Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability

Alvitta Ottley, Evan M. Peck, Lane T. Harrison,
Daniel Afergan, Caroline Ziemkiewicz, Holly A. Taylor, Paul K. J. Han and Remco Chang

Abstract— Decades of research have repeatedly shown that people perform poorly at estimating and understanding conditional probabilities that are inherent in Bayesian reasoning problems. Yet in the medical domain, both physicians and patients make daily, life-critical judgments based on conditional probability. Although there have been a number of attempts to develop more effective ways to facilitate Bayesian reasoning, reports of these findings tend to be inconsistent and sometimes even contradictory. For instance, the reported accuracies for individuals being able to correctly estimate conditional probability range from 6% to 62%. In this work, we show that problem representation can significantly affect accuracies. By controlling the amount of information presented to the user, we demonstrate how text and visualization designs can increase overall accuracies to as high as 77%. Additionally, we found that for users with high spatial ability, our designs can further improve their accuracies to as high as 100%. By and large, our findings provide explanations for the inconsistent reports on accuracy in Bayesian reasoning tasks and show a significant improvement over existing methods. We believe that these findings can have immediate impact on risk communication in health-related fields.

Index Terms—Bayesian Reasoning, Visualization, Spatial Ability, Individual Differences.

1 Introduction

As the medical field transitions toward evidence-based and shared decision making, effectively communicating conditional probabilities to patients has emerged as a common challenge. To make informed health decisions, it is essential that patients understand health risk information involving conditional probabilities and Bayesian reasoning [15]. However, understanding such conditional probabilities is challenging for patients [11]. Even more alarming, the burden of communicating complex statistical information to patients is often placed on physicians even though studies have shown that most struggle with accurate estimations themselves [11].

Still, both physicians and patients make life-critical judgments based on conditional probabilities. Deficits in diagnostic test sensitivity and specificity (intrinsic characteristics of the test itself) can lead to false negative and false positive test results which do not reflect the actual state of an individual. For low-prevalence diseases, even a highly specific test leads to false positive results for a majority of test recipients. Unless a patient fully understands the uncertainties of medical tests, news of a negative result can lead to false reassurance that treatment is not necessary, and news of a positive result can bring unjust emotional distress.

Consider the following mammography problem [17]:
"The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography.
A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?"

Misinterpretation of this and other medical test statistics can have serious adverse consequences such as overdiagnosis [30, 42, 43] or even death. However, there are currently no effective tools for mitigating this problem. Despite decades of research, the optimal methods for improving interpretation of diagnostic test results remain elusive, and the available evidence is sparse and conflicting.

Prior work indicates that visualizations may be key for improving performance with Bayesian reasoning problems. For example, results from Brase [1] and work by Garcia-Retamero and Hoffrage [16] suggest that visual aids such as Euler diagrams and icon arrays hold promise. Researchers have also explored visualizations such as decision trees [13, 28], contingency tables [7], "beam cut" diagrams [17] and probability curves [7], and have shown improvements over text-only representations. However, when researchers in the visualization community extended this work to a more diverse sampling of the general population, they found that adding visualizations to existing text representations did not significantly increase accuracy [29, 32].

Given the contradictory findings of prior research, we aim to identify factors that influence performance on Bayesian reasoning tasks. We hypothesize that these discrepancies are due to differences in problem representations (textual or visual), as well as the end users' innate ability to reason through these problems when using visualizations. In particular, we propose that the phrasing of text-only representations can significantly impact comprehension and that this effect is further confounded when text and visualization are incorporated into a single representation. Furthermore, motivated by prior work [5, 24, 40, 41, 45], we also hypothesize that individual differences (i.e., spatial ability) are mediating factors for performance on Bayesian reasoning tasks.

To test our hypotheses, we conducted two experiments to investigate how problem representation and individual differences influence performance on Bayesian reasoning tasks. The first experiment focused on text-only representations and how phrasing can impact accuracy, while the second explores how individual differences and representations that combine text and visualization affect performance.

With Experiment 1 we show that wording can significantly affect users' accuracy and demonstrate how probing[1] can help evaluate different representations of Bayesian reasoning problems. Combining techniques that have previously been tested independently, our results

- *Alvitta Ottley is with Tufts University. E-mail: alvittao@cs.tufts.edu.*
- *Evan Peck is with Bucknell University. E-mail: evan.peck@bucknell.edu.*
- *Lane T. Harrison is with Tufts University. E-mail: lane@cs.tufts.edu.*
- *Daniel Afergan is with Tufts University. E-mail: afergan@cs.tufts.edu.*
- *Caroline Ziemkiewicz is with Tufts University and Aptima Inc. E-mail: cziemkiewicz@aptima.com.*
- *Holly A. Taylor is with Tufts University. E-mail: holly.taylor@tufts.edu.*
- *Paul K. J. Han is with Maine Medical Center and Tufts Medical School. E-mail: hanp@mmc.org.*
- *Remco Chang is with Tufts University. E-mail: remco@cs.tufts.edu.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

[1] Instead of asking a single question (typically the true positive rate), you ask a series of questions designed to guide the user through their calculations.

show an increase in the accuracy of the mammography problem from the previously reported 6% [29] to 42%. Our findings demonstrate how the phrasing of a Bayesian problem can partially explain the poor or inconsistent results of prior work and provide a baseline text-only representation for future work.

In Experiment 2, we tested six different representations including a new text-only representation that uses indentation to visually depict set relations (*Structured-Text*), a storyboarding visualization that progressively integrates textual information with a frequency grid visualization (*Storyboarding*), and a visualization-only representation (*Vis-Only*). The results of our second experiment show that altering the amount of information shown in text and visualization designs can yield accuracies as high as 77%. However, we found that adding visualizations to text resulted in no measurable improvement in performance, which is consistent with prior work in the visualization community by Micallef et al. [29].

Examining our study population further, we found that spatial ability impacts users' speed and accuracy on Bayesian reasoning tasks, with high spatial ability users responding significantly faster and more accurately than low spatial ability users. Analyzing accuracy with respect to spatial ability, we discovered that users with high spatial ability tend to perform better than users with low spatial ability across all designs, achieving accuracies from 66%-100%. We discuss the implications of these findings for text and visualization design, and how these methods may impact the communication of conditional probability in the medical field and beyond.

We make the following contributions to the understanding of how phrasing, visualizations and individual differences influence Bayesian reasoning:

- We identify key factors that influence performance on Bayesian reasoning tasks and explain the inconsistent and conflicting findings of prior work.
- We show that the phrasing of textual Bayesian reasoning problems can significantly affect comprehension, and provide a benchmark text-only problem representation that allows future researchers to reliably test different designs of Bayesian reasoning problems.
- We demonstrate that a user's spatial ability impacts their ability to solve Bayesian reasoning problems with different visual and textual representations. Our findings provide guidance on how to design representations for users of varying spatial ability.

2 Related Work

There is a substantial body of work aimed at developing novel, more effective methods of communicating Bayesian statistics. Still, there is no authoritative method for effectively communicating Bayesian reasoning, and prior results are inconsistent at best. Below we survey some of these findings.

Gigerenzer and Hoffrage [17] in their seminal work explored how text-only representations can be improved using natural frequency formats. They explored the use of phrases such as *96 out of 1000* instead of *9.6%*, hypothesizing that natural frequency formats have greater perceptual correspondence to natural sampling strategies [17]. Their findings demonstrate that using natural frequency significantly improves users' understanding of Bayesian reasoning problems.

A series of studies has also been conducted to investigate the efficacy of using visualizations to aid reasoning. Various types of visualizations have been tested, including Euler diagrams [1, 25, 29], frequency grids or icon arrays [16, 25, 29, 32, 35], decision trees [13, 35, 37], "beam cut" diagrams [17], probability curves [7], contingency tables [7, 8] and interactive designs [38]. While some researchers have compared several visualization designs [1, 29, 32], many of these visualizations were proposed and tested separately. It is still not clear which best facilitates Bayesian reasoning.

For instance, recent work by Garcia-Retamero and Hoffrage [16] investigated how different representations (text versus visualization) affect the communication of Bayesian reasoning problems to both doctor and patients. They conducted an experiment where half of the participants received natural frequency formats and the other half received percentages. A further division was made within these groups; half of the participants received the information in numbers while the other half were presented with a visualization (a frequency grid). Their results confirmed the prior results of Gigerenzer and Hoffrage [17] showing that users are more accurate when information is presented using natural frequency formats. With their visualization condition, and they were able to achieve overall accuracies of 62%, one of the highest reported accuracies.

Work by Brase [1] compared various visualizations for communicating Bayesian reasoning. In a comparative study, he analyzed participants' accuracies when three different visualizations (icon arrays, Euler diagrams and discretized Euler diagrams [2]) were added to textual information. Like natural frequency formats, discrete items represented by the icon array were expected to correspond with humans' perception of natural sampling, thus improving Bayesian reasoning. In contrast, Euler diagrams were expected to enhance the perception of the nested-set relations that are inherent in Bayesian reasoning problems. The discretized Euler diagram was designed as a hybrid of the two.

Brase found that icon arrays had the best overall accuracy rate (48%), suggesting that they best facilitate Bayesian reasoning. However there were some inconsistencies with the visualization designs used by Brase [29]. For instance, the Euler diagram was not area proportional but the hybrid diagram was, and the number of glyphs in the hybrid diagram differed from the number of glyphs in the frequency grid [29]. Noticing this, researchers in the visualization community extended this work by designing a new, consistent set of visualizations and surveying a more diverse study population.

Micallef et al. [29] used a combination of natural frequency formats, icon arrays and Euler diagrams to improve the designs of Brase [1]. Instead of surveying university undergraduates, they recruited crowdsourced participants via Amazon's Mechanical Turk in an effort to simulate a more diverse lay population [29]. Their study compared 6 different visualization designs and found no significant performance differences among them. Reported accuracies for the 6 designs ranged from 7% to 21% and the control condition (a text-only representation with natural frequency formats) yielded an overall accuracy of only 6%.

They also found no statistically significant performance difference between the control text-only condition and any of the conditions with visualization designs. This finding implies that simply adding visual aids to existing textual representation did not help under the studied conditions. Their follow up work adds yet another dimension. In a second experiment, they reported significant improvements in the accuracies for their visualization conditions when numerical values for the text descriptions were removed. This finding suggests that presenting both text with numerical values and visualization together may overload the user and result in incorrect inferences.

The findings of Micallef et al. [29] suggest a possible interaction between textual information and visualization when representing Bayesian reasoning problems. One possible explanation for this interaction is that both the mental model required to interpret the textual information in a Bayesian reasoning problem and the mental model required to understand a visualization can compete for the same resources [24]. As more information is presented, a user's performance can degrade since more items will be held in the user's spatial working memory [21]. In addition to explaining the inconsistencies among prior work by exploring different wording and visualization representations, this paper aims to understand how spatial ability mediates performance on Bayesian reasoning problems.

2.1 Spatial Ability

In recent years, an overwhelming body of research has demonstrated how individual differences impact people's ability to use information

[2] These are Euler diagrams with discrete items. They were designed as hybrid diagrams that combine both the natural sampling affordance of icon arrays and the nested-set relations affordance of traditional Euler diagrams.

visualization and visualization systems [2, 5, 6, 18, 31, 33, 40, 41, 47], and a growing number of researchers have advocated for better understanding of these effects [44, 46]. One of the main factors that have been shown to influence visualization use is *spatial ability*.

Spatial ability in general refers to the ability to mentally represent and manipulate two- or three-dimensional representations of objects. Spatial ability is a cognitive ability with a number of measurable dimensions, including spatial orientation, spatial visualization, spatial location memory, targeting, disembedding and spatial perception [26, 40]. People with higher spatial ability can produce more accurate representations and maintain a reliable model of objects as they move and rotate in space.

There is considerable evidence that these abilities affect how well a person can reason with abstract representations of information, including visualizations. Vicente et al. [41] found that low spatial ability corresponded with poor performance on information retrieval tasks in hierarchical file structures. They found that in general high spatial ability users were two times faster than low spatial ability users and that low spatial ability users were more likely to get lost in the hierarchical file structures.

Chen and Czerwinski [5] found that participants with higher spatial ability employed more efficient visual search strategies and were better able to remember visual structures in an interactive node-link visualization. Velez et al. [40] tested users of a three-dimensional visualization and discovered that speed and accuracy were dependent on several factors of spatial ability. Similarly, Cohen and Hegarty [6] found that users' spatial abilities affects the degree to which interacting with an animated visualization helps when performing a mental rotation task, and that participants with high spatial ability were better able to use a visual representation rather than rely on an internal visualization.

This body of research shows that users with higher spatial ability are frequently more effective at using a variety of visualizations. Taken together, they suggest that high spatial ability often correlates with better performance on tasks that involve either searching through spatially arranged information or making sense of new visual representations. Additionally, there is evidence that high spatial ability makes it easier to switch between different representations of complex information. Ziemkiewicz and Kosara [45] tested users' ability to perform search tasks with hieraracby visualizations when the spatial metaphor implied in the task questions differed from that used by the visualization. Most participants performed poorly when the metaphors conflicted, but those with high spatial ability did not. This confirms findings that spatial ability plays a role in understanding text descriptions of spatial information [10].

In Bayesian reasoning domain, Kellen [24] found that spatial ability was relevant to the understanding of visualizations of Bayesian reasoning. He used Euler diagrams and investigated how problem complexity (the number of relationships presented in an Euler diagram) impacts users' performance. His findings suggest that spatial ability may moderate the effect of visualizations on understanding. However, his work only investigated spatial ability as it relates to Bayesian reasoning and the number of relationships depicted in an Euler diagram.

Still, like the prior reported accuracy findings, the reported results on the effects of spatial ability on understanding Bayesian reasoning have been contradictory. Micallef et al. [29] too investigated the effects of spatial ability. They compared six text and visualization conditions and one text-only condition but found no significant effect of participants' spatial abilities.

3 Research Goals

The body of existing work presented in this paper paints a complex portrait of visualization and Bayesian reasoning. First, the results of the prior works are inconsistent. The reported accuracies of the baseline text-only conditions differed significantly: Brase [1] reported 35.4%, Garcia-Retamero and Hoffrage [16] reported 26% while Micallef et al. [29] reported accuracies of only 6%. Second, prior work suggests an interaction between textual information and visualization when they are combined into a single representation.

In order to progress this important area of research, we must first identify factors that affect a user's ability to extract information from text and visualization representations of Bayesian reasoning problems. Thus, the primary research goal for this work is to disambiguate the discrepancies among prior works' results. We hypothesize that the observed discrepancies among prior work are largely due to differences in problem representations. In particular, we hypothesize that the phrasing of text-only representations impacts comprehension. Furthermore, we posit that while visualizations can be effective tools for communicating Bayesian reasoning, simply appending visualizations to complex textual information will adversely impact comprehension.

In the succeeding sections, we present the results of two experiments that were designed to investigate the interacting effect of both problem representation and spatial ability on communicating Bayesian reasoning problems. Together, these experiments address the question of how users make sense of Bayesian reasoning problems under different, and sometimes competing, representations of complex information. Our first experiment establishes a baseline, text-only condition and investigates how various forms of problem phrasing impacts accuracies. With our second experiment, we explore the interaction between textual information and visualizations when they are combined in a single representation, and the effect of users' spatial ability on their performance.

4 Experiment 1: Text-Only Representations

A survey of the prior work reveals many inconsistencies among Bayesian problems used for assessing Bayesian reasoning. Many past experiments have used their own Bayesian problems, with differing scenarios, wordings and framings. For instance, in their work, Gigerenzer and Hoffrage [17] used 15 different Bayesian problems, each with differing real-world implications and potential cognitive biases associated with them (e.g. being career oriented leads to choosing a course in economics, or carrying heavy books daily relates to a child having bad posture). Micallef et al. [29] and Garcia-Retamero and Hoffrage [16] each used three different Bayesian problems (with only one in common). Brase [1] used a single Bayesian problem not previously tested by other researchers.

Of the existing work, Brase [1] reported the highest accuracies for his text-only condition, with 35.4% of his participants reaching the correct Bayesian response. In addition to using natural frequencies, Brase [1] used **probing** as a means of evaluating the effectiveness of representations. Probing is a technique by which a series of questions are posed to the user that are designed to guide that user through the calculation process. Cosmides and Tooby [9] proposed that probing can be used to help users uncover information that is necessary for solving Bayesian inference problems and thereby improves performance. Rather than asking the participant to calculate the true positive rate from the given information directly (the task that is traditionally given), they used probing to guide their participants' Bayesian calculations. Ultimately, probing was designed to assess whether the user understands the information as it is presented instead of their mathematical skills. Following Brase [1], we examined probing as one of our study conditions.

In his study, Brase [1] also used a **narrative** - a generalizable, hypothetical scenario. Instead of presenting information about a specific disease such as breast cancer, he presented a fictional narrative, introducing a population in which individuals are exposed to a new disease (*"Disease X"*). By using a hypothetical population and a generic disease name, we hypothesize that this generalizes the problem and may have mitigated biases related to a certain disease, thus impacting accuracies.

In addition to these two techniques (probing and narrative), we adapted framing principles for reducing the complexity of text representations [9, 39]. Prior studies suggest that **framing** can significantly impact decision making with probability problems [9]. For example, saying *10 out of 100 people will have the disease* versus *90 out of 100 people will not have the disease* can elicit very different responses [39], and presenting both frames can help mitigate biases known as framing effects [39]. Using both frames also has the advantage that

Table 1. The three questions used in Experiment 1

Text$_{orig}$	10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. *Here is a new representative sample of women at age forty who got a positive mammography in routine screening.* *How many of these women do you expect to actually have breast cancer? ____ out of ____*
Text$_{probe}$	10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. *Imagine 1000 people are tested for the disease.* *(a) How many people will test positive? ____* *(b) Of those who test positive, how many will actually have the disease? ____*
Text$_{diseaseX}$	There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection found in the population. There is a test to detect whether or not a person has the disease, but it is not perfect. Here is some information about the current research on Disease X and efforts to test for the infection that causes it. There is a total of 1000 people in the population. Out of the 1000 people in the population, 10 people actually have the disease. Out of these 10 people, 8 will receive a positive test result and 2 will receive a negative test result. On the other hand, 990 people do not have the disease (that is, they are perfectly healthy). Out of these 990 people, 95 will receive a positive test result and 895 will receive a negative test result. *Imagine 1000 people are tested for the disease.* *(a) How many people will test positive? ____* *(b) Of those who test positive, how many will actually have the disease? ____*

it explicitly states relationships in the problem that are implicit in the original text.

4.1 Design

In line with our research goals of disambiguating contradictory results in previous research, our first experiment examines how these three techniques (*framing*, adding a *narrative* and using *probing*) can be combined to reduce the complexity of Bayesian reasoning problems. In the context of Bayesian problems, the term complexity can have different meanings, for instance: the number of relationships in the problem [24], the number of steps needed to solve the problem, or the amount of information to be integrated or reconstructed [9]. In the current work, we define complexity as the difficulty of extracting information. This hinges on the notion that the simplicity of a task partly depends on the *how* the information is presented. We believe that is it important first to establish a baseline text representation (i.e. **no visualization**) before we consider the effect of adding visualizations.

We conducted an online study and tested three different text-only representations of Bayesian reasoning problems:

Text$_{orig}$ For our base condition, we chose the mammography problem (see Table 1 $Text_{orig}$) since it has been used in many studies [11, 17, 16, 29, 24] and tests a skill of great importance and generalizability [11]. This specific mammography problem was used by Gigerenzer and Hoffrage [17] and Micallef et al. [29], and includes the base rate, true positive rate, and false positive rate. The expected answer was *8 out of 103*.

Text$_{probe}$ The information presented in this condition is exactly the same as $Text_{orig}$ but uses probing instead of asking for the true positive rate directly (see Table 1 $Text_{probe}$). The participant is first probed for the expected number of people who will be tested positive (*103*) then she is probed for the true positive count (*8*). The two probed questions used in the current design are:

Positive Count How many people will test positive?
True Positive Count Of those who test positive, how many will actually have the disease?

Text$_{diseaseX}$ For this condition, we adopted a narrative similar to Brase [1] for the mammography problem, and similarly to $Text_{probe}$ we used probing. In addition to using these two techniques, this condition also provides the user with both positive and negative frames of the problem. Instead of only providing the base rate, the true positive rate, and the false positive rate, the text included the true negative and the false negative rates (see Table 1 $Text_{diseaseX}$). It is important to note that no *new* data was added. The true negative and false negative rates were implicit in the $Text_{orig}$ and $Text_{probe}$ conditions.

Table 2. Demographics for Experiment 1

N	100
Gender	Female: 35%, Male: 65%
Education	High School: 13%, College: 42%, Graduate School: 25%, Professional School: 17%, Ph.D.: 1%, Postdoctoral: 2%
Age	$\mu : 33.63, \sigma : 11.8$, Range: 19 - 65

4.1.1 Participants

We recruited 100 online participants (37 for $Text_{orig}$, 30 for $Text_{probe}$ and 33 for $Text_{diseaseX}$) via Amazon's Mechanical Turk. Participants received a base pay of $.50 with a bonus of $.50 and $.05 for each correct answer on the main task and surveys respectively. The total possible renumeration was $2.80 which is comparable to the U.S. minimum wage. Participants completed the survey via an external link and performed tasks using an online experiment manager developed for this study. Using this tool, participants were regulated by their Mechanical Turk worker identification number and were only allowed to complete the experiment once. Table 2 summarizes the demographic information for Experiment 1.

4.1.2 Procedure

After selecting the task from the Mechanical Turk website, participants were instructed to navigate to a specified external link. Once there, they entered their Mechanical Turk worker identification number which was used both for controlling access to the experiment manager and for remuneration. After giving informed consent, participants were randomly presented with one of the three Bayesian reasoning problems. They were instructed to take as much time as needed to read and understand the information provided as they would have to answer questions based on the information and that bonuses will be paid for each correct answer. To separate the time spent reading the question from the time spent actually solving the problem, the question was not visible until they clicked the appropriately labeled button to indicate that they were ready. The participants were once again instructed to take as much time as needed and enter the answers in the space provided. The timer ended when the participant entered an answer. Any edits to their responses extended the recorded time. Once they submitted the main task, they completed a short demographic questionnaire.

4.2 Results

For our analysis, responses were only deemed correct if participants entered the expected response for **both** probed questions.

With the *Text$_{orig}$* condition, we successfully replicated prior results of Micallef et al. [29] with an accuracy rate of 5.4% as compared to their reported 6% for text-only representations. Modifying the original question by using probing (*Text$_{probe}$*), we presented participants with questions that were easier to understand [9]. Consistent with prior work by Cosmides and Tooby [9], we found that this small change yielded a significantly higher accuracy rate of 26.7%.

Finally, by changing the problem text with our *Text$_{diseaseX}$* condition, we successfully replicate Brase's [1] results with an accuracy rate of 42.4% as compared to his reported 53.4%. A chi-square test was conducted across all participants and revealed significant differences between the accuracy rates of the three conditions($\chi^2(2, N = 100)=13.27$ $p = 0.001$). Performing a pairwise chi-square test with a Bonferroni adjusted alpha ($\alpha = 0.017$), we found significant differences between *Text$_{orig}$* and *Text$_{probe}$* ($\chi^2(1, N = 67) = 5.9$, $p = 0.015$), and *Text$_{orig}$* and *Text$_{diseaseX}$* ($\chi^2(1, N = 70) = 13.56$, $p < 0.001$).

4.3 Discussion

In our first experiment, we found that by simply changing how the problem was presented, we observed an improvement in participants' overall accuracy from 5.4% to 42.4%. We adapted techniques such as probing, which nudges the user to think more thoroughly about the problem, adding a narrative which generalized the problem, and presenting both frames for mitigating framing effects.

Taken together this gives us insight into how lexical choices of text-only representations of Bayesian reasoning problems govern their effectiveness and may at least partially explain the poor or inconsistent accuracies observed in previous work. By using probing alone, our results showed a significant improvement over our base condition which used direct questioning. This suggests that assessment techniques for Bayesian reasoning problems should be thoroughly scrutinized.

Participants were even more accurate when the stimulus combined all three techniques (probing, narrative and framing). This finding provides initial evidence that even with text-only representations (i.e. without visualization aids), the phrasing of the problem can impact comprehension. Indeed, there were several factors that potentially contributed to the increase in communicative competence we observed for *Text$_{diseaseX}$*. For example, using the generic term *Disease X* instead of a specific disease may gave mitigated biases introduced by the mammography problem. Alternatively, the observed increase in accuracy could be attributed to the overall readability of the text or the amount of data presented in the conditions (the *Text$_{diseaseX}$* condition presented the user with slightly more explicit data than the *Text$_{orig}$* and *Text$_{probe}$* conditions). Deciphering these was beyond the scope of this project, but will be an important direction for future work.

In the following study, we further address our research goals by investigating the effect of adding visualizations for representing Bayesian reasoning tasks. We use our results from this initial experiment by adopting *Text$_{diseaseX}$* as a baseline text-only representation for evaluating different text and visualization designs.

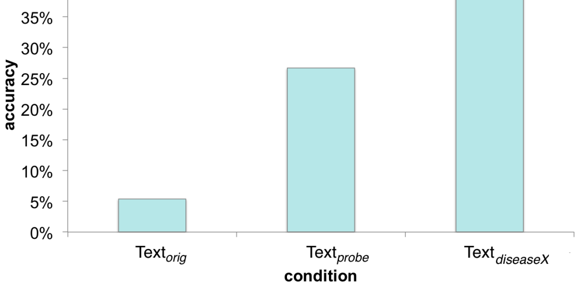


Fig. 1. Accuracies across all conditions in Experiment 1. Combining probing and narrative techniques proved to be effective for reducing the overall complexity of the text and increasing accuracy.

5 Experiment 2: Text and Visualization

Although visualization has been suggested as a solution to the Bayesian reasoning problem, recent findings suggest that, across several designs, simply adding visualizations to textual Bayesian inference problems yields no significant performance benefit [29, 32]. Micallef et al. [29] also found that removing numbers from the textual representation can improve performance. The findings of this prior work suggest an interference between text and visualization components when they are combined into a single representation.

Differing from prior work which focused mainly on comparing different visualization designs [29], our second experiment aimed to progress Bayesian reasoning research by further investigating the effect of presenting text and visualization together. We examined the amount of information presented to the user and the degree to which the textual and visual information are integrated. Grounded by the baseline condition established in Experiment 1 (Table 3 **Control-Text**), we tested representations that gradually integrate affordances of visualizations or the visualization itself.

One affordance of visualizations is that relationships that are implicitly expressed in text are often explicated in visual form. Visualizations make it easier to "see" relationships among groups. To bridge this information gap, we gradually expanded the text-only representation to explicate implied information and relationships.

Secondary to our main research goals and motivated by the prior work demonstrating a connection between spatial ability and visual design [25, 45], our second experiment also aimed to understand *how* nuances in spatial ability affect users' capacity to use different representations of Bayesian reasoning problems. Since prior research suggests that low spatial-ability users may experience difficulty when both the text and visual representations are presented [45], we hypothesize that low spatial-ability users would be more adept at using representations which integrated affordances of the visualization but not the visualization itself. On the other hand, we hypothesize that high spatial-ability users will benefit greatly from representations which merge textual and visual forms, as they are more likely to possess the ability to effectively utilize both representations.

5.1 Design

To test our hypotheses, we present the Text$_{diseaseX}$ condition from Experiment 1, using a variety of representations. Our intent was to manipulate the total amount of information presented, as well as the coupling between the problem text and visual representation. Consistent with Text$_{diseaseX}$, each condition in Experiment 2 began with an introductory narrative:

> There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection found in the population. There is a test to detect whether or not a person has the disease, but it is not perfect. Here is some information about the current research on Disease X and efforts to test for the infection that causes it.

The format of the questions asked were also consistent with the Text$_{diseaseX}$:

> (a) How many people will test positive? ____
> (b) Of those who test positive, how many will actually have the disease? ____

5.1.1 Conditions

There were a total of 6 conditions which were randomly assigned to our participants (see Table 3 for the exact stimuli).

Control-Text As the name suggests, this is our control condition and uses the same text format as was presented in the **Text$_{diseaseX}$** condition of Experiment 1.

Table 3. Table showing the 6 conditions used in Experiment 2

Control-Text

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

Complete-Text

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

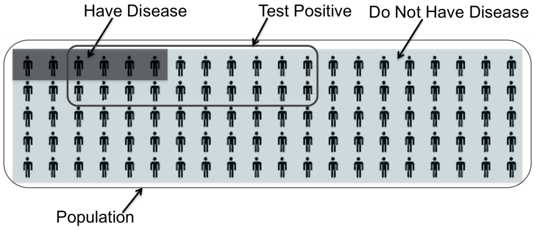
Another way to think about this is... Out of the 100 people in the population, 20 people will test positive. Out of these 20 people, 4 will actually have the disease and 16 will not have the disease (i.e., they are perfectly healthy). On the other hand, 80 people will test negative. Out of these 80 people, 2 will actually have the disease and 78 will not have the disease (i.e., they are perfectly healthy).

Structured-Text

There is a total of 100 people in the population.
Out of the 100 people in the population,
6 people actually have the disease. Out of these 6 people,
4 will receive a positive test result and
2 will receive a negative test result.
On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people,
16 will receive a positive test result and
78 will receive a negative test result.

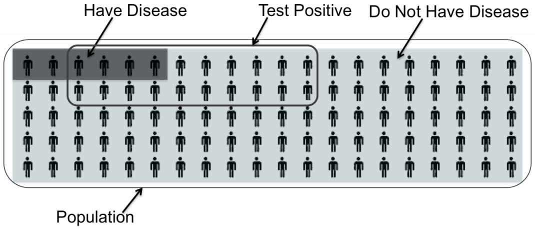
Another way to think about this is...
Out of the 100 people in the population,
20 people will test positive. Out of these 20 people,
4 will actually have the disease and
16 will not have the disease (i.e., they are perfectly healthy).
On the other hand, 80 people will test negative. Out of these 80 people,
2 will actually have the disease and
78 will not have the disease (i.e., they are perfectly healthy).

Vis-Only



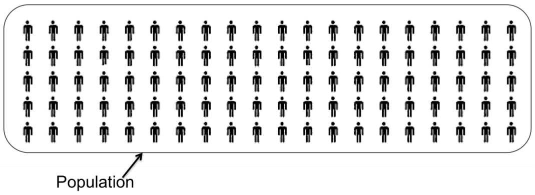
Control+Vis

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

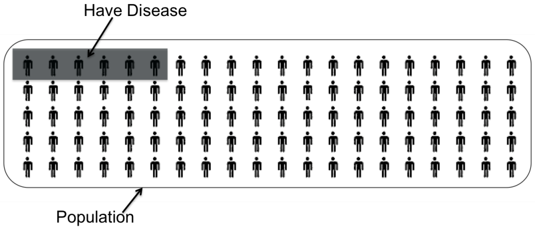


Storyboarding

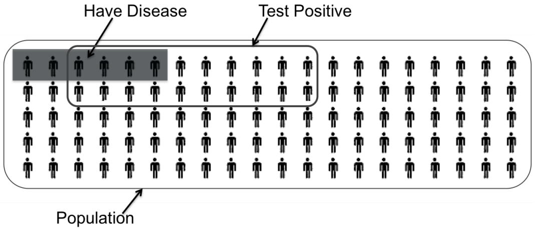
There is a total of 100 people in the population.



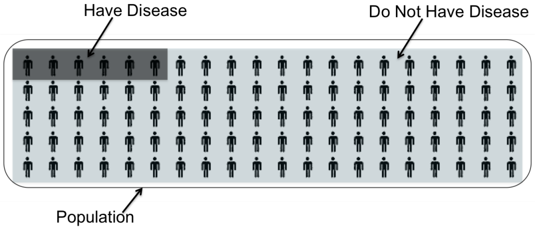
Out of the 100 people in the population, 6 people actually have the disease.



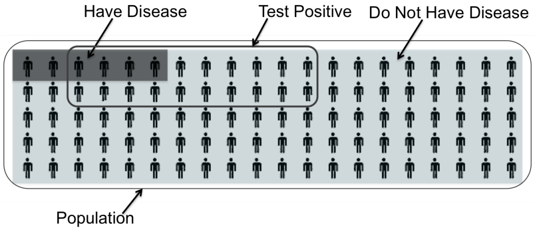
Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result.



On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy).



Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.



Complete-Text In this condition, the text is expanded to present all possible relationships and framings of the problem, which is a common affordance of visualizations. It is important to note that the text still presents the same amount of information as *Control-Text* (i.e. the base rate, the true positive rate, the false positive rate, the false negative rate and the true negative rate), however, it presents the data both with respect to having the disease and being tested positive (see Table 3 *Complete-Text*).

Structured-Text Here, we further improve the text by integrating another affordance of visualizations. Like *Complete-Text*, the text in this condition enumerates all possible relationships and framings of the problem, however, we enhanced the text by adding visual cues to the representation. Instead of using long-form paragraphs, we used indentation to clarify relationships. Similar to spreadsheets of raw data, the spatialization of the information makes the relationships more apparent.

Vis-Only With this condition, we establish a baseline for using visualizations. With the exception of the introductory narrative mentioned above, there is no additional text in this condition.

While researchers have investigated numerous visualization designs for representing Bayesian reasoning (see Section 2), there is still no consensus on which is best. In fact, recent research in the visualization community comparing the effectiveness of 6 different visualization designs found no significant difference between them [29].

That said, for this visualization-only condition, we chose to represent the information using an icon array visualization (see Table 3). A number of researchers have explored their utility for risk communication and Bayesian reasoning [1, 14, 15, 16, 19, 20, 25, 29, 32, 35] and icon-arrays are often used in the medical community for representing risk information.

The specific icon-array used in this work consists of a 5 by 20 grid of anthropomorphic figures. We adapted a sequential layout for the different sets (as opposed to a random layout which has been previously used for representing uncertainty [19]) and we used in-place labeling for ease of reference. This is similar to the design used by Brase [1].

Control+Vis Mirroring prior work [1, 29, 32] that investigated the utility of adding visualization designs to Bayesian problems, here we simply added a visualization to our control text-only representation. The information for this condition is represented using both the *Control-Text* description and the icon array visualization from *Vis-Only*.

Storyboarding This condition was designed to simplify Bayesian reasoning by gradually integrating the textual and visual components of *Control+Vis*. Such storytelling techniques are becoming increasingly popular in recent years [22, 23, 36] and have even been recently referred to as "the next step for visualization" [27].

For our *Storyboarding* design, no information was added, but the information is presented sequentially, allowing for temporal processing. The text shown in this condition is consistent with *Control-Text* and the final visualization is the same as *Vis-Only*.

5.1.2 Cognitive Ability Measures

We measured participants' spatial ability using the paper folding test (VZ-2) from Ekstrom, French, & Hardon [12]. This survey consists of 2 3-minute sessions with 10 questions each. A similar version of the test has been used as a standard technique to compare spatial ability to Bayesian reasoning skills in other studies [24, 29]. Consistent with prior work [24, 29], a participant's spatial ability score was calculated by summing the number of correct answers minus the total number of incorrect answers divided by four.

Consistent with prior studies investigating the effectiveness of Bayesian reasoning problem representations [16, 25, 29], we measured participants' numerical skills. This was measured using Brown et al.'s 6-question test [3]. Prior research has demonstrated a correlation between numerical skills and understanding natural frequencies [4] and numerical skills has been shown to correlate with one's ability to understand medical risk information [3, 34].

Table 4. Demographics for Experiment 2

N	377
Gender	Female: 34.2%, Male: 65%, Unspecified: .8%
Education	High School: 22.3%, College: 51.5%, Graduate School: 19.6%, Professional School: 4.2%, Ph.D.: 1.6%, Postdoctoral: .5%, Unspecified: .3%
Trained$^+$	Yes: 12.2%, No: 87.5%, Unspecified: .3%
Age	$\mu : 31, \sigma : 9.87$, Range: 18 to 65
Spatial Ability	$\mu : 8.60, \sigma : 5.25$, Range: -3.75 to 20
Numeracy	$\mu : 4.23, \sigma : 1.22$, Range: 0 to 6

$^+$Participants received statistics training

5.1.3 Participants

We recruited 377 participants (61-65 per condition) via Amazon's Mechanical Turk who had not completed Experiment 1. The recruitment and remuneration techniques used for this experiment follows that of Experiment 1 (see Section 4.1.1). Table 4 summarizes our participants' demographics.

5.1.4 Procedure

The procedure for this experiment also follows that of Experiment 1 (see Section 4.1.2) except for the following changes. After the main tasks, in addition to the demographics survey, participants completed VZ-2 to measure spatial ability and the numerical skill survey.

5.2 Hypotheses

Following our high-level claims that text complexity, visual representation, and spatial ability affect accuracy on Bayesian reasoning, we form the following hypotheses:

H1 Prior work shows no performance increase when visualizations are simply appended to existing text representations [29]. Therefore, removing the textual description completely from the visualization condition will mitigate this effect and *Vis-Only* will be more effective than *Control-Text* and *Control+Vis*.

H2 Since participants are given an increased amount of information in the text, users will perform better on *Structured-Text* and *Complete-Text* than on *Control-Text*.

H3 High spatial-ability users will perform better overall than low spatial-ability users.

H4 Designs that include both text and vis (*Storyboarding, Control+Vis*) will require higher spatial ability than text-only designs (*Complete-Text, Structured-Text*) and *Vis-Only*.

5.3 Results

While recent work [29] has advocated for a more continuous or fine-grained approach to assessing users' accuracy on Bayesian reasoning tasks (for instance, reporting the differences between users' responses and correct answers in terms of a log ratio), we report our accuracies in terms of the percentages of correct exact answers. Choosing this binary approach of assessing accuracy has two advantages: (1) it allows us to directly compare our results across the prior body of work as they have all (including [29]) reported their accuracies similarly, and (2) this course-grained approach is especially user-friendly for comparing representations with substantial accuracies as seen in the subsequent sections.

Consistent with Experiment 1, the proceeding analyses focus only on participants who answered both questions correctly (see Section 5.1

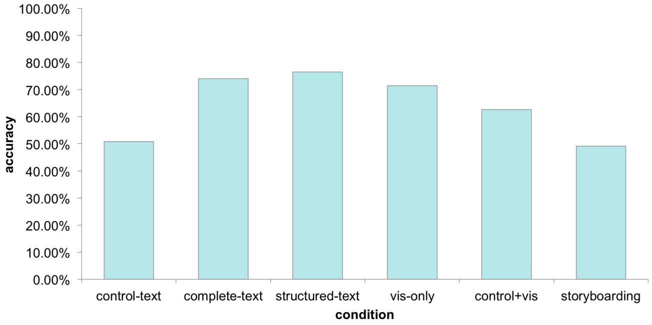


Fig. 2. Accuracies across all conditions. We found that participants were most accurate with *Structured-Text*, *Complete-Text* and *Vis-Only*.

for the exact questions asked). In an effort to further simulate a lay population, our analysis excluded participants who reported to have had statistical training.

5.3.1 Accuracy Across Designs

Across all conditions, the average accuracy was remarkably high; 63% of the participants correctly answered both questions. Figure 2 summarizes the accuracies across all conditions. *Complete-Text*, *Structured-Text* and *Vis-Only* yielded the highest overall accuracies ranging from 71% to 77%. Along with *Control-Text*, *Storyboarding* yielded the lowest overall accuracies with only 51% and 49% respectively of the participants responding correctly to the questions.

We performed a chi-square analysis to test for differences in accuracy across the six conditions. The test revealed that the percentage of participants who correctly answered both questions differed by design ($\chi^2(5, N = 330) = 17.2$, $p = 0.004$). We then performed all pairwise chi-square tests with a Bonferroni adjusted alpha ($\alpha = 0.003$) to identify the specific designs that deferred. The analysis revealed a significant differences between only *Storyboarding* and *Structured-Text* ($\chi^2(1, 114) = 8.8$, $p < 0.003$).

We found no significant difference in accuracy between *Control-Text* and *Vis-Only*, and we found no difference in accuracy among *Control-Text*, *Complete-Text* and *Structured-Text*. Consistent with prior findings [29, 32], we also found no significant difference between the *Control-Text* and *Control+Vis* conditions. This suggests that under the studied conditions, using visualizations (with or without textual information) and increasing the amount of explicit textual information in text-only designs did not improve performance, thereby rejecting both **H1** and **H2**.

5.3.2 Spatial Ability and Accuracy

To test our hypothesis that spatial ability affects participants' capacity to extract information from the different representations, we performed a binary logistic regression to predict participants who correctly answered both questions using their spatial abilities score as a predictor. A test of the resulting model against the constant model was statistically significant at $p < 0.001$, indicating that spatial ability highly correlates with participants' ability to answer the questions accurately. Prediction success overall was 71.5% (87.1% for predicting those who responded correctly and 44.6% for predicting those who responded incorrectly).

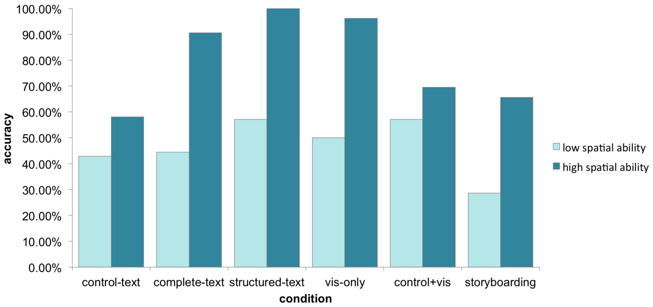


Fig. 3. Average accuracy for the low and high spatial ability groups for each design. Overall, we found that high spatial users were much more likely to correctly answer the question prompts.

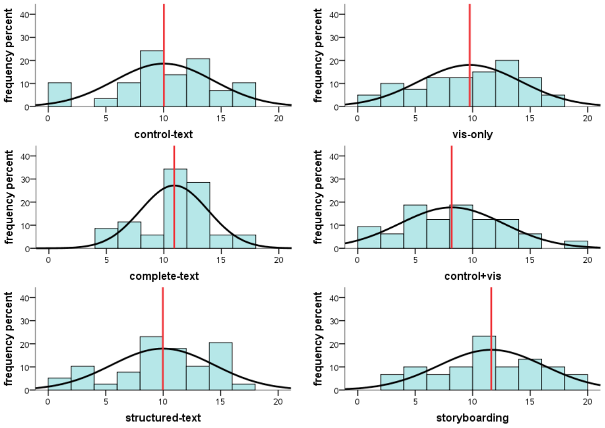


Fig. 4. Histograms showing the distribution of spatial ability scores for participants who correctly answered both questions across the six conditions. The graphs provide preliminary evidence that *Complete-Text* and *Storyboarding* may require higher spatial ability to use them.

For a more specific analysis, we split users into two groups (spatial$_{low}$ and spatial$_{high}$) based on a median split of their spatial abilities scores (spatial$_{low}$ < 9, N = 170 and spatial$_{high}$ $>= 9$, N = 160). Confirming **H3**, the overall accuracy for the spatial$_{high}$ group was 78.8% while spatial$_{low}$ was 46.9%. Figure 3 summarizes the groups' accuracies for each of the six conditions.

We then performed separate chi-square analyses testing for significant differences between the accuracies of the six conditions for the spatial$_{low}$ group and the spatial$_{high}$ group. The chi-square test for the spatial$_{high}$ group was significant ($\chi^2(5, \text{N} = 170) = 26.3$, $p < 0.001$) and multiple comparisons with the Bonferroni adjusted alpha ($\alpha = 0.003$) revealed significant differences between:

- *Control-Text* & *Complete-Text* ($\chi^2(1, N{=}63) = 8.8$, $p < 0.003$)
- *Control-Text* & *Structured-Text* ($\chi^2(1, N{=}54) = 12.7$, $p < 0.001$)
- *Control-Text* & *Vis-Only* ($\chi^2(1, N{=}57) = 11.07$, $p < 0.001$)
- *Structured-Text* & *Storyboarding* ($\chi^2(1, N{=}65) = 13.5$, $p < 0.001$)

These results indicate that for the spatial$_{high}$ group, *Structured-Text*, *Complete-Text*, and *Vis-Only* resulted in improved performance over our control condition (*Control-Text*), confirming **H2** and partially supporting **H1**. More generally, the results imply that spatial ability must be considered when evaluating the effectiveness of Bayesian reasoning designs.

Performing a similar analysis with the spatial$_{low}$ group, we found no significant difference between the accuracies for the six conditions ($\chi^2(5, N = 160) = 6.5$, $p = 0.262$). This indicates that the accuracies for the spatial$_{low}$ group were similar across the all conditions, suggesting that the proposed designs were ineffective for low spatial-ability users.

5.3.3 Spatial Ability Across Designs

Given our findings that participants' spatial ability affects their likelihood of correctly answering the question prompts using a given representation, we hypothesize that we can now use spatial ability as a tool for ranking representations based on their complexity. In particular, we use spatial ability as a proxy for measuring and comparing the extraneous cognitive load necessary to effectively use each representation. Figure 4 shows the distribution for spatial ability scores for the correct users on the six conditions.

Prior to our analysis, we removed two outliers whose spatial ability score was more that two standard deviations from the mean score for their respective conditions. We then conducted a one-way Analysis of Variance (ANOVA) to test for differences between the spatial ability scores of participants who correctly answered both questions for each of the six conditions. Our model was statistically significant (F(5, 206) = 2.57, p = 0.028) suggesting that the spatial ability scores differed across conditions.

Post hoc comparisons using Fisher's least significant difference (LSD) indicated that the mean scores of the following conditions differed significantly:

- *Control-Text* & *Control+Vis* (p = 0.042)
- *Complete-Text* & *Control+Vis* (p = 0.005)
- *Storyboarding* & *Control+Vis* (p = 0.002)

This finding supports our hypothesis that some representations may require higher spatial ability to use them. However, we partially reject **H4**. *Control+Vis* had the lowest average indicating that this representing may be most suitable for users with lower spatial ability. Conversely, we found that the average spatial ability score for correct users on *Storyboarding* was higher than all other conditions, suggesting that *Storyboarding* was the most difficult representation to use.

5.4 Discussion

The results of our Experiment 1 demonstrated how phrasing of Bayesian problems can influence performance. In Experiment 2, we examined whether we could improve problem representations by enhancing text or combining it with visualization. In an effort to bridge the information gap between text and visual representations, we studied text-only representations that clarified information that usually is more easily seen in a visualization. Our *Complete-Text* design sought to decrease this information gap by enumerating all probability relationships in the text and our *Structured-Text* design used indentations to visualize these relationships. Still, when spatial ability was not considered, we found that adding more information did not benefit users.

We observed similar results with our visualization conditions. Although we hypothesized that the *Vis-Only* design would be more effective than *Control-Text* and *Control+Vis*, our results did not support this hypothesis. Again, when spatial ability was not considered, adding visualizations (with or without textual information) did not improve performance. However, a closer examination of our results adds nuance to this finding when individual differences are considered.

5.4.1 Spatial Ability Matters

While the lack of overall difference across conditions was unexpected, factoring in the effect of spatial ability helped shed light on these findings. Across all visualizations, spatial ability was a significant indicator of accuracy and completion times. We found that users with low spatial ability generally performed poorly; the accuracy of high spatial-ability users was far higher than the accuracy of low spatial-ability users (78.8% v. 46.9%). Relative to the *Control-Text* condition, for high spatial users, the *Structured-Text, Complete-Text* and *Vis-Only* designs were extremely effective, yielding accuracies of 100%, 90% and 96% respectively. These unprecedented accuracies suggest that, for users with high spatial ability, these designs can solve the problem of Bayesian reasoning. However, it is interesting to note that effective designs were "pure" designs (i.e., they did not combine text and visualizations).

5.4.2 Text+Vis Interference

For high spatial-ability users, we found that representations that combined text and visualization (*Control+Vis, Storyboarding*) actually impeded users' understanding of conditional probability when compared to text-only (*Complete-Text, Structured-Text*) or *Vis-Only* conditions. Despite the fact that high spatial-ability users performed comparatively poorly with the *Control+Vis* design (accuracy decreased by nearly 30% when compared to *Complete-Text, Structured-Text*, and *Vis-Only*), such disparity in accuracy was not observed with low spatial ability users using *Control+Vis*. One possible explanation relies on considering the problem as a mental modeling task. Users with low spatial ability may have simply chosen the representation in *Control+Vis* (text or visualization) that best fit their understanding of the problem. On the contrary, high spatial-ability users may have attempted (and failed) to integrate the text and visualization representations in order to find the correct answer. This hypothesis would be in line with Kellen's [25] hypothesis that text and visual representations in a complex problem may compete for the same mental resources, increasing the likelihood of errors.

The *Storyboarding* design proved to be an enormous obstacle for the user. Performing analysis to investigate the spatial ability scores required to successfully extract information from the six designs revealed that *Storyboarding* demand higher spatial ability scores than the other designs. While it is intended to gradually guide users through the Bayesian reasoning problem, the different steps may have inadvertently introduced distractors to the information that the user is truly looking for and/or forced users into a linear style of reasoning that was incongruent with their mental model of the problem. This added complexity increased cognitive load to a point that accuracy for all users suffered.

Still, such storytelling techniques have been shown to be effective for communicating real world data [22, 23, 27, 36]. The tasks in this study, however, go beyond typical information dissemination, as users had to understand information known to be inherently challenging for most people. Future work could investigate the utility of storytelling techniques for similar reasoning tasks.

6 Conclusion

Effectively communicating Bayesian reasoning has been an open challenge for many decades, and existing work is sparse and sometimes contradictory. In this paper we presented results from two experiments that help explain the factors affecting how text and visual representations contribute to performance on Bayesian problems. With our first experiment, we showed that the wording of text-only representations can significantly impact users' accuracies and may partly be responsible for the poor or inconsistent findings observed by prior work.

Our second experiment examined the effects of spatial ability on Bayesian reasoning tasks and analyzed performance with a variety of text and visualization conditions. We found that spatial ability significantly affected users ability to use different Bayesian reasoning representations. Compared to high spatial-ability users, low spatial-ability users tended to struggle with Bayesian reasoning representations. In fact, high spatial-ability users were almost two times more likely to answer correctly than low spatial-ability users. Additionally, we found that text-only or visualization-only designs were more effective than those which blend text and visualization.

Ultimately, our results not only shed light on how problem representation (both in text phrasing and combining text and visualization) can affect Bayesian reasoning, but also question whether one-size-fits-all visualizations are ideal. Further study is needed to clarify how best to either adapt visualizations or provide customization options to serve users with different needs. The results from these studies can be used for real-world information displays targeted to help people better understand probabilistic information. They also provide a set of benchmark problem framings that can be used for more comparable future evaluations of visualizations for Bayesian reasoning. Further work in this domain can have significant impact on pressing issues in the medical communication field and other domains where probabilistic reasoning is critical.

7 Dataset

To facilitate future work, participants' data are made available at: http://github.com/TuftsVALT/Bayes.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1218170.

REFERENCES

[1] G. L. Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3):369–381, 2009.

[2] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 20(12), 2014.

[3] S. M. Brown, J. O. Culver, K. E. Osann, D. J. MacDonald, S. Sand, A. A. Thornton, M. Grant, D. J. Bowen, K. A. Metcalfe, H. B. Burke, et al. Health literacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient education and counseling*, 83(1):92–98, 2011.

[4] G. B. Chapman, J. Liu, et al. Numeracy, frequency, and bayesian reasoning. *Judgment and Decision Making*, 4(1):34–40, 2009.

[5] C. Chen and M. Czerwinski. Spatial ability and visual navigation: An empirical study. *New Review of Hypermedia and Multimedia*, 3(1):67–89, 1997.

[6] C. A. Cohen and M. Hegarty. Individual differences in use of external visualisations to perform an internal visualisation task. *Applied Cognitive Psychology*, 21:701–711, 2007.

[7] W. Cole. Understanding bayesian reasoning via graphical displays. In *ACM SIGCHI Bulletin*, volume 20, pages 381–386. ACM, 1989.

[8] W. Cole and J. Davidson. Graphic representation can lead to fast and accurate bayesian reasoning. In *Symp Computer Application in Medical Care*, pages 227–231, 1989.

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

[10] R. De Beni, F. Pazzaglia, V. Gyselinck, and C. Meneghetti. Visuospatial working memory and mental representation of spatial descriptions. *European Journal of Cognitive Psychology*, 17(1):77–95, 2005.

[11] D. Eddy. Probabilistic reasoning in clinical medicine: Problems and opportunities, 1982, 249-267. pages 249–267, 1982.

[12] R. B. Ekstrom, J. W. French, H. H. Harman, and D. Dermen. Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational Testing Service*, 1976.

[13] H. Friederichs, S. Ligges, and A. Weissenstein. Using tree diagrams without numerical values in addition to relative numbers improves students numeracy skills a randomized study in medical education. *Medical Decision Making*, 2013.

[14] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, 28(2):210, 2009.

[15] R. Garcia-Retamero and M. Galesic. Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social science & medicine*, 70(7):1019–1025, 2010.

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

[17] G. Gigerenzer and U. Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684, 1995.

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

[19] P. K. Han, W. M. Klein, B. Killam, T. Lehman, H. Massett, and A. N. Freedman. Representing randomness in the communication of individualized cancer risk estimates: effects on cancer risk perceptions, worry, and subjective uncertainty about risk. *Patient education and counseling*, 86(1):106–113, 2012.

[20] S. T. Hawley, B. Zikmund-Fisher, P. Ubel, A. Jancovic, T. Lucas, and A. Fagerlin. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, 73(3):448–455, 2008.

[21] M. Hegarty. Capacity limits in diagrammatic reasoning. In *Theory and application of diagrams*, pages 194–206. Springer, 2000.

[22] J. Hullman and N. Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2231–2240, 2011.

[23] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher, and E. Adar. A deeper understanding of sequence in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2406–2415, 2013.

[24] V. J. Kellen. The effects of diagrams and relational complexity on user performance in conditional probability problems in a non-learning context. In *Doctoral Thesis*. DePaul University, 2012.

[25] V. J. Kellen, S. Chan, and X. Fang. Facilitating conditional probability problems with visuals. In *Human-Computer Interaction. Interaction Platforms and Techniques*, pages 63–71. Springer, 2007.

[26] D. Kimura. *Sex and cognition*. MIT press, 2000.

[27] R. Kosara and J. Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013.

[28] L. Martignon and C. Wassner. Teaching decision making and statistical thinking with natural frequencies. In *Proceedings of the Sixth International Conference on Teaching of Statistics. Ciudad del Cabo: IASE. CD ROM*, 2002.

[29] L. Micallef, P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2536–2545, 2012.

[30] A. B. Miller, C. J. Baines, P. Sun, T. To, and S. A. Narod. Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *BMJ: British Medical Journal*, 2014.

[31] A. Ottley, R. J. Crouser, C. Ziemkiewicz, and R. Chang. Manipulating and controlling for personality effects on visualization tasks. *Information Visualization*, 2013.

[32] A. Ottley, B. Metevier, P. K. Han, and R. Chang. Visually communicating bayesian statistics to laypersons. In *Technical Report*. Tufts University, 2012.

[33] A. Ottley, H. Yang, and R. Chang. Personality as a predictor of user strategy: How locus of control affects search strategies on tree visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2015.

[34] V. F. Reyna, W. L. Nelson, P. K. Han, and N. F. Dieckmann. How numeracy influences risk comprehension and medical decision making. *Psychological bulletin*, 135(6):943, 2009.

[35] P. Sedlmeier and G. Gigerenzer. Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3):380, 2001.

[36] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1139–1148, 2010.

[37] D. Spiegelhalter, M. Pearson, and I. Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011.

[38] J. Tsai, S. Miller, and A. 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, 2011.

[39] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[40] M. C. Velez, D. Silver, and M. Tremaine. Understanding visualization through spatial ability differences. In *IEEE Visualization*, pages 511–518. IEEE, 2005.

[41] K. J. Vicente, B. C. Hayes, and R. C. Williges. Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(3):349–359, 1987.

[42] H. G. Welch and W. C. Black. Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613, 2010.

[43] H. G. Welch et al. Overdiagnosis and mammography screening. *Bmj*, 339, 2009.

[44] J. S. Yi. Implications of individual differences on evaluating information visualization techniques. In *Proceedings of the BELIV Workshop*, 2010.

[45] C. Ziemkiewicz and R. Kosara. Preconceptions and individual differences in understanding visual metaphors. *Computer Graphics Forum*, 28(3):911–918, 2009. Proceedings EuroVis.

[46] C. Ziemkiewicz, A. Ottley, R. J. Crouser, K. Chauncey, S. L. Su, and R. Chang. Understanding visualization by understanding individual users. *Computer Graphics and Applications, IEEE*, 32(6):88–94, 2012.

[47] C. Ziemkiewicz, A. Ottley, R. J. Crouser, A. R. Yauilla, S. L. Su, W. Ribarsky, and R. Chang. How visualization layout relates to locus of control and other personality factors. *Visualization and Computer Graphics, IEEE Transactions on*, 19(7):1109–1121, 2013.