Does Interaction Improve Bayesian Reasoning with Visualization?

Ab Mosca
Tufts University
amosca01@cs.tufts.edu

Alvitta Ottley
Washington University in St. Louis
alvitta@wustl.edu

Remco Chang
Tufts University
remco@cs.tufts.edu

ABSTRACT

Interaction enables users to navigate large amounts of data effectively, supports cognitive processing, and increases data representation methods. However, there have been few attempts to empirically demonstrate whether adding interaction to a static visualization improves its function beyond popular beliefs. In this paper, we address this gap. We use a classic Bayesian reasoning task as a testbed for evaluating whether allowing users to interact with a static visualization can improve their reasoning. Through two crowdsourced studies, we show that adding interaction to a static Bayesian reasoning visualization does not improve participants' accuracy on a Bayesian reasoning task. In some cases, it can significantly detract from it. Moreover, we demonstrate that underlying visualization design modulates performance and that people with high versus low spatial ability respond differently to different interaction techniques and underlying base visualizations. Our work suggests that interaction is not as unambiguously good as we often believe; a well designed static visualization can be as, if not more, effective than an interactive one.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; **Empirical studies in visualization**.

KEYWORDS

Data Analysis, Reasoning, Problem Solving, Decision Making, Interaction Design, Human-Subjects Quantitative Studies

1 INTRODUCTION

Interaction is core to visualization design and is a vital mode of communication between the user and visual system. In the visualization community, the study of interaction ranges from defining interaction [10, 21, 45], to understanding the interplay between interaction and cognition [27, 34], to leveraging user interactions to improve analytics [5, 13]. However, investigations into the value of adding interaction to a static design are rare and results are varied.

In some cases, the value-add of interaction to visualization is clear. In *The Value of Visualization* van Wijk explains "interaction is generally considered as good", and argues that it is invaluable to tasks such as allowing users to explore more data than can fit on a screen, and to customizing new visualization methods [42]. Heer and Shneiderman echo this sentiment in their taxonomy of interactive dynamics for visual analysis [21]. Additionally, it has been argued that interaction is valuable due to its ability to amplify or illustrate user cognition [27, 34, 45]. A recent study by Zhi et al. found that adding interaction to a storytelling visualization increased engagement [46]. Studies of multimedia instruction have shown that interactivity can increase deep learning and learning transfer [14, 43].

However, the literature does not uniformly support interaction as an indisputable means of improving visualization. In fact, in *The Value of Visualization* immediately after expressing the "good" aspects of interaction van Wijk states that "one could advocate the opposite: interaction should be avoided," and explains that interaction can negatively impact visualization by increasing subjectivity, and the user's perceptual and exploration costs [42]. Lam designed a framework that accounts for potential costs of interaction in information visualization, and encourages designers to weigh the cost against potential gains [25]. A study by Theis et al. [39] comparing task performance on interactive and static uncertainty visualizations found no significant difference in error rate between the two. And a study by Ragan et al. [35] comparing outcomes of a pictorial learning activity given an interactive or automatic view control found no significant differences between the two.

In this paper, we investigate the following research question: "What value can interaction add to a static Bayesian reasoning visualization?" We use a Bayesian reasoning task, because it is a well defined but difficult reasoning problem, with a clear-cut correct answer [23, 29, 32]. Moreover, Bayesian reasoning can be summarized quite succinctly by conditional probabilities and Bayes rule, however this often fails to adequately communicate the real world situation represented by these numbers [17]. As a result, there has been a plethora of research on communicating Bayesian reasoning through static visualization [2, 4, 7, 8, 15–17, 22–24, 29, 31, 32, 36, 38, 41].

In addition to being an open problem area, communicating Bayesian reasoning is an ideal test bed for interaction because interaction is not imperative to the effectiveness of a Bayesian reasoning visualization like it is for most visual analytic systems, which are built to analyze large amounts of data. Static Bayesian reasoning visualizations typically do not represent more data than

can fit on one screen. Thus, adding interaction does not add any otherwise obscured information to the visualization, it simply highlights or draws connections between information already present. This allows us to to isolate the value-add of interaction independent of data exploration and sensemaking.

Based on prior work [23, 24, 29, 32, 41], we postulate that interaction can facilitate visual Bayesian reasoning, but its effects are modulated by: (1) the interaction technique, and (2) the visualization design. Moreover, we expect to see different effects for people with high versus low spatial ability [32]. In this work, we aim to gain a better understanding of how these factors affect the value-add of interaction to a static Bayesian reasoning visualization.

To this end, we run two Amazon Mechanical Turk studies. Experiment 1 investigates the effect of adding interactive checkboxes to three different static (or base) visualizations, which range in their use of Gestalt principles to effectively depict sub-populations of interest in a Bayesian reasoning task. Analysis of Experiment 1 shows that adding interactive checkboxes to a static Bayesian reasoning visualization does not significantly impact participants' reasoning accuracy. Moreover, we do not find a case in which interaction significantly improves participants' performance. Experiment 2 investigates the same three base visualizations, and expands the number of interaction techniques tested (two types of checkboxes, drag and drop, hover, and tooltips). In our analysis of Experiment 2, we again do not find any cases in which interaction significantly improves participants' performance on Bayesian reasoning. Moreover, we find that for participants with high spatial ability, *hover* significantly decreases performance.

To summarize, we make the following contributions:

(1) We demonstrate that adding interaction to a static Bayesian reasoning visualization can (under certain circumstances) decrease users' accuracy on a Bayesian reasoning task.

(2) We provide empirical and observational evidence that the value-add of interaction to a static Bayesian reasoning visualization is dependant on two factors: design of the static visualization, and interaction technique.

(3) We show that adding interaction to a static Bayesian reasoning visualization can lower accuracy of people with high spatial ability on a Bayesian reasoning task, and generally does not effect accuracy of people with low spatial ability on a Bayesian reasoning task.

2 RELATED WORK

Interaction and *Bayesian Reasoning* are widely studied areas in visualization, but they are typically studied in isolation from each other. This paper focuses on the intersection of these two areas. By using Bayesian reasoning as a test bed for interaction techniques, we add to the body of knowledge on interactivity, and Bayesian reasoning visualizations. The following sections discuss related work studying the value add of interaction, Bayesian reasoning visualizations, and the intersection of the two.

2.1 Value add of interaction

Recent work by Dimara and Perin [10] defines interaction (in visualization) as "the interplay between a person and a data interface involving a data-related intent, at least one action from the person

and an interface reaction that is perceived as such." Similarly, Yi et al. [45] and Heer and Shneiderman [21] construct taxonomies of interactions for visualization. Others have endeavored to better explain why interaction is useful to visualization from a cognitive processing standpoint [27, 34]. In addition to theorizing and categorizing interaction, work has been done designing novel interaction techniques, for example [6, 18, 26, 44], and identifying how visualization designers can leverage interaction to learn about users and create customized visualizations [5, 13].

The majority of work on defining, categorizing, theorizing, and leveraging interaction for visualization focuses on visual analytic systems built to help users explore large amounts of data. The value-add of interaction in such cases is relatively clear; actions such as panning, zooming, and selecting subsets are indisputably essential to exploring datasets too large to reasonably fit on a single screen [21, 42]. Although there are potential costs to interaction [25, 42], in the case of visual analytic systems, these are often outweighed by benefits. Furthermore, in visual analytic systems interaction is often seen as an essential support for users' cognitive processing; it is viewed as the embodiment of sensemaking and knowledge discovery [27, 33, 34, 45]. In contrast, the value-add of interaction to static visualizations is not clear cut. Here we are specifically referring to interactions that do not reveal otherwise hidden data; i.e. they *do not add* any information to the visualization. This consideration is vital because the reasoning problems that we consider in this paper are notoriously difficult and studies suggest that adding interaction to a challenging task can result in cognitive overload [28].

There is a sampling of prior work that are relevant to the investigations in this paper. For example, Zhi et al. [46] found that adding interaction through brushing and linking to a storytelling visualization increased engagement. Theis et al. [39] compared static and interactive versions of an uncertainty data visualization and concluded based on accuracy and speed that the static visualization was preferable to its interactive counterpart. Ragan et al. [35] compared comprehension and detail recall in a pictorial learning activity across interactive versus automatic view controls, and found no significant differences between the two. Note that all of these studies include an A-B test between one static and one interactive visualization. Our work adds nuance to this body of work by investigating the value-add of interactivity with a multi-factor experimental design.

2.2 Bayesian Reasoning

An area in which Bayesian reasoning problems are prevalent is medical decision making. The standard example of a Bayesian reasoning problem in this context is the 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?*

Using conditional probabilities and Bayesian reasoning to solve this problem is difficult for patients and physicians alike [11]. Given the importance of accurate medical risk communication and understanding, numerous studies have investigated how to aid people in Bayesian reasoning. One common approach is to change the text from probability format to frequency format [17, 41]. Another approach is visualization.

Numerous studies have investigated the effect of visualization on Bayesian reasoning with a variety of different designs. Different techniques tested include Euler diagrams [4, 22, 23, 29], frequency grids or icon arrays [2, 16, 22, 23, 29, 31, 36, 41], decision trees [2, 15, 23, 36, 38], "beam cut" diagrams [17], probability curves [7], contingency tables [7, 8], double trees [2, 23, 24], flow charts [23], pipe diagrams [23], Sankey diagrams [23], and unit squares [2].

Despite all of these studies, there is still no clear consensus on the best visualization for Bayesian reasoning. Several studies compared multiple Bayesian reasoning visualizations to each other and to text [23, 29, 32]. All of these studies found that visualization did not significantly improve users' accuracy in performing a Bayesian reasoning task compared to text. Findings from Ottley et al. and Micallef et al. indicate that there may be a detrimental effect on users' ability to perform Bayesian reasoning when visualizations and numerical text descriptions are presented together [29, 32]. Ottley et al. [32] shed light on a significant performance gap between people with high and low spatial ability on a Bayesian reasoning task, and indicated that optimal visualization and text designs for people with high spatial ability differ from those for people with low spatial ability.

To the best of our knowledge, only two studies have investigated the effect of interactive Bayesian reasoning visualizations, and the results are conflicting. Tsai et al. [41] tested a frequency grid with interactive checkboxes against textual descriptions of the Bayesian reasoning problem. They found the interactive frequency grid resulted in a significant increase in accuracy compared to a textual description of the problem with statistics in probability format. Khan et al. [24] compared a static and interactive double tree diagram of the Bayesian reasoning problem. They added interaction via drag and drop, and found the interactive double tree diagram significantly decreased accuracy in performing the Bayesian reasoning task compared to static. Our goal is to provide context for these conflicting results. By identifying how static visualization design and users' spatial ability affect the value-add of interaction to a static visualization, we can shed light on confounding factors that may explain differences in prior results.

3 RESEARCH GOALS

Given conflicting prior results on how interaction impacts Bayesian reasoning visualizations, the overarching goal of this paper is to empirically test whether adding interaction to a static Bayesian reasoning visualization can improve performance on a reasoning task. We hypothesize that the mixed results of prior work are partially due to confounding factors between experiments, such as underlying visualization designs. Additionally, based on work by Ottley et al. [32] which demonstrates that spatial ability is a significant predictor of one's accuracy on Bayesian inference, we expect to see different effects of adding interaction to a Bayesian reasoning

visualization across different levels of spatial ability. Specifically, we investigate the following:

> **RQ1**: Does adding interaction to a static visualization improve accuracy on a Bayesian reasoning task?
> **RQ2**: Is the effect of interaction modulated by the effectiveness of the underlying static visualization?
> **RQ3**: Do users with different levels of spatial ability react to an interactive Bayesian reasoning visualization differently?

In the remainder of this paper we describe two crowdsourced experiments designed to begin answering these research questions. These experiments are a step towards a deeper understanding of interaction in visualization. By empirically demonstrating what specific factors lead to performance gains and losses when making a static visualization interactive, we hope to lay the ground work for better interactive visualization design with evidence-based design guidelines.

4 EXPERIMENT 1

Although interaction is commonly used in data visualization, efforts to define *interactivity* are ongoing [10]. Additionally, there is no consensus on the best visualization for Bayesian reasoning [23, 29, 32]. As a result, the design space for interactive Bayesian reasoning visualizations is large. For this experiment, we simplify the visualization design space by focusing on variants of *icon arrays* – one of the most popular and well-studied visualizations in this context [29, 30, 32]. Guided by prior work on interactive Bayesian reasoning visualizations that found positive results [41], we narrow the interaction design space to a single category, *checkboxes*. In short, this experiment examines whether adding interactive *checkboxes* to variations of static icon arrays improves accuracy in a Bayesian reasoning task.

4.1 Visualization Designs

We present users with a Bayesian reasoning problem concerning a disease in the population and the false positive and negative rates associated with testing for the disease (Section 4.2). Each stimulus is an icon array that encodes the four key sub-populations of the problem: HAVE DISEASE, DO NOT HAVE DISEASE, TEST POSITIVE, and TEST NEGATIVE. Examples of interactive and static stimuli are shown in Table 1. Below, we describe each experimental factor:

- **Base Visualization**: { *grouped, aligned, randomized* }
- **Interaction**: { *checkboxes, static* }

4.1.1 Base Visualizations. We observe three primary designs of icon arrays in the literature which are loosely based on theories for how to facilitate Bayesian reasoning. For example, some researchers propose that representing randomness can more accurately communicate the inherent uncertainly in the problem space [19]. Others hypothesize that spatially grouping visual elements aids reasoning [29]. Based on these theories, we design three variations for *icon arrays* by changing the types of contextual placements of icons (see examples in Table 3). Following guidelines of Bertin [1], background color was used to differentiate between members of the population who HAVE DISEASE versus DO NOT HAVE DISEASE, and icon color was used to differentiate between members of the population who
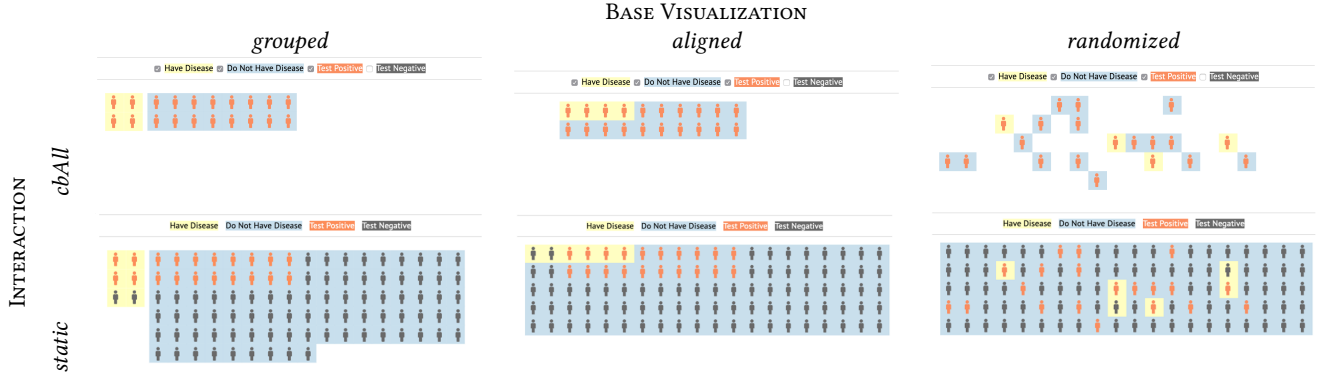
BASE VISUALIZATION

grouped   *aligned*   *randomized*

Table 1: Three interactive and static conditions used in Experiment 1. Full size images are available in supplementary materials.

TEST POSITIVE versus TEST NEGATIVE. The base visualizations differed in their use of Gestalt principles [9] to perceptually group the sub-populations in the Bayesian reasoning problem. Each base visualization design is described in detail below:

> *Grouped:* The *grouped* icon array uses spatial grouping. It shows the sub-populations HAVE DISEASE and DO NOT HAVE DISEASE in two separate grids of icons. Additionally, the TEST POSITIVE sub-population is in a block aligned at the top left of the visualization. This design is similar to the hybrid Euler-frequency grid diagram used by Micallef et al. [29].
>
> *Aligned:* The *aligned* icon array shows all icons in one 5 X 20 grid. It aligns the sub-population HAVE DISEASE in a block at the top left of the grid, and the sub-population TEST POSITIVE in a block at the top middle of the grid. A similar design was used by Brase et al. [4], and Ottley et al. [30, 32].
>
> *Randomized:* The *randomized* icon array does not spatially group any sub-populations; icons representing members of each of the four sub-populations are randomly distributed in a 5 X 20 grid. Similar designs are used in medical risk communication by Han et al. [19].

4.1.2 **Interaction**. We add *checkboxes* to traditional icon arrays which allow the user to hide or show visual elements, and create a more explicit link between the text and the graphical encodings. We chose the checkbox interaction (1) for continuity with (and replication of) prior work [41], which found positive results from adding checkboxes to a Bayesian reasoning visualization; and (2) because the checkboxes enable participants to manipulate the visualization such that it directly encodes the answer to the Bayesian reasoning question without any additional distractors i.e. users can remove all information irrelevant to answering the questions posed.

As a default, when the interactive visualizations loaded on the page all checkboxes were checked, meaning the interactive visualizations initially look identical to the static version. When using the checkboxes, one of {HAVE DISEASE, DO NOT HAVE DISEASE} as well as one of {TEST POSITIVE, TEST NEGATIVE} had to be checked for any sub-populations to show on the visualization. The top row of Table 1 shows an example of what the visualization would show if everything except TEST NEGATIVE were checked. We call this interaction technique *cbAll* (i.e. checkboxes where all boxes are initially checked).

4.2 Task

We run a between-subjects 2 {*interaction*} x 3 {*base visualization*} factor experiment. In the study, participants are asked to answer a Bayesian reasoning problem given textual and visual representations. For continuity with prior work, the textual description and question components of each stimulus were consistent with those used by Ottley et al. [32]:

> *Textual description*:
> 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 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.
> *Questions*:
> (a) How many people will test positive? _ _ _
> (b) Of those who test positive, how many will actually have the disease? _ _ _

4.3 Participants

We recruited 530 participants from Amazon Mechanical Turk. Participation was restricted to workers in the United States with an approval rating of greater than 90 percent. Participants were paid a base rate of $1.80 for participation plus a bonus of $0.10 for every correct answer.

Before analysis, participants who skipped entire sections of the experiment or did not follow instructions ($N = 3$), and participants who self-identified as colorblind ($N = 55$) were dropped from the data set. This left $N = 472$ participants distributed among stimuli as shown in Table 2. Demographics of participants are shown in Table 3.

4.4 Procedure

The experiment followed an approved protocol per *redacted for anonymity*'s IRB, and was posted as a HIT on Amazon Mechanical Turk. Workers who accepted the HIT followed a link to the experiment. After providing informed consent, participants were taken to an instruction page explaining the experiment. This page demonstrated what the legend for a *static* visualization would look like versus the *cbAll* legend. After the instruction page, participants were shown one of the six experimental stimuli. Participants could take as much time with the stimulus as they wanted before clicking a button to view the questions to answer. After completing the main task, participants were asked to complete a short demographic questionnaire, the paper folding test (VZ-2) from Ekstrom, French, & Hardon [12] to measure spatial ability[1], and to provide any additional feedback they wished.

	grouped	aligned	randomized	Total
cbAll	64	100	82	246
static	86	70	70	226
Total	150	170	152	472

Table 2: Sample size (N) for each condition in Experiment 1.

N	472
Age	18-24: 6.4%, 25-39: 64.4%, 40-49: 17.2%, 50-59: 7.6%, 60+: 4.4%
Gender	Female: 38.3%, Male: 61.0%, Non-Binary: 0.8%
Education	High School: 28.6%, Bachelors: 56.6%, Masters: 10.2%, PhD: 1.5%, Other: 3.2%
Expertise with Statistical Visualization	Novice: 21.4%, Low-intermediate: 20.1%, Intermediate: 32.3%, High-intermediate: 17.4%, Expert: 8.5%
Statistical Training 1(none) - 5 (highly trained)	1: 32.4%, 2: 22.5%, 3: 16.7%, 4: 15.3%, 5: 12.5%

Table 3: Experiment 1 participant demographics.

4.5 Research Questions

We analyze data from Experiment 1 to answer these questions:

Q1.1 Does adding interaction to a static visualization improve accuracy on Bayesian reasoning task? We investigate whether participants who use an interactive visualization will be more accurate in answering the Bayesian reasoning question than participants who saw a static visualization.

Q1.2 Is the effect of interaction modulated by the underlying static visualization design? We investigate whether differences in accuracy will be modulated by the base visualization design (*grouped, aligned, randomized*).

---

[1]Due to space constraints, spatial ability analysis for Experiment 1 is included in supplemental materials.

4.6 Findings

For analysis, participants' answers were considered correct only if they answered both parts of the two-part question correctly (this approach is consistent with prior work [30, 32]). Our analysis script is included in supplemental materials, however under the guidelines of our IRB we are unable to release the data.

4.6.1 Does adding interaction to a static visualization improve accuracy on Bayesian reasoning task? Figure 1 shows proportions of participants answering the Bayesian reasoning question correctly in the *cbAll* and *static* conditions. We observe 53% of the participants in the *cbAll* condition entered the correct answers, whereas the *static* visualization had a 57% correct response rate. We perform a 2-sample test for equality of proportions of *accuracy ~ interactive_or_static* with the null hypothesis that there is no difference in proportions of correct answers. We find no statistically significant difference in accuracy between participants using the *cbAll* versus the *static* visualization ($\chi^2(1, N = 472) = 0.69, p = 0.41$), and therefore fail to reject the null hypothesis. *Ultimately, we found no evidence that adding interaction improves accuracy on a Bayesian reasoning task.*
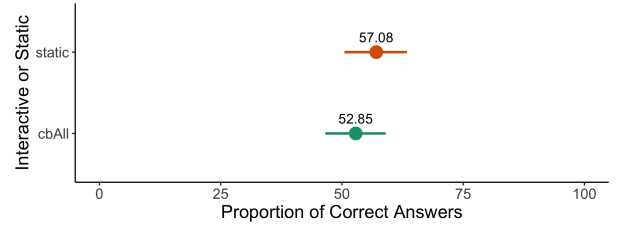

Figure 1: Proportion of participants answering the Bayesian reasoning task correctly given an interactive (*cbAll*) versus *static* visualization. Bars represent a 95% logit transformed confidence interval.
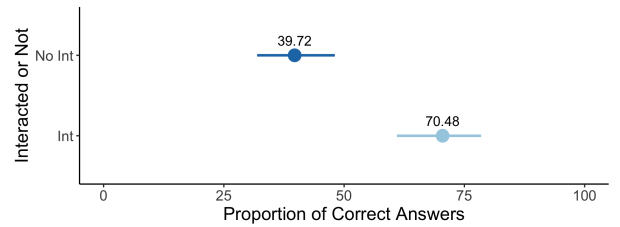

Figure 2: Proportion of participants assigned to *cbAll* answering the Bayesian reasoning task correctly grouped by if they interacted or not. Bars represent a 95% logit transformed confidence interval.

Next, we look at how many participants in the *cbAll* condition interacted with the visualization they saw. Out of the 246 participants assigned to *cbAll*, only 43% used the checkboxes on the visualization. Seventy percent of the participants who interacted with the visualization answered correctly, while only 40% of those who did not interact answered correctly, as shown in Figure 2. We perform a 2-sample test for equality of proportions of

$accuracy \sim interacted\_or\_not$ with the null hypothesis that there is no difference in proportions of correct answers. We find a statistically significant difference in accuracy between participants who did and did not interact with the *cbAll* visualization ($\chi^2(1, N = 246) = 21.63, p < 0.001$), and therefore reject the null hypothesis. The 95% confidence interval (using Wilson's score method) for the difference between the proportions is $[−43.48, −18.04]$. Taken together with the previous finding, this suggests that simply adding interaction does not guarantee use. However, *the act of interacting may improve performance on a Bayesian reasoning task.* We note future work is needed to isolate if this is an artifact of participant engagement, or an effect of interacting.
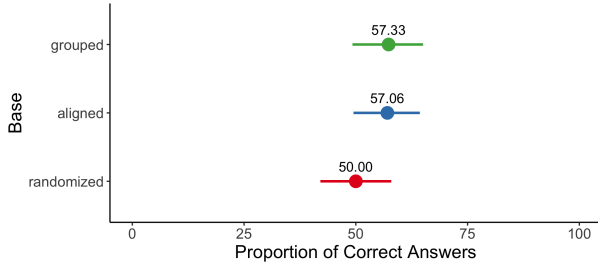

Figure 3: Proportion of participants answering the Bayesian reasoning task correctly by base visualization. Bars represent a 95% logit transformed confidence interval.
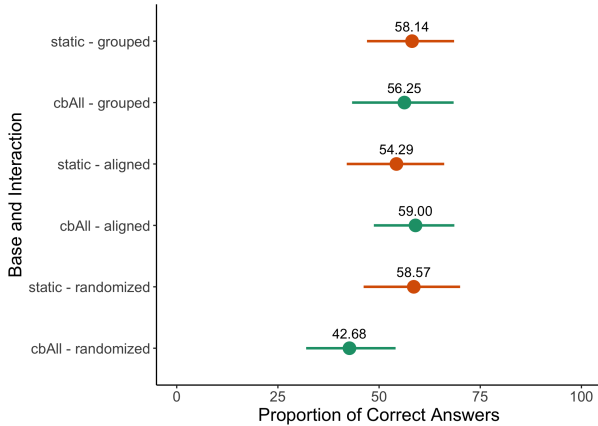

Figure 4: Proportion of participants answering the Bayesian reasoning task correctly by interaction and base visualization. Bars represent a 95% logit transformed confidence interval.

4.6.2 ***Is the effect of interaction modulated by the underlying static visualization design?*** To investigate whether base visualization design has an effect on participants' accuracy we perform a 3-sample test for equality of proportions of $accuracy \sim base\_visualization$ with the null hypothesis that there is no difference in proportions of correct answers. We find no significant difference in accuracy by base visualization ($\chi^2(2, N = 472) = 2.15, p = 0.34$), and therefore fail to reject the null hypothesis. As

shown in Figure 3, we observe near-equal proportions of correct answers for the three base representations, suggesting that the *variations in the design of icon arrays had no significant effect on accuracy.*

To investigate if there is an interaction effect between base visualization and whether a visualization is interactive or static we perform a 6-sample test for equality of proportions of $accuracy \sim \{base\_visualization\}X\{interactive\_or\_static\}$ with the null hypothesis that there are no differences in proportions of correct answers. We find no significant difference ($\chi^2(5, N = 472) = 6.42, p = 0.27$), and therefore fail to reject the null hypothesis. This suggests that *the value-add of interaction may not be modulated by base visualization design.*

Notably, Figure 4 shows proportions of correct answers are nearly identical given all combinations of base visualization and interactive or static, but that the proportion of correct answers is smaller on average for participants assigned the *cbAll* version of the *randomized* visualization. While this difference is not statistically significant, we note that practically speaking it is important to consider; particularly in the case of medical risk decision making, where small improvements on this task can lead to more informed and autonomous medical decisions.

4.7 Discussion

Although it is a common belief that interactivity adds value to visualizations, investigations into it merits can reveal essential insights about the pros, cons, or missed opportunities in interactive visualization design. In this experiment, we used Bayesian reasoning – a problem that is notoriously challenging for the general population – and showed that adding interactive checkboxes to a Bayesian reasoning visualization does not significantly improve reasoning accuracy. Moreover, in this experiment we observe cases where adding interaction decreases average performance on the Bayesian reasoning. Though these differences are not statistically significant, the lack of a clear value add suggests future work should continue to investigate the potential costs and benefits of interaction in this setting.

Our analyses suggest that there may not be a value-add to making a static visualization interactive. Moreover, our observational findings suggest that the effect of adding interactivity to a static visualization may depend on the design of the visualization itself. We used three variations of icon arrays based on theories for how to facilitate Bayesian reasoning: *grouped*, *aligned*, and *randomized*. We observed nearly identical accuracy between the interactive and static versions of the *grouped* and *aligned* designs, and an insignificant, but practically relevant, decrease in accuracy for the interactive version of the *randomized* design. We speculate that one rationale for this outcome is that the combination of a challenging Bayesian problem with randomness and interactivity may have induced an extraneously high cognitive load. Some experts caution that adding interactivity to a significantly complicated task can result in cognitive overload [28]. Due to the lack of perceptual grouping, the *randomized* base visualization induces more cognitive load than the *grouped* and *aligned* bases. Based on this and our observational findings, we postulate that adding interaction

to the *randomized* base may have caused cognitive overload in participants.

An important observation is that a sizable portion of our study population assigned to the interactive visualization did not interact with it (57%). There are a combination of factors that can explain this result. No interaction could be indicative of participants who were either confused about the task or were simply clicking through to get paid, as discussed in Section 4.6.1. Alternatively, it is possible that participants did not want to interact. Existing work indicates that people may not engage with interactive visualizations as much as previously thought [3] and there have been reports of media venues such as *The New York Times*, scaling back their creation of interactive visualization in lieu of static images[2]. While understanding if there is a value-add of interaction is an important step to user-centered interactive visualization design, we recognize that understanding users' perceived value of interaction is also crucial.

5 EXPERIMENT 2

The findings of Experiment 1 suggest that adding a *checkbox* interaction to a static Bayesian reasoning visualization has little to no effect on reasoning accuracy. Experiment 2 expands on this study by exploring the effect of different interaction techniques. Specifically, we compare the effects of adding *two types of checkboxes*, *drag and drop*, *hover*, and *tooltips* to the three *icon array* base visualizations used in Experiment 1.

5.1 Visualization Designs

Consistent with Experiment 1, each stimulus is an icon array that encodes the four key sub-populations in the Bayesian reasoning problem. Examples of stimuli are shown in Table 4. Below, we describe each experimental factor:

- **Base Visualization**: {*grouped, aligned, randomized*}
- **Interaction Technique**: {*checkboxes, drag and drop, hover, tooltips*}

5.1.1 **Base Visualizations**. We use the same three *icon array* designs as Experiment 1 (*grouped, aligned, randomized*).

5.1.2 **Interaction Techniques**. We test five different interaction techniques. We chose three of them (two types of checkboxes, and drag and drop) for consistency with prior work [24, 41], and two additional techniques representative of the well known "*overview first, zoom and filter, details on demand*" mantra of visualization design [37]. We chose hover as an interaction representative of *zoom and filter*, and tooltips as a representative of *details on demand*.

We recognize that even within the constraints of these five interaction techniques there are countless ways to design an interactive Bayesian reasoning visualization. Our goal in designing each interaction (similar to Experiment 1) was to enable participants to remove information irrelevant to answering the Bayesian reasoning question from the visualization, or to allow participants to more easily integrate information shown in the visualization and text. Below, we describe the implementation and design goals of each interaction technique:

---

[2]Why We Are Doing Fewer Interactives (Archie Tse, The New York Times): https://github.com/archietse/malofiej-2016/blob/master/tse-malofiej-2016-slides.pdf

Checkbox All: We included checkboxes in Experiment 2 for continuity with Experiment 1, and consistency with prior work [41]. We modify the interaction from Experiment 1 such that any sub-population "checked" on the legend is shown on the visualization with color, and any sub-population "unchecked" on the legend is shown on the visualization with light grey placeholders. This enabled participants to check off only populations explicitly mentioned in the Bayesian reasoning question (HAVE DISEASE and TEST POSITIVE) in order to answer it. In contrast, the implementation of checkboxes in Experiment 1 required participants to check off HAVE DISEASE, DO NOT HAVE DISEASE, and TEST POSITIVE in order to answer the Bayesian reasoning question.

Checkbox None: Checkbox None (*cbNone*) is identical to *cbAll*, except all checkboxes are unchecked by default. In other words, the page loads with only light grey placeholders shown on the visualization.

Drag and Drop: Drag and drop (*drag*) is a direct manipulation interaction. It was chosen to be consistent with interaction tested in prior work [24]. We designed *drag* with the intent of providing the same direct encoding benefits as checkboxes. In practice, *drag* functions identically to *cbNone* except that participants drag legend labels onto and off of the visualization.

Hover: Hover (*hover*) is a filter interaction. Prior studies suggest that users struggle to integrate text and visualization when performing Bayesian reasoning [29, 30, 32]. We design *hover* with intent to help users overcome this hurdle by drawing a clearer connection between text and visualization. As participants hover their mouse over areas of text describing sub-populations in the visualization, the text and corresponding sub-population are highlighted.

Tooltip: Tooltips (*tooltip*) are an example of details on demand. Similar to *hover*, we design this interaction with the intent to help users integrate text and visualization more easily. To facilitate this, we reduce the distance users need to move their eyes to integrate the text and visualization by directly overlaying the two. When a participant hovers their mouse over any icon in the visualization a text box appears describing to which of the four sub-populations that particular icon belongs.

It is relevant to note that although these techniques differ, all interaction techniques provide the same amount of information as the static visualization condition. The use of interactions does not add more information. Since Bayesian reasoning is the manipulation of 4 basic values (true positive, true negative, false positive, and false negative), the goal of the interaction techniques is to help the user in two ways: (1) isolating values of interest in the visualization, and (2) drawing connections between the visual representation and the textual description of the problem.

5.2 Task

We run a between-subjects 5 {*interaction techniques*} x 3 {*base visualizations*} factor experiment. The same textual description and
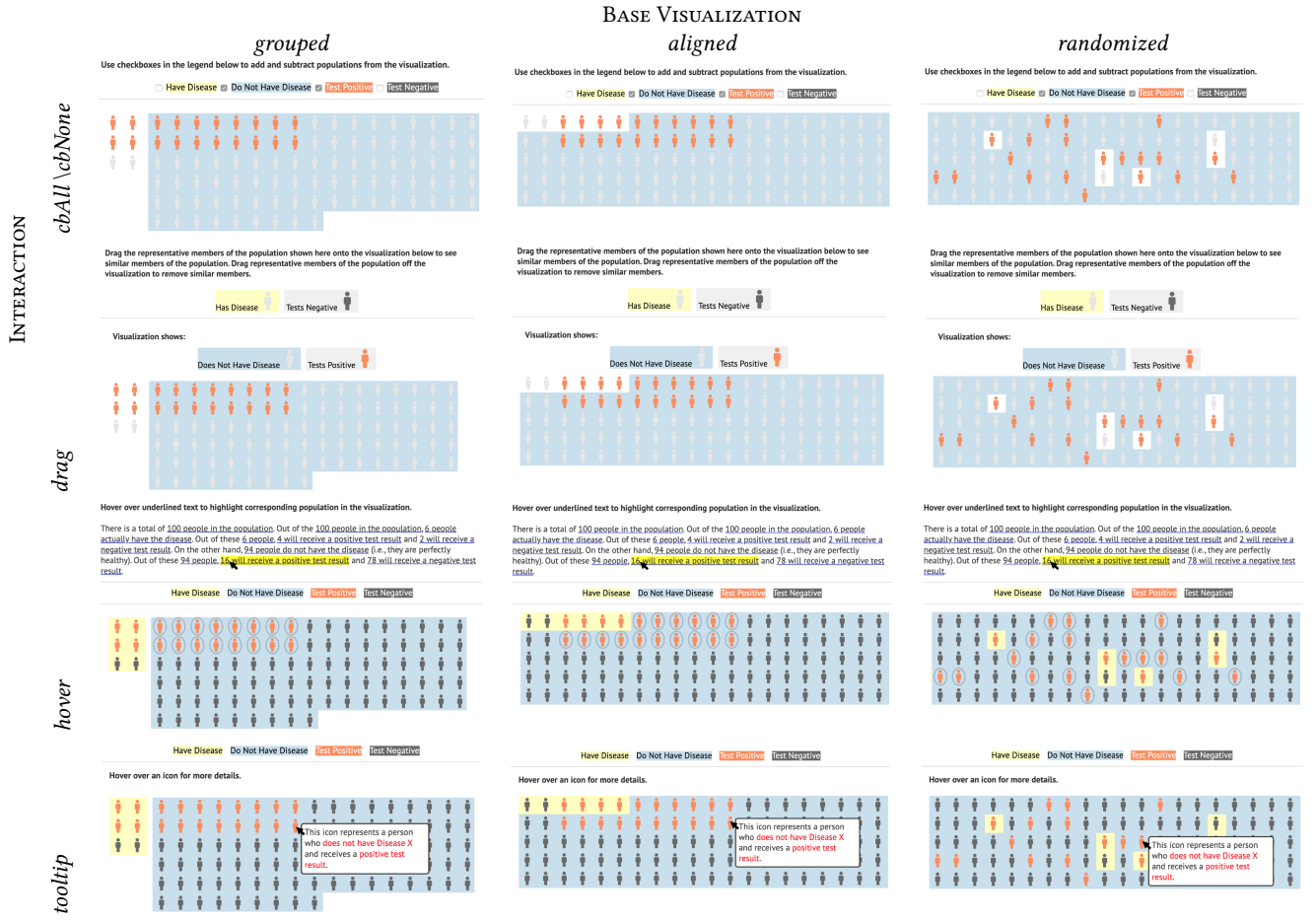
Table 4: Five interactive conditions used in Experiment 2. Full size images are available in supplementary materials.

questions as in Experiment 1 (Section 4.2) are used for this experiment.

	grouped	aligned	randomized	Total
cbAll	119	129	123	371
cbNone	147	125	122	394
drag	102	115	129	346
hover	131	136	167	434
tooltip	147	154	134	435
Total	646	659	675	1980

Table 5: Sample size (N) for each condition in Experiment 2.

N	1,980
Age	18-24: 7.8%, 25-39: 50.8%, 40-49: 20.4%, 50-59: 12.5%, 60+: 8.2%
Gender	Female: 54.1%, Male: 45.1%, Non-Binary: 0.5%
Education	High School: 27.1%, Bachelors: 49.8%, Masters: 15.1%, PhD: 1.5%, Other: 6.0%
Expertise with Statistical Visualization	Novice: 15.2%, Low-intermediate: 21.4%, Intermediate: 38.8%, High-intermediate: 18.5%, Expert: 5.4%
Statistical Training 1(none) - 5 (highly trained)	1: 29.6%, 2: 22.3%, 3: 21.8%, 4: 16.9%, 5: 8.0%

Table 6: Experiment 2 participant demographics.

5.3 Participants

We recruited 2,149 participants from Amazon Mechanical Turk. Participation was restricted to workers in the United States with an approval rating greater than 90 percent. Participants were paid a base rate of $0.80 for participation, plus a bonus of $0.10 for every correct answer.

Participants who skipped entire sections of the experiment or did not follow instructions ($N = 35$), and participants who self-identified as colorblind ($N = 134$) were dropped from the data set. This left $N = 1,980$ participants distributed among stimuli as shown in Table 5. Demographics of participants are shown in Table 6.

5.4 Procedure

Experiment 2 follows the same procedure as Experiment 1 with two exceptions. First, there is no instruction page, instead instructions are provided alongside each stimulus. And second, participants are asked to complete a an additional NASA-TLX [20] survey after the completion of the main study to measure task difficulty.

5.5 Research Questions

We analyze data from Experiment 2 to answer these questions:

> **Q2.1 Do different interaction techniques have different effects on accuracy in Bayesian reasoning?** We test whether participants who saw any interactive visualization will be more accurate in answering the Bayesian reasoning question than participants who saw a static visualization.
> **Q2.2 Is the effect of interaction moderated by interaction design and the underlying static visualization design?** We test if differences in accuracy will be modulated by base (*grouped, aligned, randomized*) visualization design, and interaction technique (*cbAll, cbNone, drag, hover, tooltip*).
> **Q2.3 How does the effect of different interaction techniques change given different spatial abilities?** We split participants into high and low spatial ability groups and repeat the analyses for **Q2.1** within each group.
> **Q2.4 Does underlying static visualization design moderate the effect of interaction techniques within spatial ability groups?** We split participants into high and low spatial ability groups and repeat the analyses for **Q2.2** within each group.

5.6 Findings

As in Experiment 1, participants' answers were considered correct only if they answered both parts of the two-part question correctly. In order to compare our interactive visualizations against a static one, we included the *static* group from Experiment 1 as an "interaction technique" in our analysis (this brings the total number of participants to $N = 2,206$). From here forward the term "interaction techniques" refers to *cbAll, cbNone, drag, hover, tooltip*, and *static*. Our analysis script is included in supplemental materials, however under the guidelines of our IRB we cannot release data.

5.6.1 *Do different interaction techniques have different effects on accuracy in Bayesian reasoning?* We perform a 6-sample test for equality of proportions of $accuracy \sim interaction\_techniques$ with the null hypothesis that there is no difference in proportions of correct answers. We find no statistically significant difference in accuracy between participants using different interaction techniques ($\chi^2(5, N = 2206) = 11.33, p = 0.05$), and therefore fail to reject the null hypothesis. As shown in Figure 5, we observe similar proportions of correct answers across interactions. This suggests that *the design of an interactive technique does not affect the value-add of interaction.*

Similar to Experiment 1, out of the 371 participants assigned to *cbAll*, only 44% actually interacted. This proportion was significantly higher for all other interactive conditions. For each interaction technique we perform a 2-sample test for equality of proportions of $accuracy \sim interacted\_or\_not$ with the null hypothesis
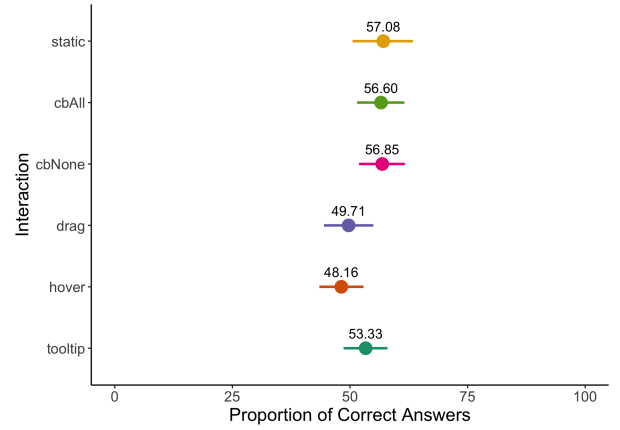

Figure 5: Portion of participants answering the Bayesian reasoning task correctly given each interaction technique. Bars represent a 95% logit transformed confidence interval.
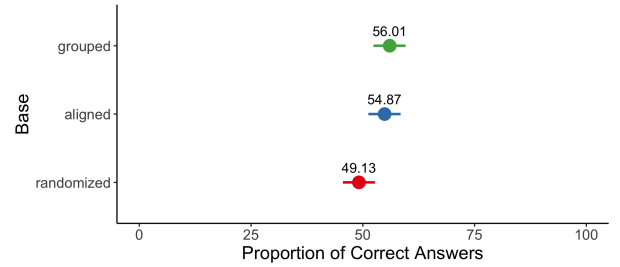

Figure 6: Portion of participants answering the Bayesian reasoning task correctly by base visualization. Bars represent a 95% logit transformed confidence interval

that there is no difference in proportions of correct answers. Interestingly *drag* is the only condition in which participants who interacted were significantly more accurate than those who did not ($\chi^2(1, N = 346) = 18.49, p < 0.001$), and therefore is the only case in which we reject the null hypothesis. The 95% confidence interval (using Wilson's score method) for the difference between proportions is [-38.69, -14.93]. Further analysis is included in supplemental materials.

5.6.2 *Is the effect of interaction moderated by interaction design and the underlying static visualization design?* We compare participants' accuracy across base visualizations with a 3-sample test for equality of proportions of $accuracy \sim base\_visualization$ with the null hypothesis that there are no differences in proportions of correct answers. We find a significant difference in accuracy by base ($\chi^2(2, N = 2206) = 8.09, p = 0.02$), and therefore reject the null hypothesis. Pairwise comparisons with a Bonferroni corrected alpha (0.02) show a significant difference between the *grouped* and *randomized* bases ($\chi^2(1, N = 1477) = 6.74, p < 0.01$). The 95% confidence interval (using Wilson's score method) for the difference between these two proportions is $[-12.10, -1.67]$. As shown in Figure 6, the proportion of correct answers in the *grouped* base is higher than that for the *randomized* base. This suggests that *design*

of the underlying static visualization has an effect on participants' accuracy in a Bayesian reasoning task.
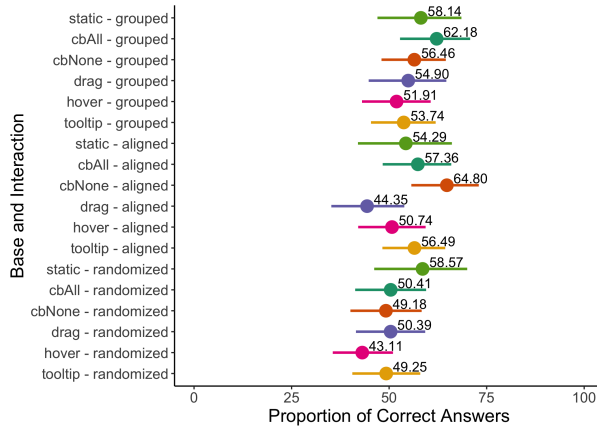

Figure 7: Portion of participants answering the Bayesian reasoning task correctly given each interaction technique and base visualization. Bars represent 95% logit transformed confidence intervals.

To investigate whether there is an interaction effect between base visualization and interaction technique we perform an 18-sample test for equality of proportions of $accuracy \sim \{base\_visualization\} X \{interaction\_technique\}$ with the null hypothesis that there are no differences in proportions of correct answers. We find no significant difference ($\chi^2(17, N = 2206) = 27.92, p = 0.05$) and therefore fail to reject the null hypothesis.

Figure 7 shows proportions of participants answering the Bayesian reasoning task correctly given each combination of base visualization design and interaction technique. Similar to Experiment 1, we observe that any interactive technique added to the *randomized* base resulted in lower accuracy on average than the static version of *randomized*. This is not the case for the *grouped* and *aligned* bases. Practically speaking, this suggests that *the combination of the randomized base design with interaction may reduce average accuracy on the Bayesian reasoning task.*

5.6.3 *How does the effect of different interaction techniques change given different spatial abilities?* Prior work demonstrates that spatial ability is a significant predictor of performance on a Bayesian reasoning task and that people with low spatial ability in particular struggle to perform Bayesian inference [32]. Based on this prior finding, we assume a difference in accuracy for participants with high versus low spatial ability and re-run our previous analyses stratified by spatial ability. We assess how people with high and low spatial ability react to the interaction techniques and base visualization designs we test.

Participants' spatial ability scores ranged from −5 to 20. The median score was 6.25. Consistent with prior work [32], we assigned participants with spatial ability scores greater than or equal to the median to the high spatial ability group, and participants with scores less than the median to the low spatial ability group. Figures 8 and 9 plot performance on the Bayesian reasoning task by spatial ability group (high or low). In both of these figures we observe

no overlaps in 95% confidence intervals for proportions of correct answers of participants with high versus low spatial ability. This supports our assumption that there is a difference in accuracy for people with high versus low spatial ability. The following sections report analyses stratified by spatial ability group.

First, we perform a 6-sample test for equality of proportions of $accuracy \sim interaction\_technique$ with the null hypothesis that there are no differences in proportions of correct answers, stratified by spatial ability group.

Within the high spatial ability group we find a statistically significant difference in accuracy between participants using different interaction techniques ($\chi^2(5, N = 1114) = 17.76, p < 0.005$), and therefore reject the null hypothesis. Pairwise tests with a Bonferroni corrected alpha (0.003) show a significant difference between the *hover* and *static* interactions ($\chi^2(1, N = 332) = 12.44, p < 0.001$). The 95% confidence interval (using Wilson's score method) for the difference between these two proportions is [9.43, 30.37]. A larger proportion of participants with high spatial ability answered the Bayesian reasoning task correctly using the *static* visualization versus the *hover* visualization (Figure 8).

Within the low spatial ability group we find no statistically significant difference in accuracy between participants using different interaction techniques ($\chi^2(5, N = 1092) = 5.72, p = 0.33$), and therefore fail to reject the null hypothesis.

These results suggest that *for people with high spatial ability, an interactive hover visualization can significantly decrease accuracy on Bayesian inference compared to a static visualization.* And that *for people with low spatial ability, adding interaction to a static visualization does not significantly effect reasoning accuracy.*
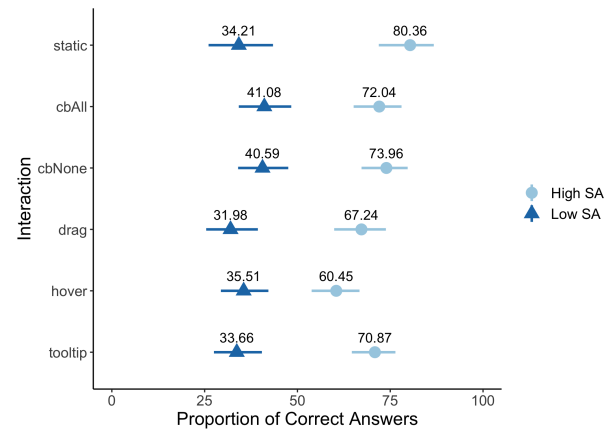

Figure 8: Portion of participants answering the Bayesian reasoning task correctly by spatial ability (SA) and interaction technique. Bars represent 95% logit transformed confidence intervals.

5.6.4 *Does underlying static visualization design moderate the effect of interaction techniques within spatial ability groups?* Within each spatial ability group we compare participants' accuracy across bases by performing a 3-sample test for equality of proportions of $accuracy \sim base\_visualization$, with the null hypothesis that there are no differences in the proportions of correct answers.
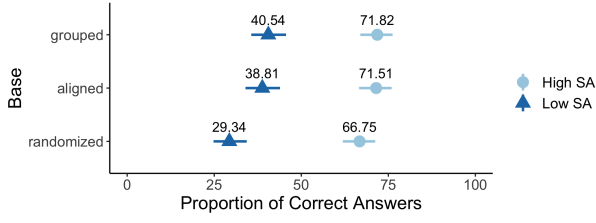
Figure 9: Proportion of participants answering the Bayesian reasoning task correctly by base visualization for each spatial ability group. Bars represent 95% logit transformed confidence intervals.

Within the high spatial ability group we find no significant difference in accuracy by base ($\chi^2(2, N = 1114) = 2.93, p = 0.23$), and therefore fail to reject the null hypothesis. Figure 9 shows near equal proportions of participants with high spatial ability answering the Bayesian reasoning questions correctly across base visualizations.

Within the low spatial ability group we find a significant difference in accuracy by base ($\chi^2(2, N = 1092) = 11.23, p < 0.005$), and therefore reject the null hypothesis. Pairwise tests with a Bonferroni corrected alpha (0.02) show participants with low spatial ability assigned the *randomized* base performed significantly worse than those assigned the *grouped* ($\chi^2(1, N = 721) = 9.43, p < 0.005$) and *aligned* ($\chi^2(1, N = 722) = 6.77, p < 0.01$) bases. The 95% confidence intervals (using Wilson's score method) for the differences in these proportions are $[-18.38, -4.01]$ and $[-16.62, -2.32]$, respectively (Figure 9). This suggests that *for people with high spatial ability, performance on a Bayesian reasoning task is not affected by design of the underlying static visualization, and for people with low spatial ability it is.*

Finally, within each spatial ability group we check for an interaction effect between base visualization design and interaction technique. To do this we perform an 18-sample test for equality of proportions of $accuracy \sim \{base\_visualization\}X\{interaction\_technique\}$ with the null hypothesis that there are no differences in proportions of correct answers.

Within the high spatial ability group we find no significant differences ($\chi^2(17, N = 1114) = 23.64, p = 0.13$), and therefore fail to reject the null hypothesis.

Within the low spatial ability group we again find no significant differences ($\chi^2(17, N = 1092) = 23.41, p = 0.14$), and therefore fail to reject the null hypothesis. As shown in Figures 10 and 10 we observe near equal proportions of correct answers in all combinations of base visualizations and interaction techniques. This suggests that *for people with high and low spatial ability the value-add of interaction is not modulated by underlying static visualization design and interaction technique.*

6 DISCUSSION

Three of the interaction techniques tested in Experiment 2 are similar to those tested in prior work (*cbAll, cbNone, drag*) [24, 41]. Our findings from both experiments suggest that neither *checkbox* interaction significantly improves performance on a Bayesian reasoning task compared to a *static* visualization. While prior work has suggested that an interactive checkbox visualization can increase
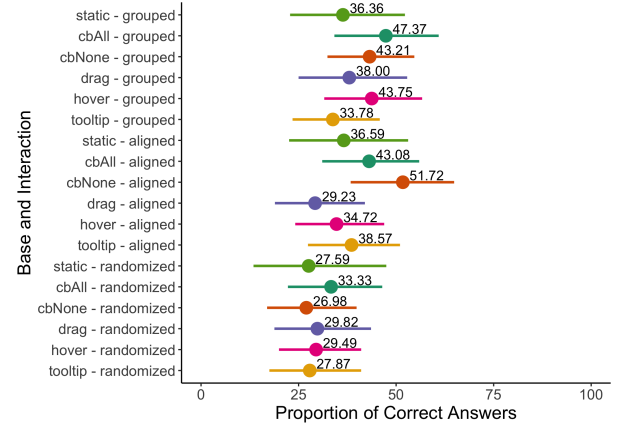

Figure 10: Portion of participants with low spatial ability answering the Bayesian reasoning task correctly given each interaction technique and base visualization. Bars represent a 95% logit transformed confidence intervals.
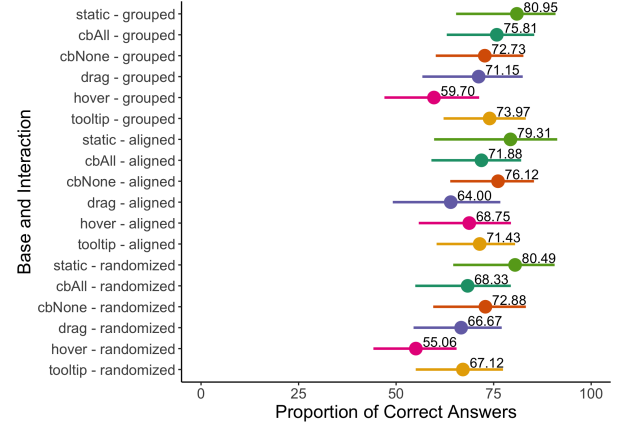

Figure 11: Portion of participants with high spatial ability answering the Bayesian reasoning task correctly given each interaction technique and base visualization. Bars represent a 95% logit transformed confidence intervals.

accuracy on a Bayesian reasoning task [41], that work compared an interactive visualization coupled with a textual description of the problem using frequencies to a textual description of the problem using probabilities. Our findings add nuance to this work by explicitly comparing an interactive and static visualization, and by using a constant wording of the text. These differences in experimental set up likely explain the discrepancy between findings in prior work and our findings. Similarly, we find that the *drag* interaction does not significantly change performance on the Bayesian reasoning task compared to a *static* visualization. These findings are consistent with prior work [24], which found adding a drag and drop interaction to a double tree diagram decreased performance on a Bayesian reasoning task, but not significantly so.

To broaden the scope of our results, we include interactions representative of the popular "*overview first, zoom and filer, details on demand*" mantra of visualization design [37] (*hover, tooltip*). Again

we find neither of these interaction techniques lead to significantly better performance on Bayesian reasoning than a *static* visualization.

Moreover, we find participants with high spatial ability have significantly worse accuracy given the interactive *hover* visualization versus *static*. We postulate that this is another case where adding interaction to an already complex task causes cognitive overload. Ottley et al. [32] showed that integrating text and visualization is especially difficult for people with high spatial ability. As *hover* draws a very explicit link between text and visualization, it is likely an extremely cognitively taxing interaction for high spatial ability participants and thus led to cognitive overload and decreased performance. We encourage future work to further investigate the nuanced relationship between cognitive load, interaction, and spatial ability.

The results of Experiment 2 coupled with those of Experiment 1 suggest a perhaps unexpected answer to the question "does interaction improve Bayesian reasoning with visualization?" Despite our best efforts, none of the interaction techniques we test significantly improve participants' performance on a Bayesian reasoning task compared to a static visualization. Moreover, we identify several scenarios that suggest interaction *decreases* participants' performance. In Experiment 1, we observe adding checkboxes where all the boxes are pre-checked (*cbAll*) to the *randomized* base visualization decreases performance of participants (Figure 4). This decrease is not statistically significant, but has practical implications. Similarly, in Experiment 2 we observe statistically insignificant decreases in accuracy across all interaction techniques added to the *randomized* design (Figure 7). In addition, in Experiment 2 we see that across visualization designs participants with high spatial ability perform significantly worse with an interactive *hover* visualization than with a *static* visualization. While all of these findings are not statistically significant, together they suggest that interaction may not be universally beneficial to static Bayesian reasoning visualizations, and that this topic warrants further investigation.

Our findings suggest that a well-designed static visualization can be as (if not more) beneficial to solving complex reasoning tasks as an interactive visualization. In cases where the user's interactions result in additional information being shown on the visualization (e.g. panning a map), the value of the interaction is undisputed. However, general claims that interaction can improve reasoning, and offer cognitive support can be called into question given the results of our experiments. We observe more than one scenario in which the use of an interactive visualization can be detrimental, and we were unable to show any cases in which use of an interactive visualization led to significant improvement in Bayesian inference. Therefore, we echo the sentiment made by researchers such as Lam [25] and van Wijk [42], and practitioners like the New York Times, and suggest a cautious use of interactivity in cases where it does not add additional information to a visualization.

7 LIMITATIONS AND FUTURE WORK

We acknowledge there are limitations to our work and that there remain open questions for future research.

Studies such as the ones presented in this paper are fundamentally limited in scope. While we made our best effort to select the most appropriate visualization designs and implement the best interactions based on existing literature, we acknowledge that there are infinite options that could be tested. It is plausible that there exists a combination of visualization and interaction techniques that can improve participants' abilities to solve the Bayesian reasoning task. However, our recommendation of a cautious use of interactivity as a reasoning aid remains true. In everyday designs of interactive visualizations, practitioners are unlikely to be able to carefully evaluate a large number of combinations of visualization and interaction designs. Our results show that, in those cases, practitioners should be cautious in adding interactions to a static visualization designed to help users perform reasoning tasks.

It is important to note that the interaction techniques used in our studies do not add new information to the static visualizations. However, doing so would not aid in Bayesian reasoning. Adding new information is not meaningful because the Bayesian reasoning is inherently based on understanding relationships between four integer values (true positive, false positive, true negative, and false negative counts). Moreover showing the user the numerical answer to the Bayesian reasoning problem does not help them better understand the reasoning process behind that number. In practice, that reasoning process, more so than a number value, is critical to decision making (e.g. in a medical decision-making scenario [19, 40]).

In addition, we acknowledge that there are many different formulations of the Bayesian reasoning problem with different levels of sensitivity, specificity and disease prevalence and that perturbing these values could result in different findings than what we have presented here. Moreover, there are numerous static visualization designs with which we could have performed this study. However, we see this work as a starting point to a principled investigation of the costs and benefits of interaction. There are countless factors that could be manipulated and tested. In performing this study we chose to bound the static visualizations tested to one category (icon arrays) as well as the formulation of the Bayesian reasoning problem. Both of these choices were made in an effort to keep as much consistency as possible with prior work in this space. Icon arrays are one of the most popular and well studied visualizations in this context [4, 29, 30, 32], and the specific formulation of the Bayesian reasoning problem used in this work has been used in a number of other studies [23, 29, 32, 41]. In future work we plan to explore the effects of different formulations of the Bayesian reasoning problem, as well as different static visualization designs.

Finally, our results suggest interaction may impede performance on high cognitive load tasks. However, to the best of our knowledge, beyond theoretical guidelines (such as the one by Lam [25]), there is no empirical work on evaluating interaction techniques based on their effect on a user's cognitive load. As a future work, we aim to investigate interaction techniques using a cognitive-load theory. The result of which we hope will provide a theoretical understanding on the outcome of the studies presented in this paper.

8 CONCLUSION

This paper empirically shows how different interaction techniques and visualization designs affect users in solving Bayesian reasoning tasks. Through two crowdsourced studies, we evaluated five interaction techniques across three different static visualization designs. The results illustrate that the effect of interaction is largely dependent on the design of the underlying static visualization, and implementation of the interaction itself. Additionally, we observe that people with different spatial abilities react to interaction differently. These findings suggest that adding interaction to a static Bayesian reasoning visualization may not be beneficial, and in some cases can be detrimental. For example, we find adding interaction to certain designs of static Bayesian reasoning visualizations can decrease users' accuracy on Bayesian inference. Similarly, we find when people with high spatial ability use a hover Bayesian reasoning visualization, they perform a Bayesian reasoning task with significantly worse accuracy than they do with a static visualization. Based on these findings we conclude that interaction may not be as unanimously beneficial as it is often believed to be; in some cases a well designed static visualization can be as, if not more, effective.

ACKNOWLEDGMENTS

REFERENCES

[1] Peter Babington. 1967. *Semiology of graphics: diagrams, networks, maps.* University of Wisconsin Press, Madison, WI, USA. Originally in French: Semiologie Graphique.

[2] Katharina Böcherer-Linder and Andreas Eichler. 2019. How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. *Frontiers in Psychology* 10 (2019), 267. https://doi.org/10.3389/fpsyg.2019.00267

[3] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. 2015. Storytelling in Information Visualizations: Does It Engage Users to Explore Data?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1449–1458. https://doi.org/10.1145/2702123.2702452

[4] Gary L. Brase. 2009. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology* 23, 3 (2009), 369–381. https://doi.org/10.1002/acp.1460 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.1460

[5] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. 2012. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Seattle, WA, 83–92. https://doi.org/10.1109/VAST.2012.6400486

[6] S. Carpendale, N. Riche, P. Isenberg, and B. Lee. 2012. Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions. *IEEE Transactions on Visualization & Computer Graphics* 18, 12 (dec 2012), 2689–2698. https://doi.org/10.1109/TVCG.2012.204

[7] W. G. Cole. 1989. Understanding Bayesian Reasoning via Graphical Displays. *SIGCHI Bull.* 20, SI (March 1989), 381–386. https://doi.org/10.1145/67450.67522

[8] William G Cole and Janet E Davidson. 1989. Graphic Representation Can Lead To Fast and Accurate Bayesian Reasoning. , 227–231 pages.

[9] S. Coren and J.S. Girgus. 1980. Principles of perceptual organization and spatial distortion: The gestalt illusions. *Journal of Experimental Psychology: Human Perception and Performance* 6(3) (1980), 404 – 412. https://doi.org/10.1037/0096-1523.6.3.404

[10] E. Dimara and C. Perin. 2020. What is Interaction for Data Visualization? *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan 2020), 119–129. https://doi.org/10.1109/TVCG.2019.2934283

[11] David M Eddy. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities.

[12] R.B. Ekstrom, J.W. French, H. H. Harman, and D. Dermen. 1976. Manual for kit of factor-referenced cognitive tests.

[13] A. Endert, P. Fiaux, and C. North. 2012. Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2879–2888.

[14] Chris Evans and Nicola J Gibbons. 2007. The interactivity effect in multimedia learning. *Computers & Education* 49, 4 (2007), 1147–1160.

[15] Hendrik Friederichs, Sandra Ligges, and Anne Weissenstein. 2014. Using Tree Diagrams without Numerical Values in Addition to Relative Numbers Improves Students' Numeracy Skills: A Randomized Study in Medical Education. *Medical Decision Making* 34, 2 (2014), 253–257. https://doi.org/10.1177/0272989X13504499 arXiv:https://doi.org/10.1177/0272989X13504499 PMID: 24085290.

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

[17] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats.

[18] Pascal Goffin, Petra Isenberg, Tanja Blascheck, and Wesley Willett. 2020. Interaction Techniques for Visual Exploration Using Embedded Word-Scale Visualizations. https://doi.org/10.1145/3313831.3376842

[19] Paul Han, William Klein, Bill Killam, Tom Lehman, and Holly Massett. 2011. 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 (03 2011), 106–13. https://doi.org/10.1016/j.pec.2011.01.033

[20] S. G. Hart. 2006. NASA-Task Load Index (NASA-TLX); 20 years later.

[21] Jeffrey Heer and Ben Shneiderman. 2012. Interactive Dynamics for Visual Analysis. *Queue* 10, 2 (Feb. 2012), 30?55. https://doi.org/10.1145/2133416.2146416

[22] Vince Kellen, Susy Chan, and Xiaowen Fang. 2007. Facilitating Conditional Probability Problems with Visuals. In *Human-Computer Interaction. Interaction Platforms and Techniques*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–71.

[23] Azam Khan, Simon Breslav, Michael Glueck, and Kasper Hornbæk. 2015. Benefits of visualization in the Mammography Problem. *International Journal of Human-Computer Studies* 83 (2015), 94 – 113. https://doi.org/10.1016/j.ijhcs.2015.07.001

[24] Azam Khan, Simon Breslav, and Kasper Hornbæk. 2018. Interactive Instruction in Bayesian Inference. *Human - Computer Interaction* 33, 3 (2018), 207–233. https://doi.org/10.1080/07370024.2016.1203264 arXiv:https://doi.org/10.1080/07370024.2016.1203264

[25] H. Lam. 2008. A Framework of Interaction Costs in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov 2008), 1149–1156. https://doi.org/10.1109/TVCG.2008.109

[26] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale. 2012. Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2689–2698.

[27] Z. Liu and J. Stasko. 2010. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov 2010), 999–1008. https://doi.org/10.1109/TVCG.2010.177

[28] Richard Mayer, Julie Heiser, and Steve Lonn. 2001. Cognitive Constraints on Multimedia Learning: When Presenting More Material Results in Less Understanding. *Journal of Educational Psychology* 93 (03 2001), 187–198. https://doi.org/10.1037/0022-0663.93.1.187

[29] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the Effect of Visualizations on Bayesian Reasoning Through Crowdsourcing. *Visualization and Computer Graphics, IEEE Transactions on* 18 (12 2012), 2536–2545. https://doi.org/10.1109/TVCG.2012.199

[30] Alvitta Ottley, Aleksandra Kaszowska, R. Jordan Crouser, and Evan M. Peck. 2019. The Curious Case of Combining Text and Visualization. https://doi.org/10.2312/evs.20191181

[31] Alvitta Ottley, Blossom Metevier, Paul Han, and Remco Chang. 2012. Visually Communicating Bayesian Statistics to Laypersons.

[32] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. J. Han, and R. Chang. 2016. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 529–538. https://doi.org/10.1109/TVCG.2015.2467758

[33] William A Pike, John Stasko, Remco Chang, and Theresa A O'connell. 2009. The science of interaction. *Information visualization* 8, 4 (2009), 263–274.

[34] M. Pohl, M. Smuc, and E. Mayr. 2012. The User Puzzle–Explaining the Interaction with Visual Analytics Systems. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec 2012), 2908–2916. https://doi.org/10.1109/TVCG.2012.273

[35] Eric Ragan, Alex Endert, Doug Bowman, and Francis Quek. 2012. How Spatial Layout, Interactivity, and Persistent Visibility Affect Learning with Large Displays. https://doi.org/10.1145/2254556.2254576

[36] Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of experimental psychology. General* 130 (10 2001), 380–400. https://doi.org/10.1037//0096-3445.130.3.380

[37] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. , 336-343 pages.

[38] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing Uncertainty About the Future. *Science* 333, 6048 (2011), 1393–1400. https://doi.org/10.1126/science.1191181 arXiv:https://science.sciencemag.org/content/333/6048/1393.full.pdf

[39] Sabine Theis, Christina Bröhl, Matthias Wille, Peter Rasche, Alexander Mertens, Emma Beauxis-Aussalet, Lynda Hardman, and Christopher M. Schlick. 2016. Ergonomic Considerations for the Design and the Evaluation of Uncertain Data Visualizations. In *Human Interface and the Management of Information: Information, Design and Interaction,* Sakae Yamamoto (Ed.). Springer International Publishing, Cham, 191–202.

[40] Lyndal J Trevena, Brian J Zikmund-Fisher, Adrian Edwards, Wolfgang Gaissmaier, Mirta Galesic, Paul KJ Han, John King, Margaret L Lawson, Suzanne K Linder, Isaac Lipkus, et al. 2013. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *BMC medical informatics and decision making* 13, 2 (2013), 1–15.

[41] Jennifer Tsai, Sarah Miller, and Alex Kirlik. 2011. Interactive Visualizations to Improve Bayesian Reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 55, 1 (2011), 385–389. https://doi.org/10.1177/1071181311551079 arXiv:https://doi.org/10.1177/1071181311551079

[42] J. J. van Wijk. 2005. The value of visualization. , 79-86 pages. https://doi.org/10.1109/VISUAL.2005.1532781

[43] Pei-Yu Wang, Brandon K Vaughn, and Min Liu. 2011. The impact of animation interactivity on novices' learning of introductory statistics. *Computers & Education* 56, 1 (2011), 300–311.

[44] Michael Wybrow, Niklas Elmqvist, Jean-Daniel Fekete, Tatiana Von Landesberger, Jarke Van Wijk, and Björn Zimmer. 2014. Interaction in the Visualization of Multivariate Networks. , XV, 237 pages. https://hal.inria.fr/hal-00974335

[45] Ji Yi, Youn Kang, John Stasko, and Julie Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *Visualization and Computer Graphics, IEEE Transactions on* 13 (12 2007), 1224 – 1231. https://doi.org/10.1109/TVCG.2007.70515

[46] Qiyu Zhi, Alvitta Ottley, and Ronald Metoyer. 2019. Linking and Layout: Exploring the Integration of Text and Visualization in Storytelling. , 675–685 pages.