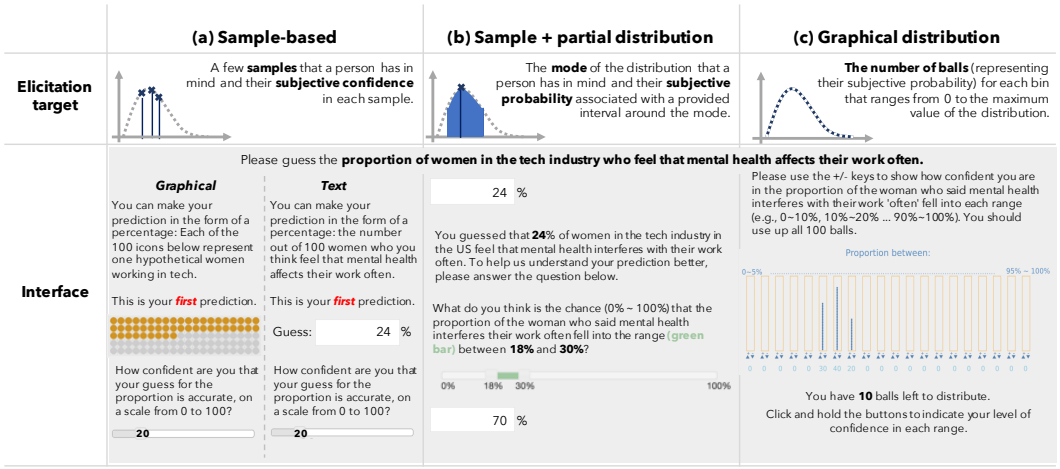


Figure 4.6: Elicitation target and interface. We developed two sample-based techniques (a), and used an interval technique [153] (b) and a graphical "balls and bins" technique [62] (c) from the literature.

Sample-based Elicitation

Evidence from research on reasoning with uncertainty (e.g., on classical Bayesian reasoning tasks [60]) and uncertainty visualization [54, 79, 78, 85, 86] indicates that people are often better at thinking about uncertainty when it is framed as frequency rather than probability. One way to elicit uncertainty is to ask people to provide one sample at a time until they have exhausted their internal representation. Imagine a person provides their expectations for the proportion of women in tech who experience mental issues often. Several possible proportions seem salient to them, including 20% and 33%. We devise a sample-based elicitation method that asks a person to articulate a small set of samples (e.g., 5), one at a time (Fig. 4.6(a)).

Even if people find it easy to reason in the form of samples, we might still expect that they perceive some samples as more likely. A sample-based elicitation technique would not prevent a person from providing the same sample multiple times, proportional to its expected probability (i.e., resampling with replacement) [18]. However, articulating the same sample multiple times can be tedious. For each sample a person provides, our technique asks for a corresponding judgment about the salience of the sample in the form of subjective confidence. Using this technique, the hypothetical person with two samples of 20% and 33% might provide 20% as a first estimate with a higher confidence (e.g., 70 on a scale of 0 to 100), and 33% as a second estimate with a lower confidence (e.g., 30). In practice, the confidence values do not need to sum to 100 as they can be normalized prior to using them to fit the responses to a distribution.

We created two versions of our sample-based elicitation technique. A **graphical sample-based elicitation interface** (Fig. 4.6 (a) left) allows participants to provide a predicted value (i.e., sample) by clicking icons in an icon array. This interface is nearly identical to the visual format used to present the observed data. However, the icon array in the elicitation interface presents 100 circles to imply elicitation in parameter space rather than 158 people icons as in the visualization of the observed

data. An analogous **text sample-based elicitation interface** (Fig. 4.6 (a) right) allows participants to provide a predicted value by entering number in a text box. As a participant provides their samples, each prior sample is appended to the bottom of the interface so that participants can review their samples and corresponding confidence values before submitting the response.

Graphical Distribution Elicitation

To conduct a Bayesian analysis in many domains (e.g., clinical trials, meteorology, etc.), analysts probe domain experts for uncertainty estimates, then use these to construct a prior distribution [112]. This approach generally assumes that people with domain knowledge possess a relatively complete internal representation of the uncertainty in a parameter. Research indicates that a graphical interface that enables constructing a distribution via placing 100 hypothetical outcomes ("balls", or circles representing hypothetical outcomes) in multiple ranges ("bins") allows people to articulate a distribution that they have been presented with more accurately than a method that asks for quantiles of the distribution [62]. We implemented a **graphical "balls and bins" elicitation interface** (Fig. 4.6(c)). Participants are prompted to add exactly 100 balls in bins that span between 0% to 100% in increments of 5% to express the distribution they have in mind. Relative to the text and graphical sample-based techniques we developed, the graphical balls and bins interface encourages a person to consider their entire subjective probability distribution at once.

Sample + Partial Distribution Elicitation

The interval technique we used in our pilot study can be considered a hybrid approach between approaches that emphasize small sets of samples and those that emphasize a full distribution (Fig. 4.6(b)). The mode that a participant provides can be thought of as the most salient sample in their priors. The subjective prob-

ability that a participant provides is analogous to the probability mass of a partial distribution.

As in our pilot study, participants are first prompted to provide a prediction (m). Participants are then asked to provide the subjective probability (sp) that the true proportion falls into the range calculated based on the mode value that they entered ($[m - m*0.25, m + m*0.25]$).

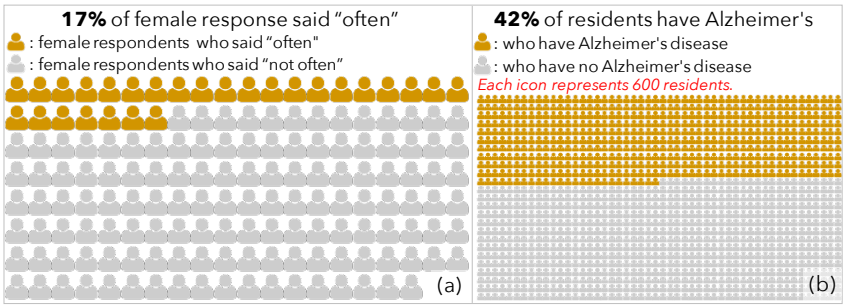


Figure 4.7: The data presentations for S1 (a) and S2 (a, b).

4.5.2 Study Design

Dataset and Presentation:

We reuse the same proportion dataset used in our pilot study (mental health outcomes among women in the tech industry) and the same icon array visualization. However, we are also interested in understanding how robust our findings are to changes in the nature of the observed data. Specifically, the sample size of the observed data directly influences how closely the normative posterior is expected to align with the data. Intuitively, as the sample size of the observed data increases, the impact of the prior distribution on the normative posterior is reduced. With a very large sample, the normative posterior will be virtually indistinguishable from the data even with a reasonably concentrated prior distribution (Fig. 4.8).

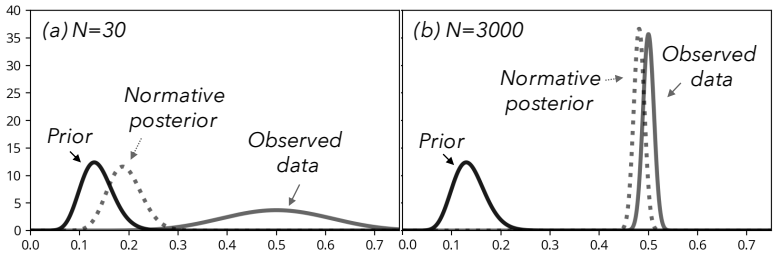


Figure 4.8: The effect of sample size on normative posteriors given the same prior and observed mode.

We therefore chose one additional large sample dataset that has been visualized in the New York Times using icon-style visualizations [17]. This dataset depicts the results of a study of chronic health conditions among assisted living center residents in the U.S. (N=750,000). We chose one type of chronic health condition (Alzheimer's disease or another form of dementia) to formulate our target proportion. We asked participants to reason about "the proportion of residents who have Alzheimer's disease or another form of dementia" in the task. We created a visualization (Fig. 4.7 (b)) that shows this proportion in a similar icon array format to that used for the mental health in tech dataset. Because of the size of the sample, we tell participants that each icon represents 600 residents of assisted living centers.

Procedure

We used the same procedure as in our pilot study (eliciting priors, presenting observed data, eliciting posteriors). However, in Study 2 we randomly assigned participants to one of the four elicitation conditions, and one of the two datasets. On the last page of the experiment, we asked a pre-registered attention-check question about the numeric range in which the observed proportion fell to exclude participants who may not have paid attention to the observed data. Participants were asked to choose an answer among three ranges (0%-30%, 30%-60%, 60%-100%).

Participants

Based on a prospective power analysis conducted on pilot data with a desired power of at least 0.8 assuming α=0.05, we recruited 800 workers with an approval rating of 98% or more (400 per dataset, 200 per elicitation condition) in the U.S from Amazon Mechanical Turk. We disallowed workers who took part in our pilot study. We excluded participants who did not respond correctly to our attention check question from the result. We posted the task to AMT until 800 participants who correctly answered the attention check question were recruited. Participants received $1.0 as a reward. The average complete time was 4.8 minutes (SD=3.35).

4.5.3 Results

Data Preliminaries

For each technique, we aimed to use the simple and most direct technique to fit a Beta distribution, so as to minimize noise contributed by the fitting process. For sample-based elicitation conditions, we used the Method of Moments [68] to estimate distribution parameters (i.e., alpha and beta) using samples provided by each participant. This method provides an estimate using the mean of the samples that participants provided ($\bar{x}$) and the variance of the samples ($\bar{v}$) to calculate beta parameters: $\alpha = \bar{x}(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1), \beta = (1-\bar{x})(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1)$. Since we asked participants to provide their subjective confidence with each sample, we calculated weighted $\bar{x}$ by multiplying the value of each sample by the corresponding confidence value. This approach does not provide a unique solution when the participant provides the same values five times or 0 confidence for all samples. In this case, we gave the participant an uninformative uniform prior ($\alpha = 1, \beta = 1$). 54 out of 400 (13.5%) participants provided responses requiring this adjustment.

An alternative way of fitting these participants' responses is to instead assume a small but varying confidence number for each response and use the Method of Moments to fit a Beta distribution. For example, if the participants provide "50"

five times [50,50,50,50,50], we can alter the response to [50 - (2 * a), 50 - a, 50, 50 + a, 50 + (2 * a)]. We conducted a sensitivity analysis to assess how aggregate and individual log KLD changes as a is varied. We observed that individual level log KLD is relatively stable (i.e., ranges from 0.09 to 1.58 as a varies from 0.001 to 0.05). However, aggregate level log KLD varied across a wider range from -2.44 to 6.543, suggesting that it may be worth trying to collect information that would help disambiguate such cases, or identifying an elicitation method that does not rely on separate weightings for each sample.

Anecdotally, we note that the percentage of "deviant" responses aligns with what Prelec has described as the typical percentage of nonsensical responses obtained through the use of the Bayesian Truth Serum mechanism [119]. For the graphical distribution condition, we also used the Method of Moments approach by considering each ball as a sample known within a 5% (the bin width). For the sample and partial distribution condition, we used the same optimization approach we used in our pilot study.

Residual Analysis and Log KLD

To assess the effect of elicitation technique on individual-level alignment with the normative Bayesian solution, we again plot residuals between normative (predicted) means and standard deviations for each participant and observed means and standard deviations (Fig. 4.9). For the tech dataset (N=158) used in Study 2, we observed a similar pattern as in our pilot study, with errors roughly equally distributed about zero for means, and around zero but with a slight bias toward the amount of variance in priors (i.e., overestimating variance in the data).

For the elderly dataset (N=750,000), residuals for means are again roughly symmetric about zero, but residuals for standard deviations are nearly entirely to the right of zero. This suggests a strong tendency for participants to be more uncertain about the true proportion than they should rationally be, given the size of the

observed dataset.

We see some small differences in residual distributions between techniques. For example, those using the graphical balls and bins interface (Fig. 4.9 fourth column) appear to be slightly more consistent (i.e., more concentrated distribution) and slightly less likely to be biased in their estimates of standard deviation of the elderly dataset (Fig. 4.9 bottom row). We counted the participants whose responses spanned more than three bins, with the number of balls on either side of the center bin differing by less than two balls. 110 out of 200 participants in this condition attempted to create a symmetric distribution across more than three bins (totaling a 15% range) for their posterior distribution. Prior work on graphical elicitation has proposed that the axes ranges of an elicitation interface may implicitly influence the predictions that people "draw" [78]. In the case of the graphical distribution interface, it is possible that participants relied on a heuristic suggesting that distributions should be roughly centered and span more than one bin. The small differences in techniques, however, are far less pronounced that the more obvious differences between participants' residuals for standard deviation for the (large) elderly dataset versus the (small) tech dataset.

Per our pre-registration we constructed bootstrapped 95% confidence intervals for the mean individual log KLD between participants' posteriors and the normative posteriors. We found that on average, the mean log KLDs from all conditions were larger than we would expect if participants are "fully Bayesian" at an individual level, further aligning with what we see in Fig. 4.9. Across both datasets, we saw no consistent effects of the elicitation techniques on alignment between participants' posteriors and the normative posteriors as measured by log KLD.

To disambiguate whether the difference between the tech dataset and the elderly dataset is due to the different domains of the data or the different sample sizes, we introduced additional datasets by manipulating sample size. We reran the study with the sample sizes switched for the two datasets (tech dataset N=720,000, elderly

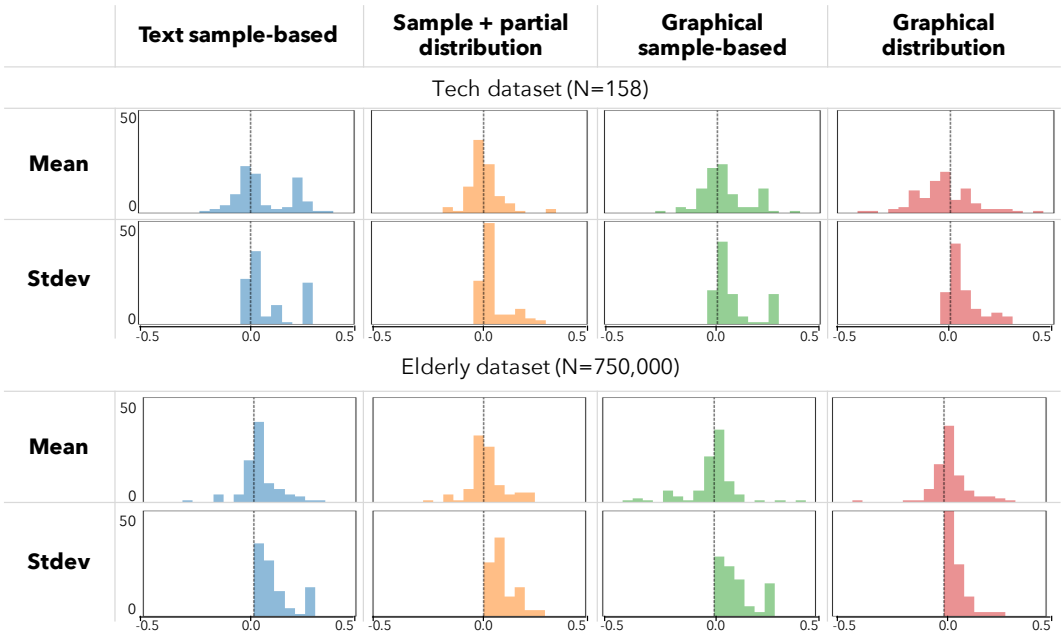


Figure 4.9: Distributions of residuals (observed-predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of normative posteriors.

dataset N=150). We observed the same pattern of results in residual plots, where elicitation techniques did not appear to reliably impact individual's residuals in means or standard deviations, but the larger sample size datasets led to residual standard deviations that were strongly biased toward greatly overestimating the amount of uncertainty one should have given their prior and the observed data. In other words participants did not weight the value of information captured by the observed elderly dataset as much as they should, given its large sample size (N=750,000). We again confirmed these results by examining log KLD. We speculate that the deviation is caused by a well-documented tendency among people to show insensitivity to sample size and its relationship to variance (sampling error) [2], and in particular to be insensitivity to large samples, or exhibit "non-belief in the law of large numbers" [15]. Prior research has demonstrated that when presented

with very large samples, experimental subjects appear to make inferences from"a sample of fixed size" [15]. The proposed formulation describes how a nonbeliever in large numbers make an inference from a distribution with the fixed sample size and highlight the distinction from a person who processes information in a Bayesian way by incorporating the sample size.

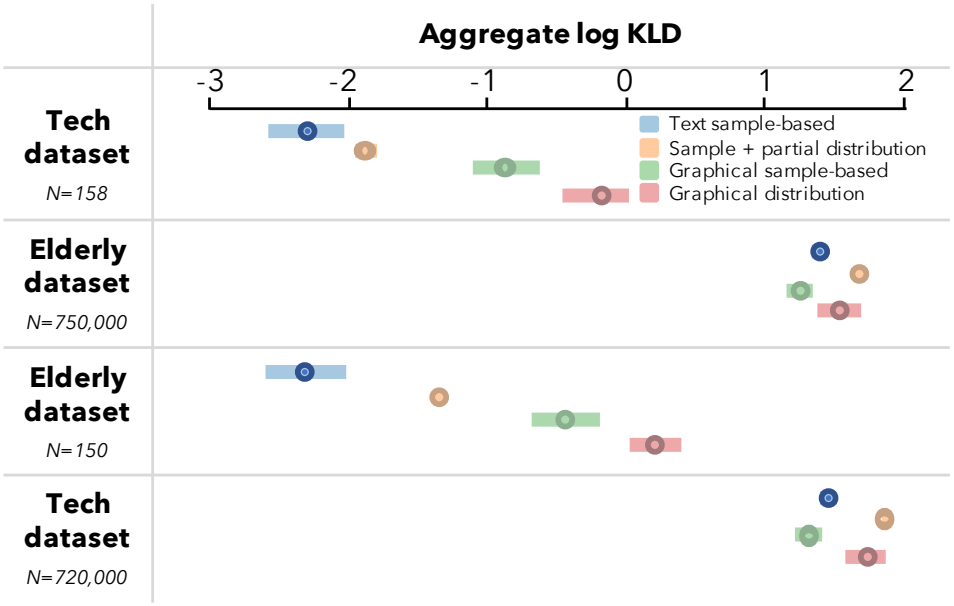


Figure 4.10: Bootstrapped 95% confidence intervals for aggregate KLDs.

We examined the aggregate level log KLD results to confirm what the residual plots suggested regarding approximate Bayesian inference for the smaller sample datasets but not for the larger sample datasets. We found that while participants' responses were consistent with an approximate or sample-based Bayesian hypothesis for the small sample size datasets, we don't see analogous evidence that participants act as sample-based Bayesians for the large sample datasets (Fig. 4.10(a, b)).

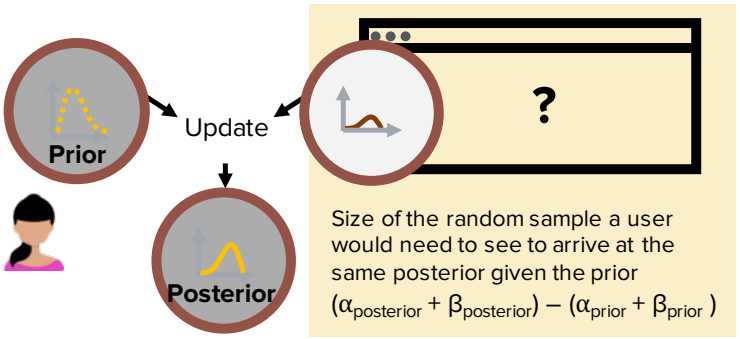


Figure 4.11: An illustration of how one's perceived sample size is calculated. First, we assume that 1) the user did a perfect Bayesian update by treating their posterior as normative posterior and 2) the user would perceive the data at face value. Then we reverse-calculate how they perceived the observed data. This can be thought of as the size of the equivalent random sample that a perfect Bayesian would need to see to arrive at their own posterior distribution.

Perceived Sample Size

One benefit of obtaining distributions rather than just expected values (e.g., [89, 90]) is that we can interpret the parameters of the fitted Beta distributions to gain insight into how participants perceived the data. For a Beta distribution, the two parameters α and β are associated with the sample size that the distribution represents. α stands for the number of successful trials, and β stands for the number of failed trials. By treating the participants' posteriors as normative posteriors and using the elicited priors, we reverse-calculated the perceived observed data distribution ($\alpha_{\text{perceived data}}$ and $\beta_{\text{perceived data}}$) for each participant (in other words, the counts in the equivalent random sample that a Bayesian would have needed to perceive to arrive at that posterior), then summed these two parameters for sample size (Fig. 4.11). An analogous metric has been used to gain insight into how women update their beliefs about the effectiveness of contraceptives [45]. Figure 4.12 shows how the perceived sample size of the observed data was roughly the same across elicitation techniques and datasets. The mean perceived sample size across all tech-

niques for the tech dataset (N=158) was 212.47 (median=41.14) whereas the mean perceived sample size of elderly dataset (N=750,000) was 359.58 (median=51.51), despite the enormous actual difference in the sample sizes of the observed data. (158 vs. 750,000).

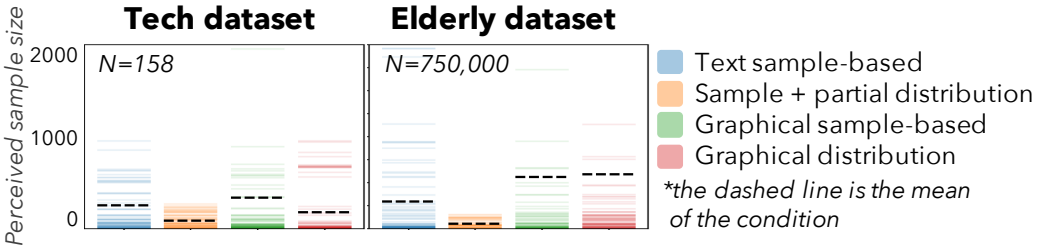


Figure 4.12: Perceived sample size as implied by participants' prior and posterior distributions. Participants perceived similar sample sizes between two very different sized datasets.

4.6 S2: Uncertainty Visualization and Prior Elicitation

We turn now to showing how a Bayesian approach can be used to evaluate how well different visualization alternatives encourage normative interpretations. To test how well the normative Bayesian approach can differentiate uncertainty comprehension and demonstrate its use for evaluation, we designed a study in which some participants are shown visualizations of uncertainty and others are not. Visualizing uncertainty is a natural way to try to help people overcome insensitivity to sample size (Study 1) by providing a more direct, visual way for them to ascertain how much they should weigh the observed data relative to their prior beliefs in formulating posterior beliefs. Through a pre-registered[2] study, we demonstrate how using the Bayesian framework allows us to quantitatively identify that an uncertainty visualization is a better design choice to reduce insensitivity to large sample sizes, compared to a visualization without uncertainty representation.

[2] http://aspredicted.org/blind.php?x=496ri9

4.6.1 Elicitation Technique and Dataset

To evaluate our questions, we used the tech dataset (N=158) and the elderly dataset (N=750,000) that we used in Study 1 (Fig. 4.7). We used the text sample-based technique from Study 1.

We used Hypothetical Outcome Plots (HOPs) [79] as our choice of uncertainty visualization, since HOPs represent a relatively "strong" (i.e., hard to ignore) representation of uncertainty that has been empirically shown to improve uncertainty comprehension among untrained participants over common static equivalents [79, 85]. HOPs convey uncertainty by animating set outcomes randomly drawn from a target distribution. To create HOPs for each dataset, we constructed a binomial distribution using parameters of the dataset (e.g., $\beta(n = 158, p = 0.17)$ for the tech dataset), then sampled multiple hypothetical modes from the distribution to present as hypothetical outcomes, using a frame rate of 400ms as suggested by prior work [79, 85] (Fig. 4.13).

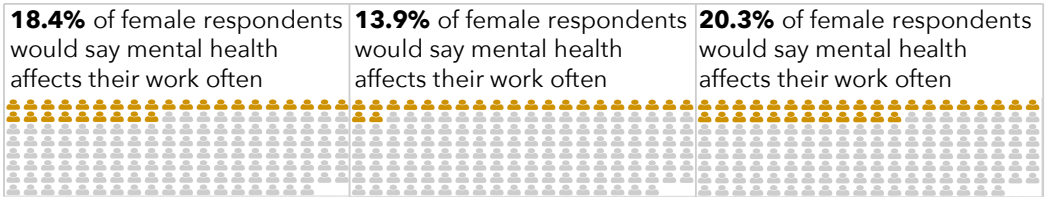


Figure 4.13: The example frames from the HOPs (tech dataset).

4.6.2 Conditions and Participants

In addition to manipulating whether or not participants viewed the uncertainty visualization, we randomly assigned participants to either a prior elicitation or a no-prior-elicitation condition. Prior study shows that participants were slightly anchored by their elicited priors when they recalled the observed data [88]. We therefore wish to examine the effect of prior elicitation on posterior beliefs. With

two interventions (uncertainty visualization, prior elicitation), we arrived at four study conditions (Fig. 4.14). Participants in the **Elicitation-Uncertainty condition** were prompted to externalize their priors before seeing the observed data, then to examine the observed data as HOPs. Participants in the **Elicitation-No uncertainty condition** were prompted to externalize their priors before seeing the observed data, then to examine the observed data as a static icon array as in our pilot study and Study 1. Participants in the **No elicitation-Uncertainty condition** were asked to examine the observed data presented with HOPs but were not prompted to externalize their prior beliefs beforehand. Lastly, participants in the **No elicitation-No uncertainty condition** were asked to examine the static observed data without being prompted to externalize their prior beliefs beforehand. Participants in all conditions used the text sampled-based elicitation interface to provide their posterior beliefs after examining the observed data.

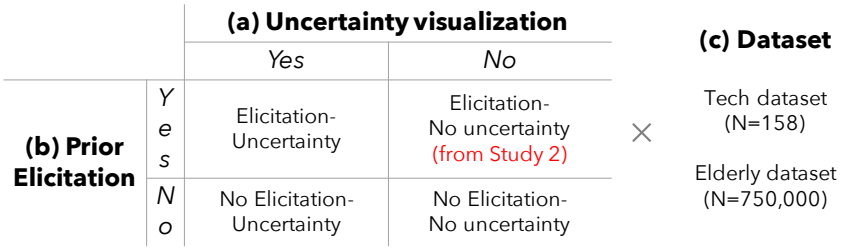


Figure 4.14: Table of Study 2 conditions.

The Elicitation-No uncertainty condition responses consisted of participants' responses from the text sample-based conditions from Study 1 (responses from a total of 200 participants, 100 per dataset). For the remaining conditions, we recruited an additional 600 participants (100 per condition, a total of 300 per dataset) in the U.S from AMT. We disallowed workers who took part in our pilot or Study 1. We excluded participants who did not respond correctly to our attention check question. We posted the task to AMT until 600 participants who correctly answered the

attention check questions were recruited. Participants received $1.0 as a reward.

4.6.3 Analysis Approach

Per our pre-registration we used a Bayesian linear regression implemented in R's *rethinking* package to evaluate the effect of prior elicitation and uncertainty visualization using a single measure (log KLD). We examined residual plots for mean and variance of participants' posterior distributions for all conditions to confirm our model interpretations below.

To compute the normative posterior for No-elicitation conditions, we used the aggregate priors from participants in the text sample-based condition in Study 1 (Tech dataset: $\alpha = 10.79, \beta = 18.99$, Elderly dataset:$\alpha = 31.25, \beta = 39.59$). We specified a model to regress the mean effect in individual log KLD on dummy variables indicating whether uncertainty visualization was shown, whether prior elicitation was prompted and which dataset was presented (tech vs. elderly). We specified identical weakly regularizing Gaussian priors for mean effects (μ: 0, σ: 1) and half-Cauchy priors (Cauchy distributions defined over positive real numbers) for scale parameters (μ: 0, σ: 1). The thick tailed Cauchy distribution tends to be slightly preferable to Gaussian distributions as a weakly regularizing prior for standard deviations [106]. We also included the (mean-centered) time that the participant spent to examine the observed data as a covariate. We present posterior mean estimates of effects with 95% confidence intervals.

4.6.4 Results

Mean completion time was 3.8 min (SD:2.4) for No-elicitation and 4.7 (SD:3.2) for Elicitation conditions.

Impacts on Individuals' Updated Beliefs

Figure 4.15 shows the posterior mean estimates for effects on log KLD. Prior elicitation had no reliable effect on the log KLD of individuals' posterior beliefs relative to the normative Bayesian posteriors (mean:-0.04, 95% CI:[-0.15,0.1]). Log KLD reliably improved when participants were exposed to uncertainty visualization, with log KLDs relative to the normative posteriors for those who viewed HOPs being on average lower by -0.15 (95% CI:[-0.29,-0.04]). Being assigned to view the large sample size dataset (i.e., elderly dataset) still had a very large impact on results at the individual level, with the average log KLD for those who viewed the large sample size dataset being on average 1.54 log KLD units larger than those for the small sample size dataset (95% CI:[1.42,1.67]). Spending more time examining the observed data reduced log KLD but not reliably (mean=-0.07, 95% CI:[-0.14,0.01]).

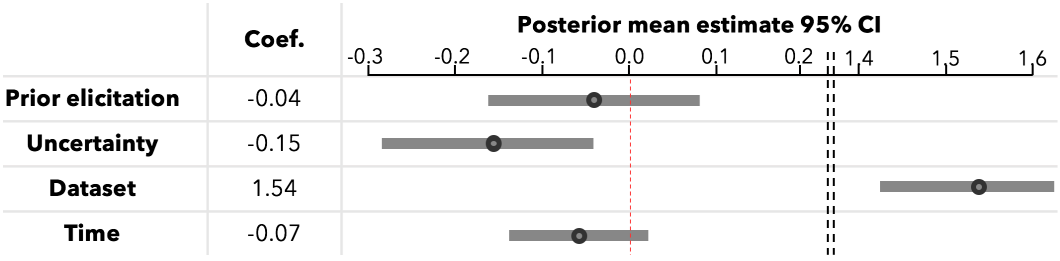


Figure 4.15: Posterior mean estimates of effects with 95% confidence intervals from a model regressing the mean effect on individual log KLD on whether uncertainty visualization was shown, whether prior elicitation was prompted and which dataset was presented. Lower values indicate a greater effect toward lowering log KLD.

Perceived Sample Size

Even though participants assigned to examine the large sample size dataset had high log KLDs relative to the small sample size dataset, viewing HOPs did impact how accurately they perceived the sample size of the observed data. Figure 4.16 shows how the predicted perceived sample size of the observed data based on the dataset and

whether uncertainty (HOPs) was presented. For the tech dataset (N=158), while the means of the No uncertainty and Uncertainty conditions were similar (326.0 vs. 327.3), the median was much closer to the actual sample size of the dataset for the Uncertainty conditions (median perceived: 166.3) than the No uncertainty conditions (median perceived: 97.2). For the elderly dataset (N=750,000), both the mean and median of the Uncertainty conditions were closer to the true observed sample size (mean perceived: 60,268.9, median perceived: 734.0) than the No uncertainty conditions (mean perceived: 809.54, median perceived: 216.1). These results suggest that presenting uncertainty information reduced participants' deviation from normative Bayesian inferences, albeit more for some participants than others.

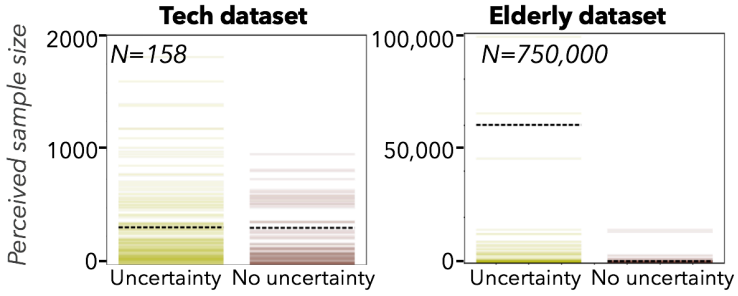


Figure 4.16: Perceived sample size for the tech and the elderly datasets. The uncertainty visualization helps participants more accurately perceived sample size in the both datasets.

4.6.5 Bayesian Inference as Evaluation Metric

Using a Bayesian inference approach, we validated the effectiveness of animated uncertainty visualization to help participants' update their beliefs in a way that more closely resembled normative Bayesian inference. Calculating a proxy for "perceived sample size" helped explain the effect of uncertainty visualization. We propose that visualization designers like journalists or professional science communicators could adopt a Bayesian approach to identify visualizations that better support uncertainty

communication and gain deeper insight into what information users appear to extract from presented data.

4.7 Discussion

Through three studies, we demonstrated how a Bayesian cognitive modeling approach can be used to interpret and evaluate how people update their beliefs after being exposed to a data presentation. With some caveats, our results suggest that in a naturalistic scenario wherein people bring prior beliefs, individuals' interpretation of visualized data can be productively viewed as a Bayesian process. Our experiments showed that on average, people's responses were consistent with a sample-based Bayesian account when examining small sample size datasets. On average, people's responses deviated from Bayesian reasoning when presented with large sample size datasets, aligning with prior evidence of a "non-belief in the law of large numbers" [15]. Some research in behavioral economics attempts to characterize this bias in synthetically designed setting [15, 11, 9], such as the bias arouses when a person predicts where a colored ball is drawn from when two boxes with a different mix of colored balls are given [9]. Future work might test a larger range of sample sizes with a realistic setting to further characterize this bias. It is also worth investigating the degree to which using single icons to represent multiples of people, such as those used in the New York Times visualization upon which we based our large sample visualization (S1), contributes to the insensitivity we observed to large sample size.

4.7.1 Model Assumptions

Our simple model of Bayesian cognition makes several assumptions. We discuss these assumptions in light of possible forms that prior beliefs and assessments of the value of data may take.

What form do viewers' prior beliefs take?

One potential critique of using Bayesian cognition for designing or evaluating visualizations might be that it is unrealistic to expect lay visualization viewers to possess meaningful prior beliefs about phenomena that tend to be presented to the public in outlets like data journalism. Our model makes the following assumptions about a viewer's prior beliefs:

1. Viewers have prior beliefs about parameters estimated by data.

2. Beliefs take the form of a probability distribution (Beta distribution).

Research in probability elicitation from experts has raised the question of whether elicited subjective distributions can in fact be interpreted as the subject's a priori beliefs or whether instead they represent an artifact of the elicitation interface [112]. Recent research in quantam cognition, which attempts to provide formalisms that can explain paradoxes sometimes observed in behavioral experiments, may also be useful for understanding how elicitation interferes with natural reasoning processes [3, 22]. We suspect that the priors and posteriors we elicited in our studies are influenced by both. While it would be impossible to definitely answer this question, understanding the extent to which elicited beliefs are sensitive to the elicitation method has implications for prior and posterior elicitation as a general technique to be used by visualization creators to evaluate designs, or in "human-in-the-loop" data analysis applications that might elicit priors in order to combine them with observed data to make predictions.

On the one hand, if people did not possess priors or were not able to articulate them, we might expect that with the sample-based elicitation techniques, which required providing samples with confidence, we would see a number of unidentifiable distributions due to all zero confidence values, or no variation in the sample values, for example. However, over 85% of participants who used sample-based techniques provided valid probability distributions. Across all elicitation techniques we also saw

that these distributions had predictive power for posterior beliefs. A Bayesian model constructed with personal priors achieved a better fit (using Watanabe-Akaike Information Criterion (WAIC) [146, 59]) than did a model constructed with the aggregate priors or assuming a uniform prior (WAIC = 2315.7, 3159.8, 3159.8, respectively).

However, on the other hand, we can imagine scenarios that might result in ambiguity in elicited distributions. For example, consider a viewer with little experience with the tech industry, whose intuitive beliefs are well described by a few mutually exclusive possibilities. They believe that the population parameter might be near 10%, based on their awareness that roughly 10% of women in the U.S. at large experience depression and their assumption that working in the tech industry is not likely to significantly change that proportion. However, they can also imagine a scenario in which the tech companies that the surveyed women work at over-represents a minority of tech companies that do practice gender bias that the viewer is aware of based on media reports. Assuming this latter scenario, they imagine the proportion would be greater than 10%, perhaps as high as 40%. What sorts of responses might this viewer provide given a location plus interval elicitation technique? Would they interpolate over the two possible proportions to arrive at a modal estimate, even if they perceived the two possible scenarios to be mutually exclusive? Would they report on only the higher probability scenario? Research in aggregation of probability distributions provides some evidence of how people intuitively combine probability distributions such as by using weighted averages of distributions [63]. The possibility that prior beliefs take this form is however acknowledged in a number of settings, from classical decision theory [141] to quantum cognition [22, 135] to proofs of guarantees of Bayesian Truth Serum [120].

Another possibility is that people have valid prior beliefs related to a parameter presented in a data visualization, but that elicitation interfaces like those used in our study assume a greater level of precision or "resolution" to these beliefs than is realistic. For example, perhaps in our proportion examples, many viewers are only

capable of expressing whether the true proportion is above or below some threshold percentage, such as above or below the base rate for women facing mental health issues. In a scenario where the parameter of interest is a trend over time (i.e., slope), viewers may have a sense of whether the slope is positive or negative, but not the ability to precisely specify a slope value to the nearest tenth of a digit, or even to the nearest integer.

Consider the visualization shown in Figure 4.17(a) presented by the New York Times titled 'Where Boys Outperform Girls in Math: Rich, White and Suburban Districts.". This visualization portrays the relationship between the gender performance gap in Math and English scores for third through eighth graders' and family income. This data was collected from nearly all U.S school districts. For simplicity, imagine we would like to elicit a viewer's subjective distribution on the trend in math scores. We can separately elicit the intercept of math scores (i.e., on average, which gender performs better in Math and by how much) and the slope (i.e., how does the gender gap worsen or get better as the average family income in the district increases). For the intercept, it is easy to imagine that some viewers, such as those with prior experience with primary school education outcomes, might possess beliefs about how much better a given gender performs, whereas some other viewers may only feel confident guessing which gender has the advantage. We can develop and test graphical prior elicitation interfaces that vary in the resolution at which they elicit beliefs. For example, in an attempt to capture differences in how confident individuals are in providing beliefs, we are devising an interface that starts by prompting a viewer for lower resolution predictions only (e.g., Do you think that girls do better at math on average? or Boys? or about the same?) but gradually allows them to progress to higher-resolution questions if they express an interest in doing so (e.g, Do you think that girls do better by 0-0.5 grade or 0.5-1.0 grade?) (Fig. 4.17(b)).

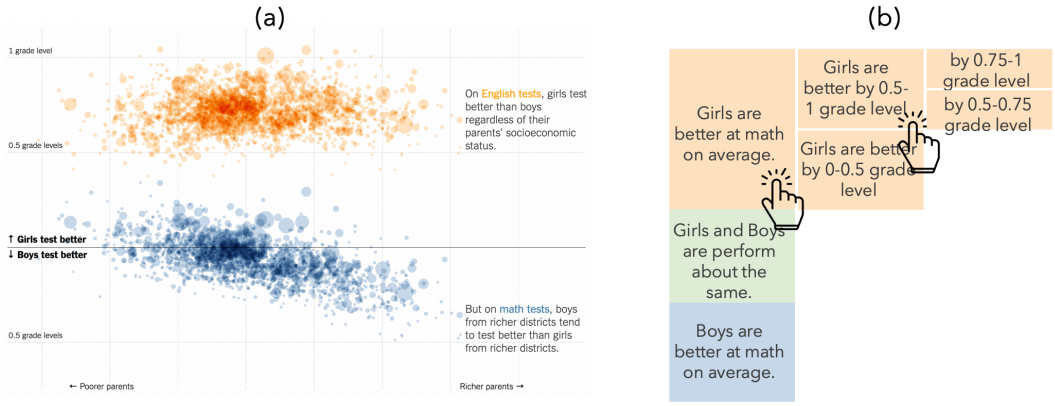


Figure 4.17: (a) A visualization depicts the relationship between the score gap for math and English by gender and the income level of school districts, presented by the New York Times, (b) an interface for eliciting viewers' prior beliefs that progressively prompts a viewer to provide move to a higher resolution prediction if they feel comfortable doing so.

How do viewers perceive the value of data?

Our model also places several assumptions on viewers' assessments of the information value of the data we presented:

1. Data are taken at "face value," such that sample size is the primary indicator of the information value of a dataset.

2. Viewers had not previously encountered the particular data we presented them with.

A possibility that is ruled out by the first assumption above is that people perceive a distribution over the credibility or trustworthiness of different data sources. Under such a possibility, it might be rational to update one's beliefs significantly less than predicted by normative Bayesian inference. While we believe that even a simple model of Bayesian inference can be highly valuable for understanding what factors affect belief updating from data presentations, we are also experimenting

with hierarchical Bayesian models that allow for the fact that viewers may have hyperpriors that specify the extent to which they trust a given information source. While the presented model accounts for prior beliefs regarding the data's *topic*, such a hierarchical model would account for viewers' prior beliefs about the data's *source*.

Wisdom of Crowds vs. Approximate Inference

We attribute our observation of non-rational individuals and rational aggregated behavior to approximate inference as described by Vul et. al [143]. In this view, the approximation process introduces errors into individual judgments, but these errors are averaged out in the aggregate analysis, yielding normative judgments.

However, the well-known wisdom of crowds effect [57] could also explain this discrepancy between individual and aggregate behavior. In this view, errors in individual judgment are introduced by individual biases rather than an approximation process, and the rational aggregate is obtained when those biases are averaged out across a diverse population.

We have adopted the sampling explanation due to existing evidence that human cognition utilizes sampling in at least some cases [19, 98, 21], but note that our proposed method of comparing elicited priors to posteriors is compatible with either scenario.

Re-Use of Samples

The notion that individuals rely on a sampling process for approximate inference introduces one additional assumption into our studies: during the prior-elicitation portion of each trial, the participant produces a set of samples from their prior. We assume that the participant will re-use these samples when updating their beliefs after observing the visualization. If the participant were to take new samples instead, then we could not compare their posterior to the elicited prior. However, we believe this is a reasonable assumption to make as others have found evidence for the re-use

of samples during behavioral tasks [74, 142].

4.7.2 Generalizing Sample Based Elicitation

Our work was not designed to identify a single, optimal technique for eliciting an untrained visualization viewer's subjective uncertainty about a proportion. With the exception of the graphical "balls-and-bins" interface, which tended to produce posterior beliefs that deviated slightly more from the normative solution, the techniques were difficult to distinguish. However we suspect that for more complex datasets and visual representations, the graphical sample-based technique is likely to have advantages due to its flexibility. Imagine observed data visualized as a line chart. A graphical sample-based technique will allow people to simply draw more lines to represent their prior distribution, while other techniques would be more cumbersome (e.g., eliciting intercepts/slopes).

For example, we have applied the graphical sample elicitation approach to evaluate interpretations of the aforementioned New York Times Graphics Department visualization, 'Where Boys Outperform Girls in Math: Rich, White and Suburban Districts," depicting the relationship between the gender performance gap in Math and English scores for third through eighth graders' and the average family income of a school district. We designed the graphical prior elicitation interface to separately elicit a viewer's prior on the intercept for each school subject (how much better or worse on average they think girls are than boys at that subject), and their prior on the slope for each school subject (how much they think score differences for that subject are impacted by the average family income of a school district). The viewer is walked through the creation of several mock visualizations (samples) in which they first position a point cloud for each subject (Math and English) by dragging the cloud of points, the slope of which is initialized to 0, to a y-position on the graph. Once they have specified the intercept for a sample, they are prompted to adjust the slope using a slider. (Fig. 4.18). The sample-based elicitation technique

makes it feasible to collect viewers' subjective distributions beyond univariate data, as the elicitation interface can be designed to replicate the format of the original visualization with multiple prompts. We note that in this particular instantiation, the prior that is elicited is conditional on providing the viewer with information about the income distribution. We believe that for Bayesian methods to be adopted in the design and evaluation of data presentations like visualizations, it will be important to provide guidance to authors on how to determine what the relevant prior is for a given data presentation and how to apply graphical sample based elicitation to obtain it.

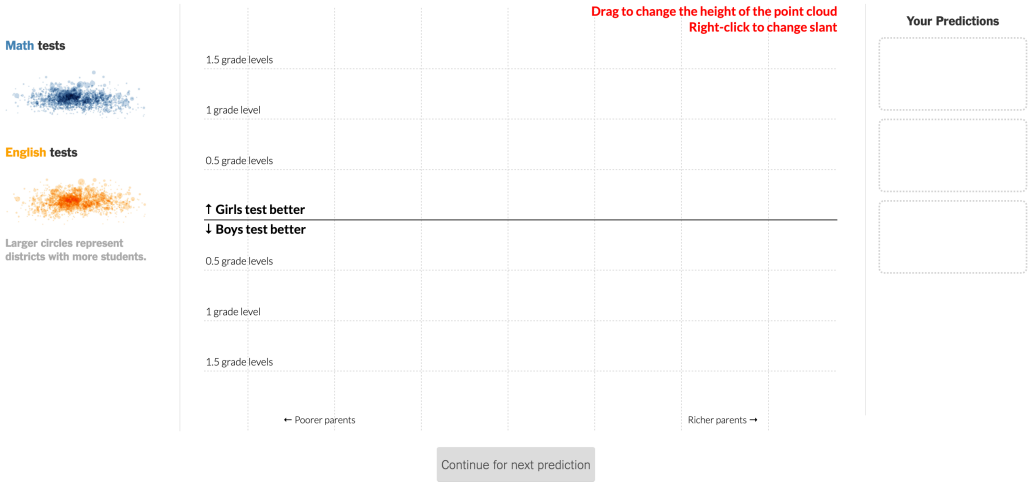


Figure 4.18: The graphical sample elicitation interface. The viewer can drag a cluster (each corresponds to Math and English respectively) to the canvas to set its intercept of the cluster by dragging the cluster and its slope by manipulating a slider that appears when the viewer right-clicks.

Our demonstrations involved presenting a proportion statistic. However even this simple scenario required reflecting on the best way to frame the elicitation of a prior. We chose to elicit prior and posterior distributions directly. These distributions are defined over parameter values (i.e., in model space). We chose to change the icons in the icon array format we used for elicitation and presentation to circles, rather than

human icons, to better align with the notion of eliciting a population proportion. Alternatively, we could have elicited the prior and posterior predictive distributions by asking participants to think about the specific value (e.g., number of women) given some sample size.

4.8 Summary

The results from the studies demonstrate the potential for using Bayesian cognitive modeling to understand how data presentations like visualizations shape beliefs. This chapter demonstrates a path toward better aligning studies of data interpretation with the undeniable effects of prior beliefs and provides a valuable framework for evaluating new presentation methods. In the next chapter, I present how the framework can be applied to better support people's uncertainty comprehension by personalizing data presentations based on an individual's prior distribution.

Chapter 5

BAYESIAN PERSONALIZATION OF VISUALIZED DATA

In chapter 3 and chapter 4, I introduce the exploration of techniques to elicit people's beliefs, the effect of the elicitation, and the methods to apply Bayesian models to improve understanding of visualization interactions. The study shows that the elicitation act itself can improve people's belief updating. In applying Bayesian inference in visualization interpretation scenario, the study result provides evidence that people undervalue the large sample data even when looking at aggregate level statistics. In this chapter, I introduce the investigation personalize data presentations based on an individual's prior beliefs (Research Question 4) to mitigate some of the problems observed in the previous chapter.

5.1 Facilitating Bayesian Update in Visualization Interpretation

Imagine viewing a visualization of poll results describing support for several political candidates among prospective voters early in an election cycle. The data indicates that candidate A has a 10% chance of winning, based on responses from around 600 people, with the chance of winning falling between 8% and 13% with high confidence (e.g., 95%). Imagine that prior to viewing this new result, you encountered the results of a similar poll which estimated support for candidate A at 20% +/- 2.5%, based on responses from around 1000 people. What should you believe after encountering the second poll?

Assuming that you have no reason to distrust either poll, you should update your beliefs proportional to the amount of new information that the second poll provides over the first. Bayesian inference formalizes this intuition by prescribing the use of Bayes rule (i.e., $p(\theta|Y) = \frac{p(Y|\theta)\cdot p(\theta)}{P(Y)}$) to calculate the *posterior beliefs* that one should

arrive at if one updates their beliefs according to laws of conditional probability. In this case, your best estimate of support for candidate A after the second poll should be around 16%, with a high probability of falling within about 2% of that value.

Within a Bayesian mathematical framework, we can reason about "rational" posterior beliefs via simple numeric intuitions. Because there is no reason to distrust either poll, informativeness is directly captured by each poll's sample size, such that the ratio of the informativeness of the second poll to to the first is 3 to 5.

We show how Bayesian inference can be used to improve visualization users' reasoning under uncertainty. Our first contribution is to propose two *Bayesian personalization* techniques that use the mathematical intuitions of Bayesian theory to guide a user's belief formation process as they interact with visualized data. Both techniques treat the user's subjective uncertainty about a parameter value before seeing newly observed data (i.e., their prior distribution) as a reference point against which the uncertainty in the observed data can be compared (Fig. 5.1b2). A *personalized uncertainty analogy* relates uncertainty in observed data to uncertainty in the user's prior. A *personalized posterior visualization* depicts the posterior beliefs predicted by Bayesian inference given the user's prior beliefs.

Does Bayesian personalization work? We present a preregistered experiment with 4,800 participants to investigate how Bayesian uncertainty analogies and posterior visualizations affect belief updating relative to non-personalized uncertainty visualizations. Our experiment compares the deviation between a person's reported beliefs and the normative Bayesian posterior distribution when they use personalized Bayesian personalizations versus more conventional representations of uncertain data like a probability-shaded interval.

Our results suggest the promise of a Bayesian approach to visualization interaction. For small datasets (N=158), both forms of personalization bring the average user's belief updating closer to normative Bayesian inference. Recent work in behavioral economics and beyond [9, 15] suggests that people tend to increasingly discount

the informativeness of data as sample size grows. However, most of these experiments have conducted with synthetically designed stimuli (e.g., coin flip) instead of using real-world stimuli. Our study in chapter 4, and this chapter show that similar bias is induced with realistic stimuli.

We find that for large samples where the topic of the data is perceived as relatively uncontentious (e.g., dementia rates among the elderly), viewing a personalized posterior visualization improves Bayesian updating. However, when the topic of the data leads participants to suspect it was manipulated (e.g., late-term abortion views), Bayesian personalization does not improve reasoning over visualizing uncertainty in the data, suggesting that people may deviate from a Bayesian standard in part as a function of how they perceive the credibility of data.

Moreover, by controlling for the *act of prior elicitation* in our study, we estimate the effect of articulating a prior as a means of prompting better belief updating relative to a Bayesian standard. We find some evidence that simply eliciting a prior from a user can encourage more Bayesian updating as evidenced by people's updating in aggregate being closer to normative Bayesian updating based on prior elicitation alone.

Together our results demonstrate that prior elicitation and personalized Bayesian approaches can help untrained visualization users better update their beliefs than the current best practice of visualizing uncertainty in observed data. We conclude by discussing the implications of our results for the presentation and analysis of data.

5.2 Motivating Bayesian Personalization

We design Bayesian personalizations that promote rational belief updating by exploiting the role of users' priors in Bayesian updating.

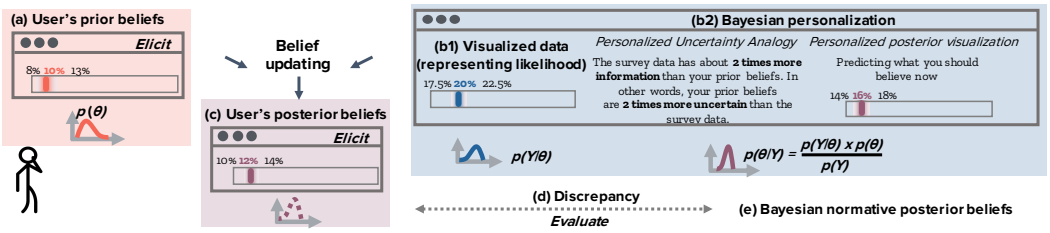


Figure 5.1: Using Bayesian inference to personalize how data is shown to improve belief updating. (a): The viewer holds prior beliefs about a parameter θ such as a disease rate in the population, which are elicited in the form of a probability distribution. (b1): The user is presented with an observed dataset Y estimating the rate, which conveys information about the likelihood function $p(Y|\theta)$. (b2): The observed data is accompanied by personalized information in the form of an uncertainty analogy or visualization of Bayesian posterior predictions derived from their prior beliefs and a normative Bayesian model. (c): The goal of Bayesian personalization is to bring the user's updated beliefs (i.e., posterior beliefs about the probability of θ given Y) closer to (d): the posterior beliefs prescribed by Bayesian inference. (e): In our experiment, we elicit posterior beliefs and use the deviation between these beliefs and the normative beliefs to evaluate the Bayesian personalizations.

5.2.1 Designing Bayesian Personalizations

Using the above intuition and conceptualization, we propose two *Bayesian personalizations* that exploit the user's prior beliefs. A *personalized uncertainty analogy* relates uncertainty in observed data to uncertainty in the user's prior, and a *personalized posterior visualization* depicts the posterior beliefs predicted by Bayesian inference, given the user's prior beliefs.

Personalized Uncertainty Analogy

The user's prior distribution captures their uncertainty about the parameter value before seeing the observed data. We can treat this "subjective" uncertainty as a personally meaningful reference against which uncertainty in the observed data can be compared. Imagine you are presented with a visualization and text telling you how much information the visualized data contains relative to how informed you

were about the topic already: "Your prior beliefs have 2 times more information than the data." To generate the multiplicative factor, we compare κ (a proxy for sample size defined as $\alpha + \beta$) in the prior distribution (κ_{prior}) to the sample size of the observed data (κ_{data}). To avoid multipliers less than one, we always chose the distribution (Beta corresponding to likelihood or participant's prior) for which κ was lower as the reference distribution. For example, if κ_{data} is greater than κ_{prior}, we calculated the multiplier as $\kappa_{data}/\kappa_{prior}$ (e.g., *Your prior beliefs have 2 times more information than the data*), calculating the multiplier as $\kappa_{prior}/\kappa_{data}$ in the case where κ_{prior} was greater.

Personalized Posterior Visualization

An even more direct way to guide a user toward Bayesian inference is to present them with the normative belief distribution calculated using their prior beliefs and the likelihood. Imagine that in addition to an observed dataset, you are presented with a visualization of the normative posterior calculated using your prior distribution, along with a brief explanation of how it was derived (i.e., by combining the information in their prior beliefs with that in the observed data).

5.3 Experiment: Bayesian Personalizations

We designed and preregistered a large crowdsourced between-subjects experiment to investigate whether personalizing the presentation of visualized data using Bayesian theory can improve how well a user updates their beliefs in light of new data. Specifically, our experiment evaluates how participants' beliefs change as a result of the Bayesian personalizations compared to more conventional depictions of proportion estimates as might be seen in science, the media, or government.

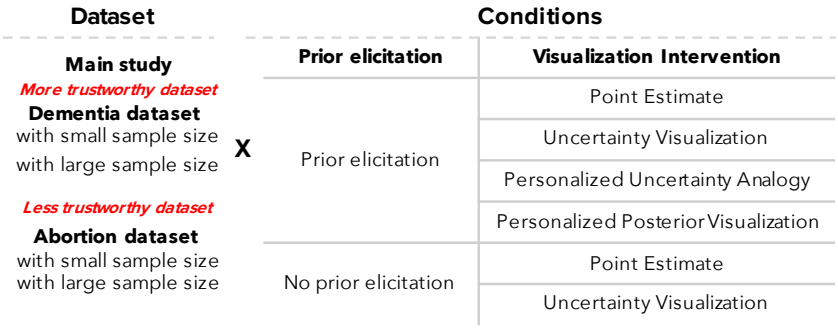


Figure 5.2: The study conditions and datasets.

5.3.1 Study Conditions and Research Questions

We tested four approaches to conveying uncertainty around a parameter estimate (Fig. 5.2).

- **Point Estimate**: Participants view a point estimate of the observed proportion with the size of the sample in text.
- **Uncertainty Visualization**: Participants view a point estimate of the observed proportion along with a probability density shaded interval in which the estimate is expected to fall with high probability (95%).
- **Personalized Uncertainty Analogy**: Participants view the uncertainty visualization alongside a personalized uncertainty analogy. A brief explanation of how the analogy was generated (e.g., "We directly compared the sample size of the study to the sample size implied by your prior beliefs.") is also presented.
- **Posterior Visualization**: Participants view the uncertainty visualization alongside a visualization of the normative posterior distribution. A brief explanation of how the posterior was arrived at (including an analogy expression comparing the uncertainty in the participant's prior beliefs to that of the data as above) is presented.

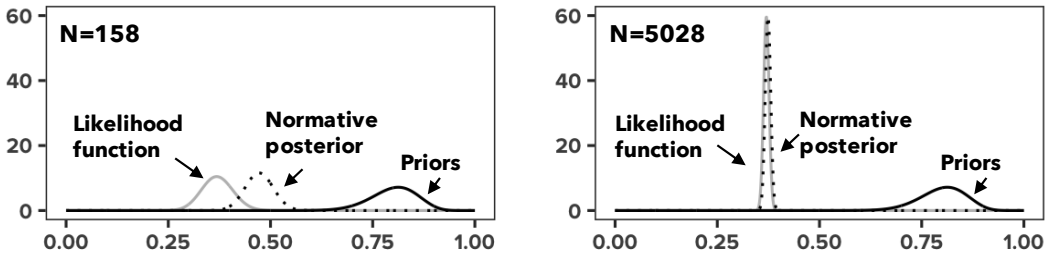


Figure 5.3: Illustration of how normative posterior beliefs (dashed) are influenced by the sample size of the observed data (represented by the likelihood in gray) given a prior distribution (solid). Assuming a relatively weak prior, when the sample size is small, the normative posterior distribution is located between the likelihood and the prior. Assuming the same prior and a large sample observed dataset, the normative posterior distribution is nearly identical to the likelihood function.

Robustness to Varying Sample Size

As Fig. 5.3 left shows, a weak prior belief distribution still has a demonstrable impact on the normative posterior beliefs when the observed data is relatively small (N=158). For a larger sample (N=5208) the normative posterior distribution is nearly identical to the observed data (Fig. 5.3 right). By varying sample size, we use our experiment to investigate whether a tendency for people's posterior beliefs to deviate more substantially from the normative posterior distribution for large samples found in previous chapter holds for our participants as well. We chose 158 and 5,208 as samples in the low thousands are common in presentations of poll or survey results that people encounter in everyday life.

Robustness to Topic Controversy

Besides misunderstanding uncertainty, not trusting that a dataset is a faithful depiction of reality is another possible reason for the deviation between one's posterior beliefs and the normative Bayesian posterior.

To investigate the impact of the perceived "controversialness" of data on the

effects of Bayesian personalization, we identified two datasets that vary in how likely they are to be perceived as having been manipulated. We recruited 200 Mechanical Turk workers in the U.S. with approval ratings of 97% and above. Participants viewed pairwise combinations of six datasets: the proportion of 1) residents of U.S. assisted living centers residents who have Alzheimer's or other dementia, 2) corn production relative to other grain production in the U.S, 3) patients in the U.S who misuse opioids prescribed for chronic pain, 4) foreign-born residents in the U.S., 5) adults in the U.S who think third trimester abortion should be illegal regardless of circumstances, and 6) adults in the U.S who support the death penalty. In a first session, on each trial the participant saw a pair of dataset descriptions (i.e., a summary of the variable) side by side. Participants were asked to choose one dataset that "seems more likely to be tampered with or manipulated to persuade" using a radio button. Participants viewed a total of 15 pairs (trials). In the second session, participants viewed the same 15 pairs but where the original proportion from the source is presented with a 95% highest density interval calculated by for an assumed sample size of 158. We randomized the order of pairs in both sessions.

We ranked the datasets by perceived manipulation using the sum of participants' votes per dataset. The proportion of U.S. assisted living centers residents who have Alzheimer's obtained the fewest votes across both questions, while the proportion of Americans who believe long-term abortions should be illegal unilaterally obtained the most.

Impact of Prior Elicitation

It is possible that prior elicitation itself may affect how "Bayesian" a person appears to be, for example if it encourages the user to be more sensitive to uncertainty in the data. We include two conditions for which we do *not* elicit prior beliefs–No Elicitation-Point Estimate and No Elicitation-Uncertainty Visualization–and use them to evaluate the impact of elicitation on deviation from normative Bayesian

belief updating. Though individual-level updating with and without elicitation cannot be directly compared without eliciting the individual's prior, an aggregate-level analysis, in which we assign No Elicitation conditions a common prior learned from many participants, allows us to observe how elicitation appears to change updating at an aggregate level.

5.3.2 Experiment Design and Procedure

We ran our experiment as a between-subjects study. Participants were randomly assigned to one of the six elicitation and visualization conditions and one of four datasets (small or large dementia dataset or small or large abortion data) (Fig. 5.2). We pre-registered our conditions, sample sizes, and analysis[1]. An introductory page described the dementia datasets (originally from the U.S. National Center for Health Statistics [17]) as having been collected by a national health agency, and the abortion datasets (originally from FOX News [12]) as having been collected by a media outlet.

Prior Belief Elicitation

Participants assigned to elicitation conditions first provided their prior beliefs (Fig. 5.4). We designed an interface that prompted the participant to enter their best estimate of the parameter of interest (e.g., the percentage of assisted-living center residents in the US have Alzheimer's or dementia, Fig. 5.4a), following prior research in proportion prior elicitation from experts [153]. A two-handled slider then appeared, representing an interval around the value they provided as their estimate, with endpoints at 0 and 100%. Participants were asked to specify a range around the value by dragging the end of the interval until its width aligned with how uncertain they felt about the true percentage (Fig. 5.4d). Participants were explicitly told that if their estimate represented a truly random guess, then their interval should span

[1] https://aspredicted.org/blind.php?x=sq3xz8, https://aspredicted.org/blind.php?x=2uc84m

from 0 to 100%; otherwise they should adjust the ends of the interval to make it smaller (Fig. 5.4b). When the participant interacted with either handle, we updated the concentration parameter (κ) based on the handle's value and the mode, then calculated the other handle's location to reflect the 95% interval of the new Beta distribution. Specifically, κ is inversely proportional to the width of the elicited interval. Text above the slider reflected the specified prior (e.g., *You think the percentage is almost certainly no less than 15% and no more than 33% and it's most likely around 23%*, Fig. 5.4c).

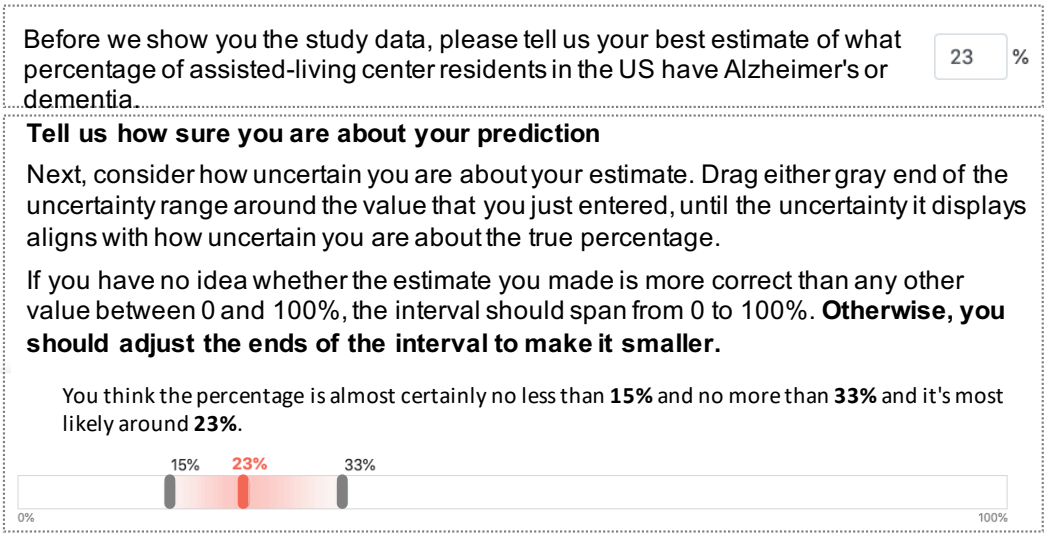


Figure 5.4: The elicitation interface. First, the participant enters a point estimate (top), then they specify how certain they are about their estimate by dragging either end of the interval (bottom). When the participant interacts with either handle, the other handle updates to accommodate the updated Beta distribution.

Presentation of Observed Data

After prior elicitation, all participants examined the observed data. To create the visualization stimuli, we used the proportions from the original source of the datasets (dementia dataset: 42%, abortion dataset: 37%) and varied the sample size that a

participant was assigned (small: 158, large:5208). Participants in the Point Estimate conditions saw the point estimate of the proportion plotted with the number of successes and sample size in text only (Fig 5.5a). Participants in the Uncertainty Visualization and Bayesian personalization conditions saw the point estimate plotted with an interval depicting the lower and upper bound of the corresponding Beta distribution for the Binomial likelihood function, with shading proportional to probability density (Fig 5.5b).

Presentation of Bayesian Personalizations

After viewing the observed data and prior visualization, participants in the personalization conditions then clicked to examine the personalizations, which appeared below the visualization of the observed data. For participants in the Analogy condition, we presented an analogy in text (Fig. 5.5c). For participants in the Posterior Visualization condition, we presented a visualization like our uncertainty visualization of the observed data, but where the distribution shown is the Beta distribution corresponding to the predicted posterior from our Bayesian model (Fig. 5.5d).

Posterior Belief Elicitation and Post-Task Questions

All participants then submitted their posterior beliefs on the next screen. On a final screen, participants were asked demographic questions (gender, education level, and age), and how likely they thought it was that the data was manipulated on a five point Likert scale with endpoints labeled Not at all likely (1) and Extremely likely (5). The final screen asked participants what proportion corresponded to the observed data they had been shown via multiple choice (Below 30%, between 30% to 60%, above 60%) as a preregistered exclusion criteria to filter participants who were not paying attention from analysis.

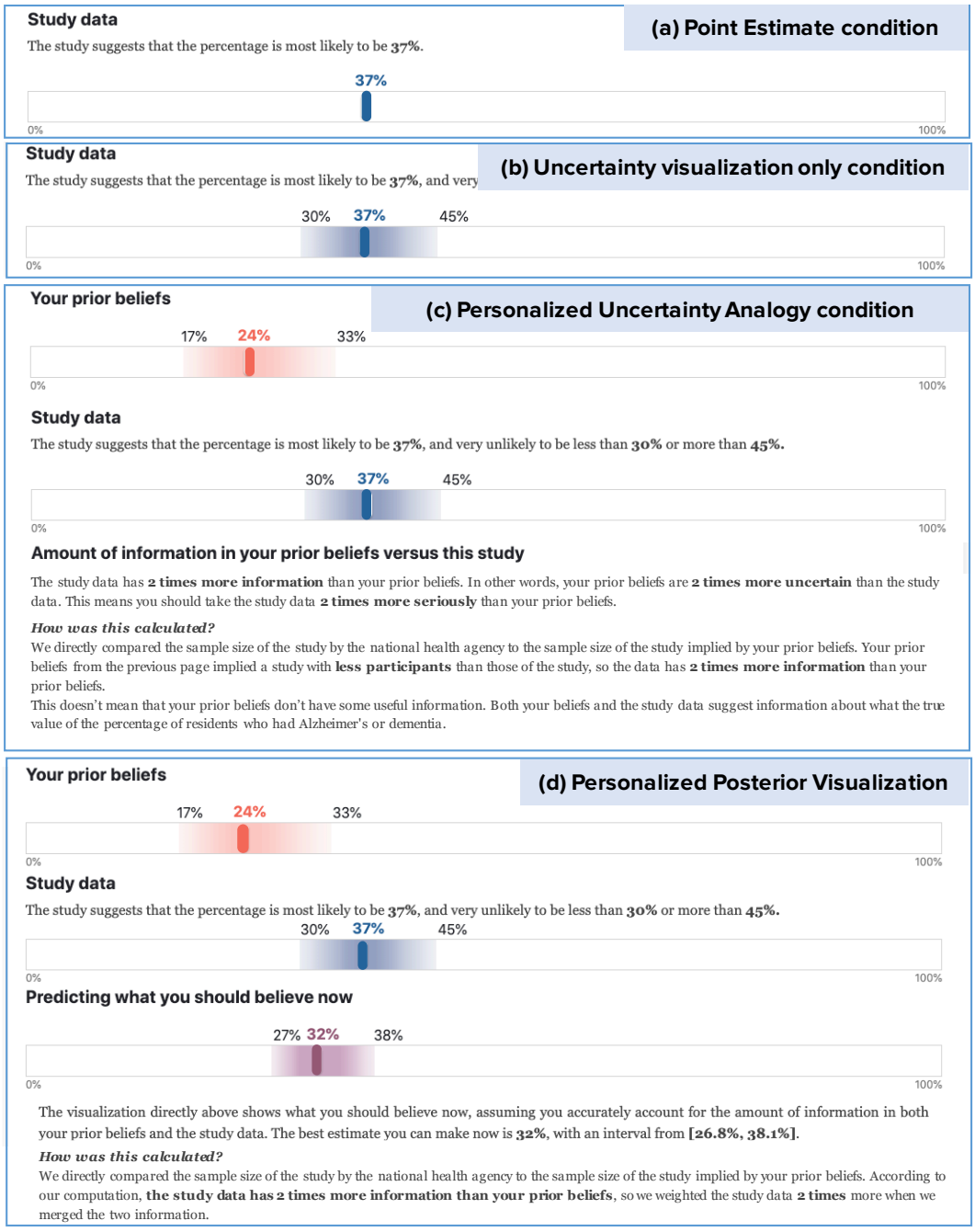


Figure 5.5: Conditions in our experiment, including visualizing observed data as a point estimate with sample size, using a high probability interval with shading to visualize uncertainty in the observed data only, providing an uncertainty analogy based on the participant's prior, and providing a predicted posterior visualization based on the user's prior.

Participants

We recruited participants on Amazon Mechanical Turk, removing those who failed the preregistered exclusion criteria question (total 182), and recruiting more until each condition had 200 participants (total 4,800). We made the HIT available to U.S. workers with an approval rating of 97% or more. The HIT carried a reward of $0.8, which we calculated to ensure that the majority of workers would receive minimum wage according to pilot study completion times.

5.4 Results

5.4.1 Data Preliminaries

The average completion task time was 3.6 min (SD: 6.6). We observed no reliable differences in participants' demographics between conditions.

To analyze participants' responses, we fit the elicited beliefs to a Beta distribution. We treat the elicited point estimate as the mode of a Beta distribution (ω) and the width of the interval as the concentration parameter (κ) to fit a distribution using optimization as suggested by prior work [153]. To compute each participant's normative posterior distribution, we used the relationship between the posterior Beta parameters and those of the prior and likelihood deriving from Bayes' rule to calculate the normative Beta posterior distribution (Eq. 4.1).

5.4.2 Outcome Measures

We treat the deviation between the participant' actual posterior beliefs and the normative posterior beliefs as a proxy for *how well* the participant appears to have interpreted the information contained in the observed data and combined them with their knowledge they already had. We analyzed the deviation in two ways. First, to provide intuition for how participants updated in terms of the familiar notions of a distribution's location and variance, we compared the *location* (i.e., mean) and

the *variance* of each participants' posterior distribution to those of the normative posterior distribution.

Second, we pre-registered an analysis using KL Divergence (KLD) to measure the difference between a participant's stated posterior beliefs and the normative posterior distribution from our Bayesian models. KLD captures the information loss when representing a target distribution p with a second distribution q [95].

5.4.3 Overview of Updating by Location vs. Variance

We analyzed qualitative differences in how participants updated their beliefs across datasets and visualization conditions.

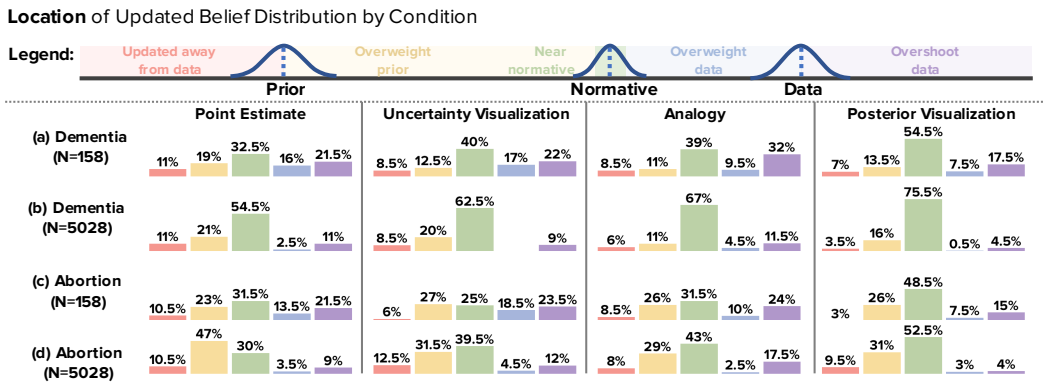


Figure 5.6: Categorization of the location of participants' updates relative to the predictions of normative Bayesian inference for that participant. Each participant was categorized according to the relationship between the mean of their posterior distribution relative to that of their prior distribution, the normative posterior distribution, and the likelihood function.

Location of Updated Belief Distribution by Condition

We categorize participants into five "update types" based on the location (i.e., mean) of their posterior distribution relative to their prior distribution, the normative posterior for that participant, and the likelihood (Fig. 5.6). The legend shows a

hypothetical participant for which the mean of their prior distribution was smaller than that of the likelihood; our analysis also includes the opposite case (i.e., the mean of the participant's prior was greater than the mean of the likelihood). Bottom: Each participant was categorized according to the relationship between the variance of their posterior distribution relative to that of the normative posterior distribution.

We use *near normative* when the location of the participant's posterior is within a relatively small window of the normative posterior (e.g., +/- 2%). We use *overweight prior* for cases where a participant overweighted their prior distribution relative to the predictions of normative Bayesian updating, and *overweight data* for cases where the participant's posterior fell between the prior and likelihood but was closer to the likelihood than predicted by normative Bayesian updating. While most participants' posterior distributions fell, as we might expect, somewhere between their prior distribution and the likelihood, we use *updated away from data* for cases where participant's posterior moved in an opposite direction from the likelihood as well as their prior. We use *overshoot data* for cases where the location of the participant's posterior surpassed or "overshot" the observed data.

Figure 5.6 characterizes participants' updating behavior by dataset and visualization condition according to these categories. Overall, the *near normative* type was the most frequent across datasets and conditions, suggesting that people are approximating Bayesian updating in terms of the location of their distributions. Participants in the Point Estimate conditions (first column in Fig. 5.6) were the least likely to fall in the *near normative* category, and those in the Posterior Visualization conditions (last column) were the most likely to.

Overweighting one's prior was, however, more common in two conditions: the Point Estimate for the large abortion dataset and Uncertainty Visualization for the small abortion dataset. The greater tendency among participants to perceive the abortion dataset as having been manipulated may have led participants to adhere more strongly to their prior beliefs.

Similarly, when comparing the ratio of the *overweight prior* type between dementia datasets (row a and b) and abortion dataset (row c and d), more participants overweighted their priors when they examined abortion datasets.

Figure 5.6 also indicates that the analogy conditions resulted in the highest ratio of people who overshot the likelihood across datasets. The vast majority (roughly 95%) of our participants had more uncertain priors compared to the likelihood, leading to multipliers greater than one. It is possible that imprecise mental calculations led analogy participants to overcorrect.

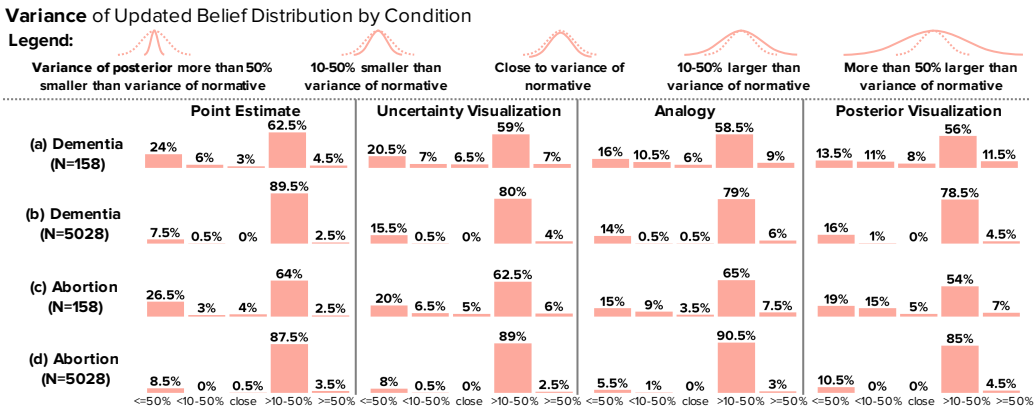


Figure 5.7: Categorization of the variance of participants' updates relative to the predictions of normative Bayesian inference for that participant. Each participant was categorized according to the relationship between the variance of their posterior distribution relative to the normative posterior distribution.

Variance in Updated Beliefs by Condition

To contextualize how the amount of uncertainty implied by participants' posterior beliefs compared to the amount predicted by normative inference, we categorized patterns in variance updates (Fig. 5.7). Because the deviation in elicited posterior versus normative posterior variance was considerably larger than that for means, we categorized participants as *close to normative* if the participant's posterior was

within 10% of the variance of the normative posterior. We similarly categorized participants whose posterior variance was more than 50% smaller than the variance of the normative posterior, as well as 10-50% smaller, 10-50% larger, or more than 50% larger.

Comparing the distribution across categories in Figure 5.7 to that in Figure 5.7, it is clear that participants' deviations from normative inference are driven primarily by non-Bayesian updating of the variance of their beliefs. Additionally, in contrast to the results on location updating, we see no clear advantages of the Bayesian personalizations in reducing errors in variance updating. Regardless of the specific dataset, most participants provided posterior beliefs the variance of which was 10%-50% higher than the variance of the normative posterior. Hence, participants remained more uncertain about the parameter value than they should have in general. Possible drivers of this pattern include unmodeled predictors (e.g., a person's relative trust in data relative to a Bayesian), error in elicitation, or non-Bayesian updating.

Variance results again suggest a difference between the dementia datasets (row a and b) and the abortion datasets (row c and d). Specifically, around 30% of participants assigned to examine the dementia datasets were more certain than the normative posterior. However, for those who saw the abortion datasets, this number dropped to less than 15% of participants. Similar to the location results, this may suggest that the controversial nature of people's values around abortion rights led participants to discount the information in the dataset and maintain more uncertainty in their beliefs.

5.4.4 Preregistered Models: Updating by Log KLD

Per our pre-registration, we specified four Bayesian linear regressions, one for each dataset we presented to participants (dementia N=158, dementia N=5208, abortion N=158, abortion N=5208). These regressions estimate differences in the *distribu-*

tions of KLD, a singular measure of deviation between each participant's updating and normative Bayesian updating, by condition.

$$
\begin{aligned}
kld &\sim \text{dlnorm}(\mu, \sigma) \\
\mu &= \mu_{intcp} + \mu_{post\_vis} * PostVis \\
&+ \mu_{anlg} * Analogy + \mu_{point\_est} * PointEst \\
\log(\sigma) &= \sigma_{intcp} + \sigma_{post\_vis} * PostVis \\
&+ \sigma_{anlg} * Analogy + \sigma_{point\_est} * PointEst \\
\mu_{intcp}, \mu_{post\_vis}, \mu_{anlg}, \mu_{point\_est} &\sim dnorm(0, 5), \\
\sigma_{intcp}, \sigma_{post\_vis}, \sigma_{anlg}, \sigma_{point\_est} &\sim dnorm(0, 2.5)
\end{aligned}
$$

Each model consisted of two submodels. The first submodel predicted **bias** (mean error) in log KLD, capturing how closely participants' response distributions aligned with the normative Bayesian prediction by condition. We use log KLD in our analysis to reduce the impacts of outliers we observed across conditions on our estimates, as KLD grows rapidly as the two distributions diverge more.

The second submodel regressed **dispersion** (variance) in log KLD in log space on the same variables, capturing how much variation there was between participants' deviations from normative inference in a condition. In addition to lower bias, lower dispersion (i.e., more consistent) estimates of log KLD means a technique reduces noise.

We implemented each model in R's rethinking package [105], using weakly-informed Gaussian prior distributions centered around 0 for bias and dispersion. We used dummy variables to indicate whether the participant was shown an uncertainty visualization, an analogy, or a posterior visualization.

We report the result for each condition and dataset relative to a participant in the Uncertainty Visualization condition, as visualizing uncertainty is arguably the best choice a designer could make outside of personalization. We provide coefficients

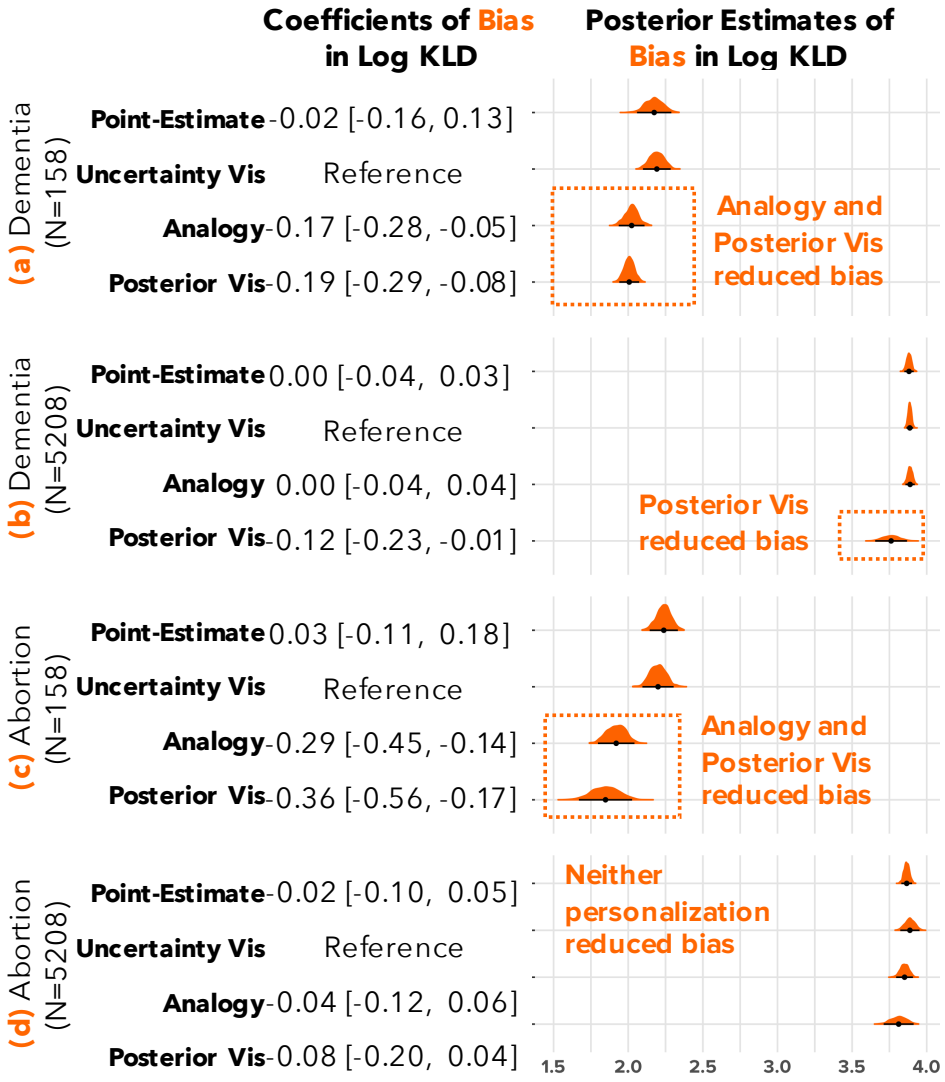


Figure 5.8: Posterior estimates of bias (mean error) of log KLD with 95% credible interval by condition. Results for the dementia datasets are presented in the top row, and for the abortion datasets in the bottom row. Annotations describe effects relative to visualizing uncertainty in observed data (Uncertainty Vis).

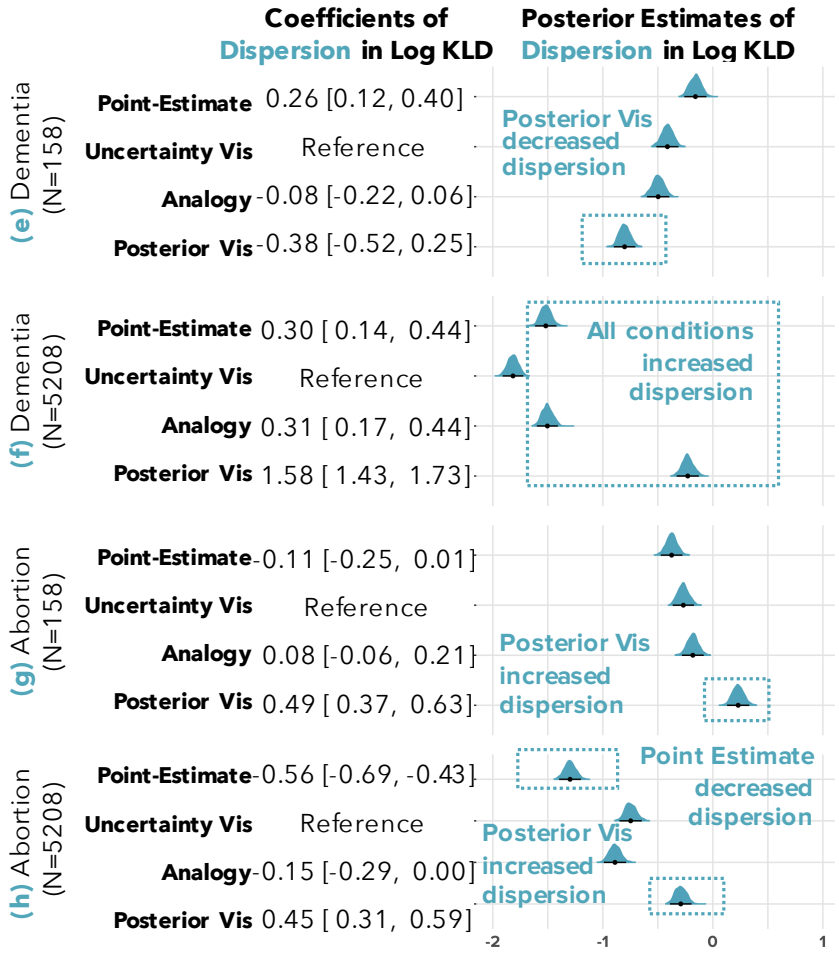


Figure 5.9: Posterior estimates of dispersion (standard deviation) of log KLD with 95% credible interval by condition. Results for the dementia datasets are presented in the top row, and for the abortion datasets in the bottom row. Annotations describe effects relative to visualizing uncertainty in observed data (Uncertainty Vis).

for both submodels in Figure 5.8 and Figure 5.9 left. For readers familiar with statistical significance, we say that a condition has a reliable effect over uncertainty visualization when its 95% Percentile Interval (PI) (reported in text) does not overlap with 0 (which would indicate the possibility of no effect). We visualize posterior estimates of expected bias and dispersion in log KLD by condition (Fig. 5.8 and Fig. 5.9, right).

To further contextualize the size of the effects in bias and dispersion, we also report Cohen's d [33] and Common Language Effect Size (CLES [107]), measures of standardized effect size, using our model results. Cohen's d captures the number of standard deviations by which two means differ, while CLES describes what percentage of the time a randomly drawn sample from one distribution would have a higher value than a randomly drawn sample from the second distribution. To calculate effect size on our model estimates, we first constructed an aggregated posterior distribution for each condition, using the bias posterior estimates from the bias submodel and dispersion posterior estimates from the dispersion model. We compute effect size by comparing the distribution of the personalization conditions with that of the Uncertainty Visualization condition.

Dementia Dataset

Small sample (N=158): Relative to the Uncertainty Visualization condition, both Bayesian personalizations reliably decreased **bias** in log KLD by similar amounts (-0.19, -0.17 respectively; Fig 5.8**a**). Viewing a Point Estimate was not distinguishable in log KLD compared to viewing an Uncertainty Visualization.

Our characterization of updating by location and variance (Sec. 5.4.3) suggested that the personalized Posterior Visualization helped participants correctly update the location of their beliefs. Hence, the **bias** reduction in log KLD may be driven by better location updating among Posterior Visualization participants. On the other hand, our earlier analysis (Fig. 5.6, Fig. 5.7) indicates that the location updating of

participants in the Analogy condition and the Uncertainty Visualization condition for the small dementia dataset are similar. Hence the reliable improvement in updating we observe for the Analogy condition may be driven more by better variance updates than better location updating.

Our **dispersion** submodel indicates that the Posterior Visualization led to more consistent values of log KLD among participants compared to Uncertainty Visualization, with an estimated reduction in dispersion of 0.39 (Fig 5.9e). Seeing an Analogy did not noticeably affect dispersion compared to the Uncertainty Visualization. However, viewing a Point Estimate increased dispersion in log KLD relative to Uncertainty Visualization.

Cohen's d for the Posterior Visualization was 0.33, equivalent to a CLES of 59%. Hence, a participant from Posterior Visualization conditions will have lower log KLD than a participant from the Uncertainty Visualization condition 59 out of 100 times when we randomly select a participant from each condition. Cohen's d for the Analogy personalization was 0.27, equivalent to a CLES 57%.

Large sample (N=5208): Relative to the Uncertainty Visualization condition, viewing a Posterior Visualization reliably reduced **bias** in log KLD, but viewing an Analogy or Point Estimate had no observable effect (Fig. 5.8b).

While highly variant, the distribution of **bias** in log KLD for the Posterior Visualization condition does not overlap with the distributions of expected **bias** for the non-Bayesian conditions (Fig. 5.8b right). However, the distribution of expected **bias** for the Analogy condition is not distinguishable from the Point Estimate and Uncertainty Visualization conditions. Again, our earlier analysis of location and variance updates (Fig. 5.6, Fig. 5.7) suggests that participants in the Posterior Visualization conditions were better at updating the location of their posterior.

All conditions reliably increased **dispersion** in log KLD relative to Uncertainty Visualization (Fig. 5.9f)

Cohen's d for the Posterior Visualization was 0.21, equivalent to a CLES of 56%.

Abortion Dataset

Small sample (N=158): Similar to the small dementia dataset, the Analogy and Posterior Visualization both reliably reduced **bias** in log KLD relative to the Uncertainty Visualization (Fig. 5.8**c**) while the Point Estimate condition was not reliably different.

Compared to the small sample dementia dataset, being in the Posterior visualization condition resulted in higher estimated **dispersion** in log KLD (Fig. 5.9**g**).

Cohen's d for the Analogy and Posterior Visualization were 0.35 (CLES 59%).

Large sample (N=5208): In contrast to the large dementia dataset, neither the Posterior Visualization nor the Analogy condition reliably reduced **bias** in log KLD for the large abortion dataset (Fig. 5.8**d**). A Point Estimate also did not reliably differ from Uncertainty Visualization. We suspect that any effects of Bayesian personalization were too small to observe in light of the rather large discrepancies we observed between participants' posterior beliefs and the predictions of normative Bayesian inference with regard to variance (Fig. 5.7).

We see slightly different patterns compared to the large sample dementia dataset when it comes to effects on **dispersion** in log KLD. Viewing an Analogy slightly decreased **dispersion** in log KLD while viewing a Point Estimate had a stronger decreasing effect (Fig. 5.9**h**).

5.4.5 Conceptual Replication of Sample Size Effect

Our results conceptually replicate a difference in how closely the updates of untrained participants resemble Bayesian updating when shown a small versus a large dataset observed in behavioral economics [9, 15] and what we observe in our own study in the previous chapter. While participants assigned large datasets appear to update closer to normative Bayesian inference when we look at location of posterior beliefs (e.g., compare row a and b, and row c and d in Fig 5.6), the opposite is true when we look at the variance of their posterior beliefs, where deviation from normative Bayesian

inference is substantial. The average bias in log KLD across participants was 0.90 (median:0.93, IQR:0.23, KLD: 11.24) for small datasets, and much higher for large datasets (mean: 1.67, median:1.68, IQR:0.04, KLD: 49.7), similar to the previous chapter's observations for a small sample (n=158) and much larger (n=750k) sample. This result suggests that visualization authors should use Bayesian personalization or other approaches to help users recognize the informativeness of large samples.

5.4.6 Effect of Prior Elicitation

Our results show that conditional on a user specifying their prior, Posterior Visualization and sometimes Uncertainty Analogy better promote Bayesian updating than simply visualizing uncertainty in the observed data. However, given that the status quo in most interactive visualization is not to elicit a prior, one might ask how the act of prior elicitation *itself* impacts updating. Do users become more sensitive to uncertainty in observed data when they explicitly consider their subjective uncertainty about a parameter value?

Comparing an individual's posterior to a normative posterior with and without elicitation is not possible, as without a prior we would have no way of computing the normative posterior. We instead use an aggregate analysis approach similar to that used in prior work on Bayesian cognition [66] and to our approach to computing effect size using CLES. Specifically, we compare participants' log KLD in aggregate between No Elicitation and Elicitation conditions. For each of the four datasets, we constructed a single aggregate prior distribution by finding the median value of α and β respectively across all participants in Elicitation conditions (Fig. 5.10 (1)). Then we calculated a corresponding aggregate normative posterior distribution using the aggregate prior distribution (Fig. 5.10 (2)). We constructed the aggregate posterior distributions from Elicitation-Uncertainty Visualization, Elicitation-Point Estimate, No elicitation-Uncertainty Visualization, and No elicitation-Point Estimate and compared the distribution with the aggregate normative distribution (Fig. 5.10

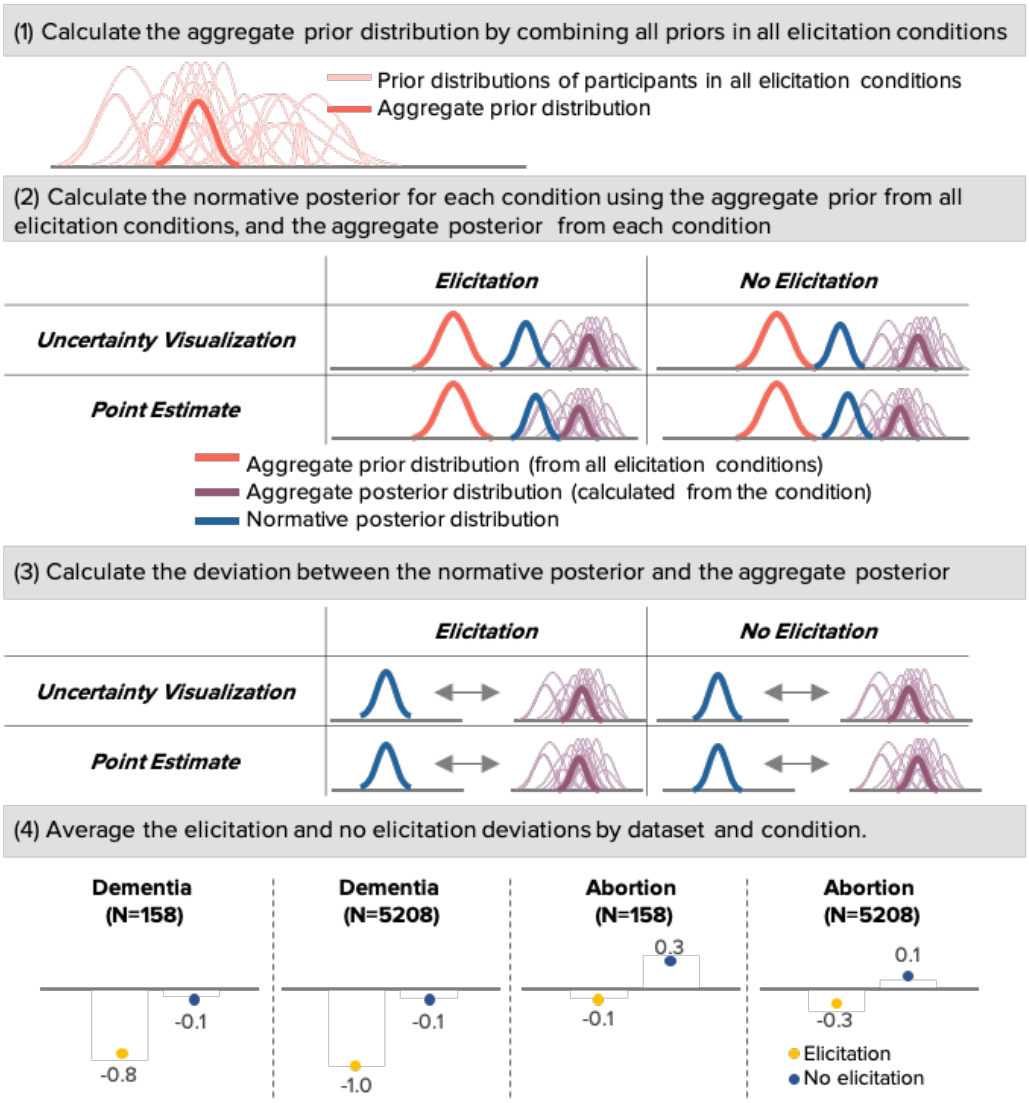


Figure 5.10: Process used to calculate aggregate prior and posterior beliefs and the corresponding normative posterior for each dataset and condition (1-3). Elicitation conditions yielded a lower value compared to No Elicitation condition across all four datasets (4), suggesting that eliciting prior beliefs alone may improve inference.

(3)). To calculate the average effect of eliciting a prior, we pool the Point Estimate and Uncertainty Visualization condition results by dataset for more power.

Across the board, elicitation conditions yielded lower log KLD, suggesting prior elicitation alone may improve updating (Fig. 5.10 (4)).

5.5 Discussion

Evaluating users' interactions with data against rational Bayesian updating paves the way for more sophisticated models of visualization interaction. We reflect on the takeaways from our results and implications of Bayesian evaluation and personalization for visualization.

5.5.1 Bayesian Personalization as Design Strategy

Our results suggest that when visualizations present estimates based on small samples for inference, Bayesian personalization can help untrained users update their beliefs more like Bayesian agents. It can also reduce heterogeneity in updating behavior across a group of users. Our results suggest that any reduction in heterogeneity from Bayesian personalization is likely to be greater in settings where the alternative presentation is a point estimate with sample size, which produced the most dispersion in our results perhaps as a result of users inferring different levels of uncertainty [75].

Bayesian personalization had a small to moderate but reliable effect on reducing bias in updating for small samples, even when data was perceived as moderately likely to have been manipulated, with a randomly drawn user of a Bayesian personalization having an estimated 52% - 59% probability of updating better than someone who only viewed a visualization of uncertainty in the observed data. When Bayesian personalizations were compared to Point Estimates, which are perhaps the most conventional approach to presenting data, Bayesian personalizations were slightly more effective (estimated CLES from 55% to 61%). Using users' prior beliefs as an entry

point into communicating uncertainty via Bayesian personalization may therefore be helpful in common small sample scenarios like presentations of poll results, where people's misinterpretations of uncertainty in data often have implications for their decisions.

The benefits of Bayesian personalization for large sample scenarios are less clear-cut. For the dementia dataset that participants perceived as less likely to be manipulated, visualizing a predicted Bayesian posterior better aligned participants' posterior beliefs on average with Bayesian inference. This effect, similar to the effects of Posterior Visualization that we observed for small datasets, appears to be driven mostly by the Bayesian personalization helping people more accurately update the location of their beliefs. It is important to note that while reliable, the effect of Posterior Visualization for the large dementia dataset may be too small to be of practical significance. For a large dataset, small errors in how a person aggregated the evidence in their prior and the likelihood can have large effects on the deviation between their posterior beliefs and the normative posterior.

If one of two distributions is sharply peaked, which is the case of the observed data of the large sample datasets, KLD will yield a high value even if the two distributions are relatively close in terms of location.

The Analogy condition did *not* improve inference for the large dementia dataset. It is possible that people struggled to use large multipliers to arrive at the normative posterior implied by the analogy, as larger numbers are associated with less precise mental representations and more error in mental calculation [44]. The median multiplier that participants in the Analogy conditions saw for small samples (3) was much smaller than the median of those seen for large samples (72).

For the abortion dataset, which participants rated as slightly more likely to be subject to manipulation, neither Bayesian personalization improved inferences. This may be due to participants discounting the informativeness of the data based on their perceptions that it might have been manipulated. We conducted an ex-

ploratory analyse to explore the strength of this association. We assessed how well participants' Likert ratings (1-5) predicted their deviation from normative Bayesian inference as measured by KLD. We found that for the large abortion dataset, a stronger belief that the data was manipulated predicted a higher KLD, as evidenced by Pearson's correlation coefficient of 0.75. For both the dementia and abortion small datasets, the relationship was still present but not as strong, ranging from 0.45 to 0.57. The large dementia dataset had a strongly negative Pearson's correlation coefficient of -0.94. However, the range of observed KLD for the large dementia dataset was considerably smaller than for the others.

Prior Elicitation as Beneficial

While presenting the observed data using shading to show probability density lowers the dispersion compared to the Point-Estimate condition, visualizing uncertainty has no observable effect on mean error in log KLD (bias).

One possible explanation is that interacting with the prior elicitation interface better prepared participants to reason about uncertainty in the observed data. Our aggregate level analysis suggests that, when priors and observed data are presented using a common, shaded interval representation, prior elicitation alone may improve reasoning about uncertainty. This result, if confirmed in future comparisons of uncertainty comprehension with and without prior elicitation, suggest researchers and authors consider eliciting subjective uncertainty as an alternative or complement to visualizing uncertainty in observed data.

Generalizing Bayesian Personalization

The Bayesian models we employed are quite simple, and the approach we used can also be applied beyond Beta distributions. For example, if a Gaussian (i.e., Normal) distribution is assumed to generate the observed data with two parameters (μ_{data} and σ_{data}), than the mean and the standard deviation of the normative posterior are

prescribed by $\mu_{posterior} = \frac{\mu_{data} \times \sigma^2_{prior} + \mu_{prior} \times \sigma^2_{data}}{\sigma^2_{prior} + \sigma^2_{data}}$, $\sigma_{posterior} = \frac{\sigma^2_{prior} \times \sigma^2_{data}}{\sigma^2_{prior} + \sigma^2_{data}}$ (cf. Eq. 4.1).

Here, the mean of normative posterior is the weighted average between the mean of the prior and the observed data weighted by the amount of information in each distribution (i.e., inverse variance $1/\sigma^2$). The standard deviation of the posterior is the sum of the information from the prior and the observed data (i.e., $1/\sigma^2_{prior} + 1/\sigma^2_{data}$). Bayesian personalization could be applied similarly to our application to Beta distributions, with appropriate changes to the elicitation interface to elicit subjective Gaussian distributions rather than Beta.

We believe that the potential for Bayesian personalization to be used as a design strategy in visualization-based analysis and communication settings extends far beyond the demonstration we presented here. For example, while we use an individual's prior from a single belief update to drive personalization, recent work from economics suggests that how a person updates their beliefs in light of new data is a stable individual trait [10, 11, 47, 110]. Personalizing data representations based on an individual's "update type" (e.g., tendency to overweight vs. underweight their prior or data) may be beneficial in visual analytics or communication settings.

5.5.2 Characterizing Belief Updating Using These Results

Comparing our analysis of location updates to that of variance updates as a whole (Sec. 5.4.3), it is clear that people are much better at providing posteriors beliefs that are located (i.e., have a mean that is) approximately near the location of the normative Bayesian posterior beliefs than they are at providing posterior beliefs that are appropriately certain. We suspect that independent of the sample size of the data they are presented with, people may diverge from Bayesian-like inference whenever they perceive a conflict between their prior beliefs and a presented estimate. For example, the simple Bayesian model would prescribe a posterior mode representing the weighted average of the two distributions' modes and less variance than either distribution. However, we suspect based on our results that many people might

behave approximately Bayesian in updating the location of their beliefs, but remain considerably less certain that the information-pooling Bayesian would due to the perceived conflict between their prior beliefs and the data. Future work might also extend our approach to better detect other qualitatively non-Bayesian behaviors people might exhibit, like deciding to reject either their prior or the observed data when they perceive a conflict between the two.

The characterization of belief updating that emerges from our results also paints a clear picture of how people underupdate given large samples. This result corroborates the previous chapter's finding that people's posterior beliefs deviate sharply from the expectations of Bayesian inference when they are shown large samples in a data interpretation context, and the findings of more abstract behavioral economic studies [9, 15].

5.5.3 Limitations

We measured participants' uncertainty comprehension by quantifying how much their updated beliefs deviated from a normative Bayesian standard. Alternative measures of uncertainty comprehension should be investigated for comparison in future work.

Bayesian personalizations had reliable, but small effects, impacted in part by high variance in how people updated their beliefs. Using a within-subject design could provide better estimates of the relative impact of personalized Bayesian personalizations on individuals.

By explicitly suggesting to a user how they should update their beliefs in light of new data, Bayesian personalization poses interesting questions about when Bayesian inference is the most appropriate normative standard. For example, under what conditions should a user who is distrustful of a data source be guided to integrate the new information into their prior beliefs? While this question is beyond the scope of our work, we believe that there are a number of cases where valid data is rejected

irrationally by users, such as when distrust in the source of a media report (e.g., a Conservative leaning publication) leads a Democrat to reject new information that is in fact trustworthy. Regardless of one's ethical stance, in an era where assuming that people will see data as fact is increasingly unrealistic, a Bayesian framework provides the kind of powerful toolset that can allow researchers to begin asking questions about when, how, and how should people "listen" to new information.

5.6 Summary

In this chapter, I showed how personalizing the presentation of visualized data using Bayesian inference can help laypeople update their beliefs in ways that align more closely with normative Bayesian inference. The study result shows that presenting a personalized Uncertainty Analogy or Posterior Visualization improved belief updating for proportion estimates compared to typical presentations of uncertainty for small datasets, and, in some cases, for large datasets for which people tend to deviate more from normative inference. Visualizing uncertainty in the data via a shaded interval did not show comparable benefits over simply showing a point estimate. Further, an aggregate level analysis of updating suggested that prior elicitation alone may improve Bayesian reasoning.

Chapter 6

SUMMARY

As opposed to the conventional "data only" view of visualization interactions, this dissertation sheds light on ways of considering people's beliefs in understanding visualization interactions. This dissertation contributes new techniques to elicit beliefs, a methodology to investigate visualization interactions in light of beliefs and empirical findings that demonstrate the positive impact of belief-driven approach in visualization interactions.

6.1 Summary of Findings

6.1.1 RQ1: What are the requirements to design belief elicitation techniques to integrate ones' beliefs into visualization interaction?

We design graphical and interactive techniques that elicit users' probabilistic and non-probabilistic beliefs. Inspired by theories in Cognitive and Education Psychology, we develop a technique that prompts a user to graphically articulate the non-probabilistic beliefs, and then shows the data alongside their predictions. We characterize the design space for applying graphical predictions in various chart types.

To elicit probabilistic beliefs, we survey Bayesian cognition literature. Vul et al. demonstrate that people often reach an optimal decision from very few samples [143]. Inspired by the hypothesis, we design a technique that prompts a user to respond with a discrete value multiple times (e.g., 3-5 times) about a parameter (e.g., what is the proportion of elderly who has Alzheimer's in the U.S assisted living centers).

Our exploration of graphical elicitation techniques paves the way for studying people's updating behaviors, especially in a scenario where people are interacting with data. While our work is modeled on a data communication context, the tech-

niques can be used in analysis applications. For example, since Bayesian analysis requires analysts to set a prior distribution, our techniques can be used to elicit prior beliefs in a Bayesian analysis setting and to lower the bar for articulating beliefs in a distribution format.

6.1.2 RQ2: What is the effect of eliciting prior beliefs on how much people process and engage with data?

We evaluate our techniques with several goals in mind. Regarding the techniques introduced in chapter 3 to elicit non-probabilistic beliefs, we demonstrate the efficacy of techniques in recalling the data, a common measure of comprehension in educational psychology. The study shows that externalizing and examining prior beliefs alongside data can improve a viewer's ability to recall visualized data. This finding shows an exciting possibility for visualization practitioners. For example, a visualization designer can simply ask people to graphically sketch their beliefs then show it alongside the data to have some positive impact on engaging them further with data, when the designer expects dataset to be unfamiliar or moderately familiar to their readers.

We evaluate the effect of the different techniques to elicit people's probabilistic beliefs on supporting their belief updating. To summarize the findings, the sample-based graphical technique (with the specific fitting process we used) compared to full-distribution and partial distribution techniques yield slightly lower aggregate deviation, which suggests the sample-based technique may emulate the closest process of how people reason about uncertainty. Since the effect of fitting processes that we used for each technique can be varied, follow up research is needed to further tease apart the effect of different elicitation approaches and the fitting process to make the findings robust.

While the advantage we observed isn't huge, the sample-based technique is worth exploring since it has two nice properties that make it easy to deploy in real-world

visualization interaction. First, a user doesn't need to articulate their responses in a probabilistic manner but directly articulate their beliefs with a few samples with the same unit as data. Second, it's easier than other techniques to scale to multiple parameters.

We also observe that the elicitation act itself can improve people's belief updating. In chapter 4, the analysis shows an improvement in aggregate level belief updating when participants were prompted to articulate their beliefs before seeing the data, where the interface matched with the representation of the data. While the finding manifests consistently across datasets and sample size, the aggregate measure shows the benefit in a less direct manner. Future research has to design more direct measures to observe individual-level benefits. For example, measuring the performance between elicitation conditions and non-elicitation conditions on uncertainty comprehension tasks (e.g., "If you replicate this study 100 times, how likely is it that the observed data will fall between X% and Y%") can be one way to observe the effect of elicitation at an individual level.

6.1.3 RQ3: What models should we use to understand visualization interpretation in light of a user's prior beliefs, data, and their final beliefs?

I present Bayesian inference as a lens to probe people's belief updating process while interacting with visualizations. By analyzing the discrepancy between the user's updated beliefs and the normative Bayesian answer, researchers can gain a more systematic understanding of how a user interprets data. For example, the study shows that individuals' updated beliefs often deviated sharply from the normative Bayesian answer. This tendency is exacerbated when participants examine a large sample dataset. The finding implies that visualization designers need to consider how they can help people recognizing the amount of information a large dataset provides.

Another interesting finding is that the aggregate of all individuals' beliefs was

very close to the normative Bayesian answer when the data contains a small number of samples. This finding aligns with a prominent study in Cognitive Science that shows that people in aggregate appear to be optimal Bayesian when predicting daily quantities (e.g., total baking times) [66].

Using the Bayesian framework, we propose a quantification that captures the user's "perceived uncertainty" of the data by assuming the user's final beliefs were rationally constructed according to the Bayesian standard. This proposed quantification provides insights about how people interpret and perceive the uncertainty and the amount of information in the observed data compared to a simple evaluation metric, such as accuracy in reading out the data. We further demonstrate the validity of the proposed Bayesian framework as an evaluative framework. I believe that the way of quantifying the perceived amount of information opens up a new problem space by providing a new perspective of evaluating visualization interpretation. For example, the quantification enables researchers to identify that participants' perceived sample size is still far from the actual sample size (ref. 4.6.4). I can envision that visualization designers can explore multiple design alternatives to identify which visualization supports the most accurate perception of the underlying uncertainty by applying the proposed quantification.

6.1.4 RQ4: Can we intervene in people's belief updating process to improve users' data comprehension?

We evaluate Bayesian personalizations to promote Bayesian reasoning by communicating uncertainty related to one's prior beliefs. We find that seeing a personalized Uncertainty Analogy or Posterior Visualization help belief updating compared to typical presentations of uncertainty for small datasets, and, in some cases, for large datasets. I can envision that visualization designers can create various interventions to promote better reasoning using people's prior beliefs. For example, the designer can highlight the part where the user has strong beliefs but far from reality, and

provide additional context to persuade the user.

6.2 Future Directions

Through multiple projects, I present how belief-driven approaches can have a positive impact on people's data and uncertainty comprehension and rational belief updating. Future work in belief-driven data interaction will further investigate methods that leverage people's beliefs in designing visualization systems to enable efficient data exploration and decision making under uncertainty. I layout a few directions where I envision the belief-driven approach will be beneficial.

6.2.1 Belief-driven visualization recommender system

Given the sheer amount of data being collected, it is common for analysts to do an exploratory visual analysis of large datasets with tens of different fields or more. This scenario can be supported by recommender systems to reduce the burden of creating visualizations.

When the analyst conducts data analysis, especially exploratory data analysis, the analyst often uses visualization recommender systems that suggest a series of automatically created visualizations to reduce the burden of creating multiple visualizations.

I would like to build a system that incorporates the analyst's beliefs in recommending visualizations that prevent the analyst's bias and make the sense-making process efficient. To contextualize the idea, imagine an analyst who wants to explore a dataset about demographics of a movie award nominee, which contains more than 20 columns including gender, ethnicity, birthplace, birth date, and when they were nominated (Fig 6.1).

If the analyst would like to fully understand the entire patterns that might exist in a dataset, they may need to go through more than thousands of visualizations. The state of the art recommender systems would help the analyst save the cognitive

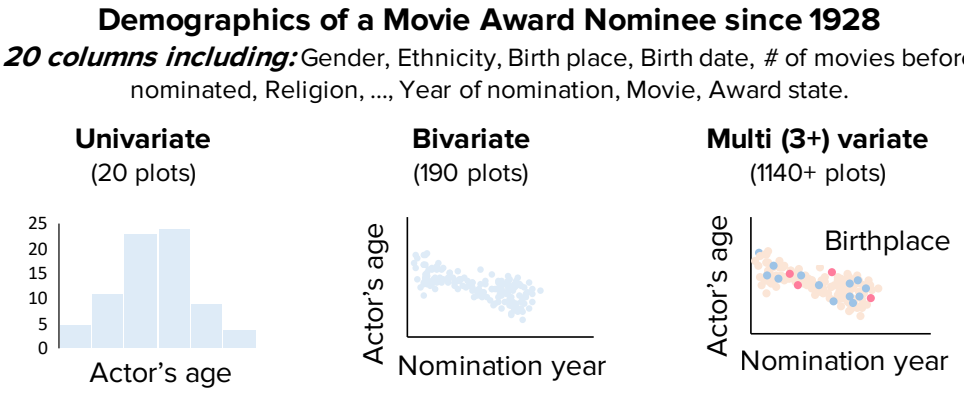


Figure 6.1: People often start the exploration by looking at a univariate chart that describes single variable [91]. Then move on to examine more complex one as the analysis progresses to understand more complex relationship between the variables.

cost by providing recommendations that are closely related to what the analyst is looking at the moment [150], instead of suggesting random charts. Another existing system carefully curates the order of the visualizations to minimize the burden of reading them consequently [92].

One remaining challenge is that the analyst still has to go through multiple charts. None of the existing approaches directly incorporate analyst beliefs to make the process more efficient, and lack considerations of preventing human bias such as confirmation bias that the analyst may confirm a spurious pattern as the real pattern because that was what the analyst believed.

I can envision a visualization recommendation system that takes into account the analyst's beliefs to make the process more efficient. I can build a feature that focuses on showing patterns that the analyst may not be aware, or patterns against the analyst's expectation to maximize the information gain while they are interacting with data. An exciting research question will be *when* to elicit the analyst's beliefs on *which parameter* to minimize the interruption of the user's flow but maximize the information gain from the user's input.

Another useful application will be detecting bias and supporting the analyst to avoid them by explicitly eliciting their beliefs. For example, I can elicit beliefs on a few parameters first, and infer new beliefs on other parameters from the pre-elicited beliefs. When the analyst is interacting with a visualization that the system detects to be similar to the analyst beliefs, the system can alert them to carefully examine the visualization to avoid the chance to introduce confirmation bias.

6.2.2 Supporting human-machine decision-making process with belief elicitation and visualizations

The performance of a human-AI team is influenced by how well the human's expectation of the AI aligns with its actual capabilities [13]. However, supporting the process for humans to construct an accurate mental model remains unexplored. The Bayesian framework applied to visualization interpretation offers a promising way to support this process by surfacing the necessary components to perform the alignment operation.

Imagine a user examines a medical image with a model's prediction of how different regions provide evidence for the diagnosis. The user, likely to be a doctor who has their own beliefs, will try to combine the model's prediction and their own expertise to make the final decision. I envision a system that prompts the user to articulate their beliefs, combines the beliefs with the model's prediction using Bayesian models, and visualizes the final beliefs with a depiction of how it is derived. By making the belief updating process *explicit* and *visual*, the user may opt for the better rational choice as the final decision. The system can also help the user reflect on the model's capabilities as well as their own by examining their previous interactions. In addition to interaction design, the utility of this approach also depends on *representations* of uncertainty that both humans and models can understand. Future research will investigate how to represent human beliefs in formats that can be incorporated into an AI model and vice-versa.

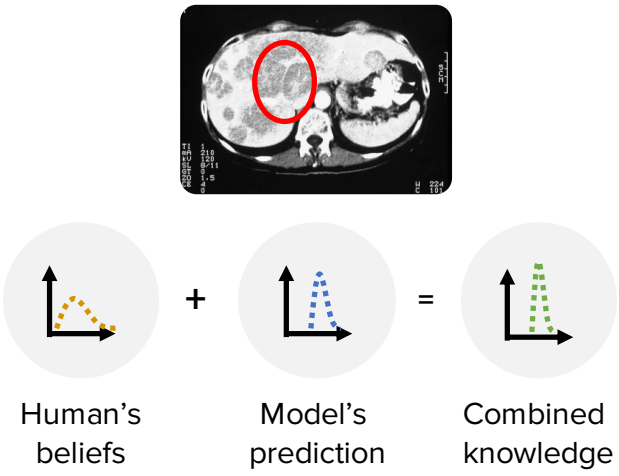


Figure 6.2: An example scenario where an expert can make a professional decision based on the model's prediction and their beliefs as an expert. The interface can explicitly visualize and describe how their expertise and the new evidence (i.e., model's prediction) can be merged rationally to inform the user to make more logical choice.

6.3 Closing Remarks

Visualizations have become essential media for navigating the world as more data is generated to approximate phenomena around us. In order to design visualizations that are maximally informative, it is critical to focus not only on how well a person will decode the information, but also on how it will influence the person's beliefs and further inform their decision. A belief-driven approach, such as I have demonstrated, can provide a realistic perspective on understanding visualization interactions and designing visualization systems.

BIBLIOGRAPHY

[1] Osmi mental health in tech survey 2016. https://www.kaggle.com/osmi/mental-2016%2011%2014ealth-in-tech-2016, 2016. Accessed: 2018-05-01.

[2] Tversky A. and Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science*, 34(2):1124—1131, 1974.

[3] Diederik Aerts. Quantum structure in cognition. *Journal of Mathematical Psychology*, 53(5):314–348, 2009.

[4] Shaaron Ainsworth and Andrea Th Loizou. The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, 27(4):669–681, 2003.

[5] Gregor Aisch, Amanda Cox, and Kevin Quealy. You draw it: How family income predicts children's college chances. *The New York Times*, May 28, 2015, *http://nyti.ms/1ezbuWY*, 2015.

[6] Chris Alcantara and Chiqui Esteban. 2016 election exit polls. *The Washington Post*, Nov. 29, 2016, https://www.washingtonpost.com/graphics/politics/2016-election/exit-polls/„ 2016.

[7] Vincent Aleven and Kenneth R Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science*, 26(2):147–179, 2002.

[8] Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education*, pages 246–255. Citeseer, 2001.

[9] Sandro Ambuehl and Shengwu Li. Belief updating and the demand for information. *Games and Economic Behavior*, 109:21–39, 2018.

[10] Pavel Atanasov, Jens Witkowski, Lyle Ungar, Barbara Mellers, and Philip Tetlock. Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, 160:19–35, 2020.

[11] Ned Augenblick and Matthew Rabin. Belief movement, uncertainty reduction, and rational updating. *UC Berkeley-Haas and Harvard University Mimeo*, 2018.

[12] Victoria Balara. Fox news poll: Voters split on abortion, but majority wants roe v. wade to endure, 2019.

[13] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[14] Lucy Bastin, Peter F Fisher, and Jo Wood. Visualizing uncertainty in multi-spectral remotely sensed imagery. *Computers & Geosciences*, 28(3):337–350, April 2002.

[15] Daniel J 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.

[16] Katerine Bielaczyc, Peter L Pirolli, and Ann L Brown. Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and instruction*, 13(2):221–252, 1995.

[17] Matthew Bloch and Hannah Fairfield. For the elderly, diseases that overlap. *The New York Times*, Apr 15, 2013, 2013.

[18] Elizabeth Bonawitz, Stephanie Denison, Alison Gopnik, and Thomas L Griffiths. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, 74:35–65, 2014.

[19] Elizabeth Bonawitz, Stephanie Denison, Thomas L Griffiths, and Alison Gopnik. Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in cognitive sciences*, 18(10):497–500, 2014.

[20] David Borland and Russell M Taylor Ii. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[21] David Bourgin, Joshua Abbott, Tom Griffiths, Kevin Smith, and Ed Vul. Empirical evidence for markov chain monte carlo in memory search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[22] Jerome R Busemeyer and Peter D Bruza. *Quantum models of cognition and decision*. Cambridge University Press, 2012.

[23] Matt Canham and Mary Hegarty. Effects of knowledge and display design on comprehension of complex graphics. *Learning and instruction*, 20(2):155–166, 2010.

[24] Stuart K Card and Jock Mackinlay. The structure of the information visualization design space. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pages 92–99. IEEE, 1997.

[25] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[26] Patricia A Carpenter and Priti Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75, 1998.

[27] Yang Chen, Scott Barlowe, and Jing Yang. Click2annotate: Automated insight externalization with rich semantics. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 155–162. IEEE, 2010.

[28] Andrea Cheshire, Linden J Ball, and CN Lewis. Self-explanation, feedback and the development of analogical reasoning skills: Microgenetic evidence for a metacognitive processing account. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, ed. BG Bara, L. Barsalou & M. Bucciarelli*, pages 435–41. Citeseer, 2005.

[29] Michelene TH Chi. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology*, 5:161–238, 2000.

[30] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182, 1989.

[31] Michelene TH Chi, Nicholas Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

[32] 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, September 1984.

[33] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[34] Patricia Cohen. Hiring slowed in september as unemployment rate fell to a 50-year low. *The New York Times*, Nov 1, 2019, *https://nyti.ms/2AFPCq7*„ 2019.

[35] William G Cole and Janet E Davidson. Graphic representation can lead to fast and accurate bayesian reasoning. In *Proceedings. Symposium on Computer Applications in Medical Care*, pages 227–231. American Medical Informatics Association, 1989.

[36] William G Cole and Janet E Davidson. Graphic representation can lead to fast and accurate bayesian reasoning. In *Proceedings. Symposium on Computer Applications in Medical Care*, volume 20, pages 227–231. American Medical Informatics Association, 1989.

[37] Cristina Conati and Kurt Vanlehn. Toward computer-based support of metacognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education (IJAIED)*, 11:389–415, 2000.

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

[39] Michael Correll, Dominik Moritz, and Jeffrey Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 642. ACM, 2018.

[40] Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

[41] Nelson Cowan. *Attention and memory: An integrated framework*. Oxford University Press, 1998.

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

[43] Trevor J Davis and C Peter Keller. Modelling and visualizing multiple spatial uncertainties. *Computers & Geosciences*, 23(4):397–408, 1997.

[44] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. OUP USA, 2011.

[45] Adeline Delavande. Measuring revisions to subjective expectations. *Journal of Risk and Uncertainty*, 36(1):43–82, 2008.

[46] Nathan F Dieckmann, Robin Gregory, Ellen Peters, and Robert Hartman. Seeing what you want to see: How imprecise uncertainty ranges enhance motivated reasoning. *Risk analysis*, 37(3):471–486, 2017.

[47] Jeff Dominitz and Charles F Manski. Measuring and interpreting expectations of equity returns. *Journal of Applied Econometrics*, 26(3):352–370, 2011.

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

[49] Ruth B Ekstrom, John W French, Harry H Harman, and Diran Dermen. Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational testing service*, 1976.

[50] Paolo Federico, Albert Amor-Amorós, and Silvia Miksch. A nested workflow model for visual analytics design and validation. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 104–111, 2016.

[51] Paolo Federico, Markus Wagner, Alexander Rind, Albert Amor-Amorós, Silvia Miksch, and Wolfgang Aigner. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. 2017.

[52] Monica GM Ferguson-Hessler and Ton de Jong. Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction*, 7(1):41–54, 1990.

[53] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 144. ACM, 2018.

[54] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 144. ACM, 2018.

[55] Baruch Fischhoff, Daniel Kahneman, Paul Slovic, and Amos Tversky. For those condemned to study the past: Heuristics and biases in hindsight. *Foundations of cognitive psychology: Core readings*, pages 621–636, 2002.

[56] Bennett L Fox. A bayesian approach to reliability assessment. 1966.

[57] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

[58] Rocio Garcia-Retamero and Ulrich Hoffrage. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, 83:27–33, 2013.

[59] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

[60] Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

[61] Daniel G Goldstein and Gerd Gigerenzer. Fast and frugal forecasting. *International Journal of Forecasting*, 25(4):760–772, 2009.

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

[63] Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. Investigating representation alternatives for communicating uncertainty to non-experts. In *IFIP Conference on Human-Computer Interaction*, pages 256–263. Springer, 2015.

[64] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

[65] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229, 2015.

[66] Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

[67] Carolina E Hagberg, Annika Mehlem, Annelie Falkevall, Lars Muhl, Barbara C Fam, Henrik Ortsäter, Pierre Scotney, Daniel Nyqvist, Erik Samén, Li Lu, et al. Targeting vegf-b as a novel treatment for insulin resistance and type 2 diabetes. *Nature*, 490(7420):426–430, 2012.

[68] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

[69] Mary Hegarty. Diagrams in the mind and in the world: Relations between internal and external visualizations. In *Diagrammatic representation and inference*, volume 8, pages 1–13. Springer, 2004.

[70] Mary Hegarty and Sarah Kriz. Effects of knowledge and spatial ability on learning from animation. *Learning with animation: Research implications for design*, pages 3–29, 2008.

[71] Mary Hegarty, Sarah Kriz, and Christina Cate. The roles of mental animations and external animations in understanding mechanical systems. *Cognition and instruction*, 21(4):209–249, 2003.

[72] Mary Hegarty and Kathryn Steinhoff. Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning and Individual differences*, 9(1):19–42, 1997.

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

[74] Kathleen L Hourihan and Aaron S Benjamin. Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(4):1068, 2010.

[75] Jessica Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, January 2019.

[76] Jessica Hullman, Eytan Adar, and Priti Shah. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 17, pages 1461–1470. ACM, IEEE, 2011.

[77] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. Reproduced as Figure 6, 2018.

[78] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. Improving comprehension of measurements using concrete re-expression strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 34. ACM, 2018.

[79] 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.

[80] Edwin Hutchins. *Cognition in the Wild*. MIT press, 1995.

[81] Sirkka L Jarvenpaa. Graphic displays in decision making the visual salience effect. *Journal of Behavioral Decision Making*, 3(4):247–262, 1990.

[82] Wilson Andrews Josh KatZ and Jeremy Bowers. Elections 2014: Make your own senate forecast. *The New York Times*, Sep 2, 2014, *http://nyti.ms/1plfIyv*„ 2014.

[83] Christopher Kaeser. A day in the life. *The Wall Street Journal*, June 24, 2015, 2015.

[84] Daniel Kahneman and Patrick Egan. *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York, 2011.

[85] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics*, 2018.

[86] 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 CHI Conference on Human Factors in Computing Systems*, pages 5092–5103. ACM, 2016.

[87] Daniel Kersten and Alan Yuille. Bayesian models of object perception. *Current opinion in neurobiology*, 13(2):150–158, 2003.

[88] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 38–48. ACM, 2016.

[89] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1375–1386. ACM, 2017.

[90] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation. *IEEE transactions on visualization and computer graphics*, 24(1):760–769, 2018.

[91] Yea-Seul Kim, Nathalie Henry Riche, Bongshin Lee, Matthew Brehmer, Michel Pahud, Ken Hinckley, and Jessica Hullman. Inking your insights: Investigating digital externalization behaviors during data analysis. In *Proceedings of the 2019 ACM Interactive Surface and Spaces*. ACM, 2019.

[92] Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, and Jeffrey Heer. Graphscape: A model for automated reasoning about visualization similarity and sequencing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2628–2638. ACM, 2017.

[93] David Kirsh. Thinking with external representations. *AI & Society*, 25(4):441–454, 2010.

[94] Stephen M Kosslyn. Understanding charts and graphs. *Applied cognitive psychology*, 3(3):185–225, 1989.

[95] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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

[97] Bongshin Lee, Greg Smith, Nathalie Henry Riche, Amy Karlson, and Sheelagh Carpendale. Sketchinsight: Natural data exploration on interactive whiteboards leveraging pen and touch interaction. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis)*, pages 199–206. IEEE, 2015.

[98] Stephan Lewandowsky, Thomas L Griffiths, and Michael L Kalish. The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive science*, 33(6):969–998, 2009.

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

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

[101] Charles F Manski. Survey measurement of probabilistic macroeconomic expectations: progress and promise. *NBER Macroeconomics Annual*, 32(1):411–471, 2018.

[102] Bill Marsh. Are we in the midst of a sixth mass extinction? *The New York Times*, June 1, 2012, 2012.

[103] Richard E Mayer. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 43, 2014.

[104] Nina McCurdy, Julie Gerdes, and Miriah Meyer. A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics*, 25(1):925–935, 2018.

[105] RICHARD MCELREATH. *RETHINKING AN R PACKAGE FOR FITTING AND MANIPULATING BAYESIAN MODELS VERSION 1.56*, 2016.

[106] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2018.

[107] Kenneth O McGraw and SP Wong. A common language effect size statistic. *Psychological bulletin*, 111(2):361, 1992.

[108] Danielle S McNamara, Tenaha O'Reilly, Michael Rowe, Chutima Boonthum, and IB Levinstein. istart: A web-based tutor that teaches self-explanation and metacognitive reading strategies. *Reading comprehension strategies: Theories, interventions, and technologies*, pages 397–421, 2007.

[109] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2536–2545, 2012.

[110] Markus M Moebius, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. Managing self-confidence: Theory and experimental evidence. Technical report, National Bureau of Economic Research, 2011.

[111] Hedwig M Natter and Dianne C Berry. Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology*, 19(1):123–135, 2005.

[112] 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.

[113] Alvitta Ottley, Blossom Metevier, PK Han, and Remco Chang. Visually communicating bayesian statistics to laypersons. In *Technical Report*. Tufts University, 2012.

[114] Alvitta Ottley, Evan M Peck, Lane T Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A Taylor, Paul KJ 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, 2015.

[115] 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.

[116] Steven Pinker. A theory of graph comprehension. *Artificial intelligence and the future of testing*, pages 73–126, 1990.

[117] Peter Pirolli and Margaret Recker. Learning strategies and transfer in the domain of programming. *Cognition and instruction*, 12(3):235–275, 1994.

[118] Kristin Potter, Mike Kirby, Dongbin Xiu, and Chris R Johnson. Interactive visualization of probability and cumulative density functions. *International journal for uncertainty quantification*, 2(4):226–249, 2012.

[119] Drazen Prelec. Filtering survey respondents with bayesian truth serum: Application to the 2018 us house elections. Presentation at Kellogg Business School Northwestern University, April 2019.

[120] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532, 2017.

[121] Alexander Renkl. Learning from worked-out examples: A study on individual differences. *Cognitive science*, 21(1):1–29, 1997.

[122] Alexander Renkl, Robin Stark, Hans Gruber, and Heinz Mandl. Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary educational psychology*, 23(1):90–108, 1998.

[123] Marguerite Roy and Michelene TH Chi. The self-explanation principle in multimedia learning. *The Cambridge handbook of multimedia learning*, pages 271–286, 2005.

[124] R Ryan. Self-explanation and adaptation. *Psychology*, 1996.

[125] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2015.

[126] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014.

[127] Priti Shah, Richard E Mayer, and Mary Hegarty. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, 91(4):690, 1999.

[128] Nate Silver. Final election update: There's a wide range of outcomes, and most of them come up clinton. *The New York Times*, Nov 8, 2016, *https://fivethirtyeight.com/features/final-election-update-theres-a-wide-range-of-outcomes-and-most-of-them-come-up-clinton/*, 2016.

[129] Keith Stenning and Jon Oberlander. A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive science*, 19(1):97–140, 1995.

[130] Elsbeth Stern, Carmela Aprea, and Hermann G Ebner. Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction*, 13(2):191–203, 2003.

[131] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. Inferring causal networks from observations and interventions. *Cognitive science*, 27(3):453–489, 2003.

[132] Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[133] Joshua B. Tenenbaum Thomas L. Griffiths and Charles Kemp. Bayesian inference. In Keith J. Holyoak and Robert G. Morrison, editors, *The Oxford handbook of thinking and reasoning*. Oxford University Press, Oxford, 2012.

[134] J Gregory Trafton, Susan B Trickett, and Farilee E Mintz. Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science*, 10(1):89–106, 2005.

[135] Jennifer Trueblood and Emmanuel Pothos. A quantum probability approach to human causal reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[136] Jennifer Tsai, Sarah Miller, and Alex Kirlik. Interactive visualizations to improve bayesian reasoning. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 55, pages 385–389. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[137] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

[138] Jarke J Van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86. IEEE, 2005.

[139] Jarke J Van Wijk. The value of visualization. In *Visualization, 2005. VIS 05. IEEE*, pages 79–86. IEEE, 2005.

[140] Fernanda B Viegas and Martin Wattenberg. Communication-minded visualization: A call to action. *IBM Systems Journal*, 45(4):801, 2006.

[141] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[142] Edward Vul and Harold Pashler. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647, 2008.

[143] Griffiths T. Vul E., Goodman N. and Tenenbaum J. One and done? optimal decisions from very few samples. *Cognitive science*, 38(2):1124—1131, 2014.

[144] Xiaoyu Wang, Dong Hyun Jeong, Wenwen Dou, Seok-won Lee, William Ribarsky, and Remco Chang. Defining and applying knowledge conversion processes to a visual analytics system. *Computers & Graphics*, 33(5):616–623, 2009.

[145] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[146] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010.

[147] Martin Wattenberg and Jesse Kriss. Designing for social data analysis. *IEEE transactions on visualization and computer graphics*, 12(4):549–557, 2006.

[148] 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.

[149] Craig M Wittenbrink, Alex T Pang, and Suresh K Lodha. Glyphs for visualizing uncertainty in vector fields. *IEEE transactions on Visualization and Computer Graphics*, 2(3):266–279, 1996.

[150] 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 *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2648–2659. ACM, 2017.

[151] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. The sandbox for analysis: concepts and methods. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 801–810. ACM, 2006.

[152] Yifan Wu, Larry Xu, Remco Chang, and Eugene Wu. Towards a bayesian model of data visualization cognition, 2017.

[153] Yujun Wu, Weichung J Shih, and Dirk F Moore. Elicitation of a beta prior for bayesian inference in clinical trials. *Biometrical Journal*, 50(2):212–223, 2008.

[154] Jeff Zacks and Barbara Tversky. Bars and lines: A study of graphic communication. *Memory and Cognition*, 27(6):1073–1079, 1999.