Interactive Instruction in Bayesian Inference

Azam Khan[1,2], Simon Breslav[1], Kasper Hornbæk[2]

[1] *Autodesk Research, Canada*

[2] *University of Copenhagen, Danmark*

ABSTRACT

An instructional approach is presented to improve human performance in solving Bayesian inference problems. Starting from the original text of the classic Mammography Problem, the textual expression is modified and visualizations are added according to Mayer's principles of instruction. These principles concern coherence, personalization, signaling, segmenting, multimedia, spatial contiguity, and pre-training. Principles of self-explanation and interactivity are also applied. Four experiments on the Mammography Problem showed that these principles help participants answer the questions at significantly improved rates. Nonetheless, in novel interactivity conditions, performance was lowered suggesting that more interaction can add more difficulty for participants. Overall, a leap forward in accuracy was found, with more than twice the participant accuracy of previous work. This indicates that an instructional approach to improving human performance in Bayesian inference is a promising direction.

CONTENTS

1. INTRODUCTION
2. RELATED WORK
    2.1. Textual and Numeric Format
    2.2. Visualization
    2.3. Principles of Multimedia Instruction
    2.4. Interactivity
3. EXPERIMENTAL DESIGN
    3.1. Procedure
    3.2. Limitations
    3.3. Variables
    3.4. Results Overview
4. EXPERIMENT 1: TEXT ONLY
    4.1. Hypothesis
    4.2. Results
    4.3. Discussion: Experiment 1
5. EXPERIMENT 2: DIAGRAM
    5.1. Hypotheses

5.2. Results
5.3. Discussion: Experiment 2
6. EXPERIMENT 3: INTERACTIVE FOCUS
6.1. Hypotheses
6.2. Results
6.3. Discussion: Experiment 3
7. EXPERIMENT 4: INTERACTIVE DIAGRAM
7.1. Hypotheses
7.2. Results
7.3. Dragging Mistakes & Box Arrangment
7.4. Discussion: Experiment 4
8. DISCUSSION
8.1. Text-diagram versus Text-only
8.2. New Text diagram versus New Text-only
8.3. Interactive Focus versus New Text-only
8.4. Interactive conditions versus New Text diagram
8.5. Principles of Multimedia Instruction
9. CONCLUSION

1. INTRODUCTION

Human decision making under uncertainty has been studied for decades. Due to poor accuracy in answering these kinds of problems, researchers have investigated two primary modifications to early works in Bayesian inference in hopes of improving human judgment: the use of frequency formats in the textual descriptions of the problems, and the addition of visualizations of the numeric values and/or the structure of the problem. The classic Mammography Problem (Eddy, 1982) is often used in studies as a representative instance of a class of Bayesian inference problems and we adopt this problem in our decision making study as well.

The Mammography Problem is a good testbed problem for studying judgement performance as it hits a number of interacting challenges for people. The high level of difficulty typically yields a low performance for correct responses of only 20%, yet the results have been replicated many times without significant variation. While a base-rate bias (Tversky & Kahneman, 1981) and the numeric format of probabilities (Gigerenzer & Hoffrage, 1995) have been proposed as major barriers keeping people from correctly answering the problem, the underlying cognitive mechanisms at play in this vexing problem are still unknown.

The high level of adoption of the frequency format in the study of Bayesian inference problems (e.g., “10 out of 100”), over the use of normalized probabilities (e.g., “0.1”) or percentages (e.g., “10%”), indicates the general acceptance of frequency format as a useful modification in problem formulations (Gigerenzer & Hoffrage, 1995) albeit with some contradictory evidence (Ayal & Beyth-Marom, 2014; J. S. Evans, Handley, Perham, Over, & Thompson, 2000). However, despite the large number of studies on the augmentation of Bayesian problems with visualizations, there has not been a single type of graph or chart that has generally been adopted as a useful modification (Khan, Breslav, Glueck, & Hornbæk, 2015; Micallef, Dragicevic, & Fekete, 2012).

In addition to numeric format and accompanying visualizations, as two classes of modifications to the expression of Bayesian problems, we propose the adoption of the *Principles of Instruction* as a third avenue of worthwhile investigation. A significant body of research was amassed on human information processing in learning that Richard Mayer summarized into ten principles to help guide the design of multimedia learning (Mayer, 2008).

In this paper, we contribute new designs of the textual, graphical, and interactive expression of the classic Mammography Problem by applying instructional science principles. We evaluate these designs through a series of crowdsourcing experiments that demonstrate novel insights into the expression of Bayesian inference problems. Specifically, we apply the instructional principles of coherence, personalization, signaling, segmenting, multimedia, spatial contiguity, and pre-training as defined by Mayer (Mayer, 2008). We also apply a form of the Self-Explanation Principle (Wylie & Chi, 2005) and a proposed Interactivity Principle (Wang, Vaughn, & Liu, 2011). By using established guidelines, our approach provides the benefit, over ad hoc methods, of probing the problem and developing novel expressions of the problem in a systematic way. And as the Principles of Instruction are based on the Cognitive Theory of Multimedia Learning (Mayer, 2005), exploring these principles on Bayesian inference problems may provide new insights into the cognitive demands of this class of decision making problem.

2. RELATED WORK

The classic Mammography Problem began with Casscells et al. (1978) and Eddy (1982) to test understanding of probabilistic reasoning in medicine. The surprisingly poor performance of medical students and professionals in answering this question led to the broader interest in this problem in the decision making and visualization communities. One common version of the textual form of the problem is:

> At age forty, when women participate in routine screening for breast cancer, 10 out of 1000 will have breast cancer. However, 8 of every 10 women with breast cancer will get a positive mammography, and 95 out of every 990 women without breast cancer will also get a positive mammography.
>
> Given a new group 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?

While the Mammography Problem may seem to be very specific, it represents an important class of Bayesian inference problems that occur for many people in daily life. As will be shown, by adopting an instructional approach we come much closer to helping people solve this important problem in real-world critical decision making scenarios including medical and financial decision making.

As this class of problem has been studied by many researchers for decades, it should be noted that it generally follows the original scenario which does not explicitly state any prior beliefs to be updated. Therefore, even though this problem class is often said to test "Bayesian reasoning", it has been recommended that this problem be called "statistical inference" (Mandel, 2014). However, as this dismisses the applicability of the mechanics of Bayes Theorem in understanding and calculating an answer, we adopt the hybrid name "Bayesian inference" for the Mammography Problem class.

2.1. Textual and Numeric Format

Previous work has shown how various ways of transforming the textual representation and the numeric format of the problem can affect Bayesian inference performance (J. S. Evans et al., 2000; Gigerenzer & Hoffrage, 1995; Ottley et al., 2016). This paper also presents a number of transformations to the Mammography Problem and questions what effect these changes may have. So, to preserve a meaningful way to measure the effect of these changes compared to previous work, we do not change the format for the numeric answer that participants are asked to provide. We therefore present participants with the traditional frequency format of numeric information to support direct comparisons to previous work (e.g., "(subset) out of (set)") (Gigerenzer & Hoffrage, 1995). In this way, we avoid introducing a confounding factor by asking a different question from the traditional Bayesian inference question.

2.2. Visualization

The use of static graphs, charts, and diagrams to augment Bayesian inference problems has been studied in human-computer interaction (Brase, 2008; Breslav, Khan, & Hornbæk, 2014; Cole, 1989) and information visualization (Micallef et al., 2012; Ottley, Metevier, Han, & Chang, 2012) but with generally poor performance (Calvillo, DeLeeuw, & Revlin, 2006). Furthermore, in a recent study, it was shown that static visualizations in these problems offer no benefit over completeness of information in a text-only form (Khan et al., 2015). That is, while visualizations generally provide increased performance over text alone, this same level of performance increase can be achieved by providing a table of all the values involved in finding the correct answer without any type of diagram.

An ongoing debate (Brase, 2008) between frequency coding proponents and nested-set proponents, together with the questionable benefits of static visualization (Khan et al., 2015), motivates the focus of the current paper. In this work, we examine an alternate direction in improving Bayesian inference by examining the problem from an instructional perspective.

2.3. Principles of Multimedia Instruction

We review the ten principles of multimedia instruction (Mayer, 2008) (see Figure 1) and describe how they may apply to problems involving Bayesian inference. Note that we focus on instruction for problem comprehension and are not testing learning *per se*, for example retention or generalization to novel problems. As our target population of interest is the general public receiving medical or financial information on which to base critical decisions, our focus is on improving naïve statistical inference in untrained participants given only a single problem instance.

Principle	Definition
Reducing Extraneous Processing	
Coherence	Reduce extraneous material.
Signaling	Highlight essential material.
Redundancy	Do not add on-screen text to narrated animation.
Spatial Contiguity	Place printed words next to corresponding graphics.
Temporal Contiguity	Present corresponding narration and animation at the same time.
Managing Essential Processing	
Segmenting	Present animation in learner-paced segments.
Pre-training	Provide pre-training in the name, location, and characteristics of key components.

Modality	Present words as spoken text rather than printed text.
Fostering Generative Processing	
Multimedia	Present words and pictures together rather than words alone.
Personalization	Present words in conversational style rather than formal style.

Figure 1: Principles of Multimedia Instruction (Mayer 2008)

We chose not to include narration or animation in our treatment of the Mammography Problem. Therefore, we did not apply the Principles of Redundancy, Temporal Contiguity, or Modality. For the remaining Principles, we adapted them from the original multimedia context to our decision making purposes.

Coherence: *While it may seem that extra supportive information in a lesson would be beneficial to learning, it has been shown that better learning outcomes are achieved when extraneous material is excluded*. In the Mammography Problem, several opportunities exist to simplify the classic problem text using coherence. On the other hand, adding visualization and interactivity into the problem may cause extraneous processing thereby reducing participant performance.

Signaling: *When it is not possible to further remove material, it may still be possible to highlight essential material with headings or other methods to emphasize critical information*. In the Mammography Problem, we apply **bold styling** to text that could be emphasized and to essential numbers. We also use Signaling to visually highlight, through a yellow flash, an area of the screen where content has changed. For example, in the experiments presented here, when the "Next Step" button is pressed, the next question and answer area are displayed and the background of that area is briefly drawn in a bright yellow color to help indicate which screen content has changed guiding the attention of participants toward essential material.

Spatial Contiguity: *Poor layout can also cause extraneous processing in participants. When captions or labels are not placed together with the object to which they apply, a split-attention effect occurs as the participant must attend to both locations and conceptually integrate the presentation*. In the double-tree diagram (see Figure 6) that we employed in both static and interactive experimental conditions, we placed the labels inside boxes representing the nodes in the tree. In this way, the labels were always with the node to which they refer, avoiding the typical arrow pointing from a displaced label to a part of the diagram.

Segmenting: *When the learning material is too complex, essential processing could overwhelm the participant thereby reducing learning performance. Segmenting breaks down a continuous presentation into a number of individual segments, consisting of one or two sentences each*. At the end of each segment, a "Next" button appears that the participant can press to advance the lesson and have the next segment presented. This allows the participant to control the pace of the lesson so that they can cognitively fully represent each part before moving on to the next part. We use this within a single page, cumulatively revealing new parts of the problem.

Pre-training: *Describing each component before describing a whole system can be beneficial to learning. Participants who are already familiar with the components can apply more cognitive capacity to building a cause-and-effect model of the system*. We use this principle in the design of the interactive diagram we present later where the nodes are interactively placed into a double-tree diagram. In this way, participants can focus on each node label while forming a mental model of the whole system.

Multimedia: *This principle states that participants learn better from words and pictures than from words alone*. We use this principle in the design of our experimental conditions and in the development of our hypotheses.

Personalization: *Participants learn better when words are in a conversational style than in a formal style. This principle recommends the transformation of "the" to "your" in text and narration, and the use of less formal terms*. We apply this principle to the Mammography Problem text in reducing technical jargon, using subjective phrasing and making the text gender-neutral. We also modified implied connotations. For example, "positive mammography" may sound desirable if it is not known that a mammography is a test for breast cancer. Also, there is disagreement in the use of terms such as "chance" and "probability" and whether participants interpret terms as representing single-event probabilities or as frequencies. As Brase (Brase, 2008) showed that either may be used, we preferred "chance" as other terms could be too technical.

In summary, by considering the Mammography Problem from an instructional perspective, a number of potential issues may be revealed and controlled. Other research has commented on how sensitive the problem may be to the specific wording (Ottley, et al., 2016) and how the text and format of both the problem and the question is often "deliberately altered in a number of respects" (J. S. Evans et al., 2000). In particular, Ottley et al. (2016) studied problem representation and the effect of individual differences, for example, spatial abilities, and tested how these issues influence Bayesian inference performance. But by adopting the Principles of Instruction, we introduce a more controlled approach to justify alterations to the text as part of the intentional design of the experimental conditions. Furthermore, by modifying the problem according to these principles, we show how they may be used as a set of guidelines to simplify any given problem text.

2.4. Interactivity

It has been proposed that an *Interactivity Principle* be considered together with the multimedia Principles of Instruction as the work of Mayer involved interactivity in some studies but that it was not discussed at length (Wang et al., 2011). A beneficial interactivity effect has been reported (C. Evans & Gibbons, 2007) but it was found to be more pronounced for learning transfer than retention. Furthermore, cognitive limits are the foundation of the Principles of Instruction implying that when these limits are passed, learning will be decreased despite adding more learning activities (Mayer, Heiser, & Lonn, 2001). On the topic of Bayesian inference, simple interactivity has been touched upon in a single study. When participants could show or hide four problem subsets visualized on an otherwise static frequency grid, performance increased significantly (Tsai, Miller, & Kirlik, 2011). However, the Mammography Problem was not tested and the question format was not disclosed so results cannot be compared to previous work.

The concept of interactivity has been discussed in the instructional multimedia domain (C. Evans & Gibbons, 2007), the information visualization community (Yi, Kang, Stasko, & Jacko, 2007), and is a central topic in the field of human-computer interaction (HCI). In cognitive modeling research, related to multimedia learning, the term "interactivity" is loosely related to the level of interactivity between the participant and the multimedia

system, but is more broadly used to describe interactions between sub-models in a cognitive architecture (Domagk, Schwartz, & Plass, 2010; Reed, 2006).

For our purposes, we employ low-level HCI interactivity to add value to otherwise static visualizations. Specifically, we use simple highlighting and focus (the Signaling Principle) (Dix & Ellis, 1998), control of pace (the Segmenting Principle), and control of objects (the Pre-training Principle) (Paas, Van Gerven, & Wouters, 2007; Wang et al., 2011). In contrast, high-level interactivity includes control of parameters and complex analysis functionality (Wang et al., 2011).

The interactivity techniques of highlighting and pace control are used throughout the experiments presented here and will be described in the following section. However, control of objects is only employed to add interactivity to an otherwise static diagram, and so, will be described in a later section. We also applied the Principles of Multimedia Learning (Wylie & Chi, 2005) in an interactive focus task design (see Figure 8) to be discussed in Section 6.

3. EXPERIMENTAL DESIGN

Four experiments were designed to evaluate the benefits of applying instructional principles to the Mammography Problem. In this section we will briefly outline our procedure and variables. See our Supplemental Material for a more detailed discussion of the experimental design and limitations.

3.1. Procedure

The experiment was carried out using Mechanical Turk (MTurk), a crowd-sourcing service. We used a between-subjects design, with each participant completing a single trial to control against learning effects and fatigue. Each experiment contained two conditions, for each condition the participant was shown four pages. First, an introduction page was shown describing the rules of experiment and a button to enter the Full Screen mode of the browser.

The second page contained the content of the specific condition being tested. For this page we follow the guidelines for page layout, screen resolution and window size to ensure legibility and visibility of all content (no scrolling) (Khan et al., 2015). We used Segmentation and broke down the problem and question presentation into a number of steps. All steps were shown on a single page so each step simply revealed more of the page. In this way, participants controlled the pace of the presentation through simple interaction by using the "Next Step" button. During each step, the participants had to completely fill out the exercise before moving on to the next step. As mentioned previously, we used Signaling through highlighting to direct the participants' attention to the area of the screen where the new content appeared (Khan, Matejka, Fitzmaurice, & Kurtenbach, 2005).

The third page contained a catch question based on the recommendations for crowdsourced studies of Downs et al. (Downs, Holbrook, Sheng, & Cranor, 2010) where 39% of participants were disqualified. On average, in our study, the catch question disqualified 32% of participants per condition. The final page requested demographics information including gender, level of education, category of occupation, level of statistics training, and general comments.

We kept our survey active until approximately 100 participants per condition completed the survey *and* correctly answered the catch question. Participants were compensated $1.00

USD for their participation. The qualification requirement for the study included a Human Intelligence Tasks (HIT) Approval Rate ≥ 95% and Number of HITs Approved ≥ 50. MTurk workers were restricted to only one trial and only one of the HITs. In total 1,107 unique participants were tested across four experiments, however, analysis is reported only on 749 of these participants, those who answered the catch question correctly. Each participant performed a single test from a single condition. Average completion time across the entire group of participants was 5 minutes 19 seconds.

3.2. Limitations

A key limitation of using a crowd-sourcing service in an experiment is control of the experimental environment, including the time and place that the experiment is actually performed by participants. The process starts with the posting of an experiment and requesting the number of participants desired. While the time of the posting may be controlled, the time of day when a participant will actually perform the experiment cannot be. Furthermore, it is difficult to place restrictions on this as participants from multiple time zones may be involved.

Also, the question of generalization is a limitation of the present study. That is, while we feel the results presented here are compelling to support the further investigation of instruction in Bayesian inference problems, we have only tested the Mammography Problem in a number of different forms. To test how the results presented here may generalize we could investigate other Bayesian inference problems that have appeared in the literature such as the Cab Problem and the Economics Problem (Micallef et al., 2012).

3.3. Variables

In all experiments, the independent variable was the condition. An exact answer is the value directly entered by the participant. A combined final probability answer is calculated by dividing the exact numerator with the exact denominator. The weakness of the combined answer is that incorrect thinking may still result in a correct combined answer. For simplicity of reporting the results, we only report exact answers, as they most clearly convey correct thinking. Thus, the main dependent variable measured is EXACT ∈ { *true*, *false* }, and is *true* when the numerator = 8 and the denominator = 103.

3.4. Results Overview

We present the results overview here to avoid duplication throughout the paper (Figure 2). It shows the overall findings of the four experiments as well as the relation between the conditions, where some are extended to form other conditions. The most striking aspect of the results is that the application of the principles of instruction tripled performance in the text-only conditions (experiment 1, from 5% to 16%) and, when a static diagram was added to the new text, performance tripled again (experiment 2, 16% to 51%). We discuss the results further in the sections of each experiment.

Exp.	Condition	N	Exact	Exact (%)	χ^2 (df=1)	φ	Odds Ratio	Time (s)
1	Text-only	92	5	5%	4.7*	0.2	0.3	(M=240, SD=205)
	New Text-only	91	15	16%				(M=227, SD=254)
2	Text-diagram	98	32	32%	6.3*	0.2	0.5	(M=285, SD=136)
	New Text-diagram	99	50	51%				(M=334, SD=248)
3	Interactive Focus	95	35	37%	0.04	0.03	0.9	(M=272, SD=145)
	New Text-diagram + Interactive Focus	93	32	34%				(M=450, SD=256)
4	Interactive Diagram	89	31	35%	0	0	1.0	(M=362, SD=221)
	Interactive Diagram + Focus	92	32	35%				(M=432, SD=233)

Figure 2: Overview of results for all 4 experiments. Column N shows total number of participants, Column Exact shows the number of participants that answered correctly and column Exact (%) shows the number as a bargraph as a percentage of N. Error bars in the column Exact (%) represent 95% confidence interval. Column χ^2 shows Chi-square test with Yates' continuity correction, * indicated statistical significance with p < 0.05. Column ϕ shows phi coefficients. Time reports mean and std. dev. of time taken in seconds to fill out the main page of the survey.

4. EXPERIMENT 1: TEXT ONLY

The purpose of the first experiment was to test if changing the text of the problem and question used in previous work would impact participant performance. The experiment consisted of two conditions, each containing different versions of the text. The first condition, TEXT-ONLY, used text taken directly from Micallef et al. (Micallef et al., 2012), and the second condition, NEW TEXT-ONLY, used the modified version of the text (see Figure 4 and Figure 5).

To create the text used in the NEW TEXT-ONLY condition, we applied the Principles of Instruction (Mayer, 2008) and the Set-inclusion Cue (J. S. Evans et al., 2000). In Figure 3, we show the principles being used together with the text changes made. Text was either removed, added, or styled (in bold) in a series of stages.

Principle	Mammography Problem and Question Text
Original Text from (Micallef et al., 2012)	You know the following information: • 10 out of every 1000 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? Your answer: ____ out of ____
Coherence Principle (Extraneous material is excluded. Consistency.)	You know the following information: • 10 out of every 1000 women ~~at age forty who participate in routine screening~~ who have a mammography, 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? Your answer: ____ out of ____
Set-inclusion Cue (J. S. Evans et al., 2000)	You know the following information: • Of 1000 women who have a mammography, 10 women have breast cancer. • Of the 10 women with breast cancer, 8 women will get a positive mammography. • Of the 990 women without breast cancer, 95 women will also get a positive mammography. Here is a new sample of women who got a positive mammography. How many of these women do you expect to actually have breast cancer? Your answer: ____ out of ____
Personalization Principle (Remove technical jargon, Gender-neutral,	You know the following information: • Of 1000 people tested for skin cancer, 10 people will actually have skin cancer. • Of the 10 people with skin cancer, 8 people ~~will~~ got a test result that says they have cancer. • Of the 990 people without skin cancer, 95 people will also get a test result that says they have cancer.

Implied connotations.)	Here is a new sample of people who got a test result that says they have skin cancer. How many of these people do you expect to actually have skin cancer? Your answer: ____ out of ____
Signaling Principle (Highlighting essential material. Clarify misleading information.)	You know the following information: • Of **1000** people tested for skin cancer, **10** people will **actually** have skin cancer. • Of the **10** people with skin cancer, **8** people got a correct **test result** that says they have cancer. • Of the **990** people without skin cancer, **95** people incorrectly get a **test result** that says they have cancer. Here is a new sample of people who got a **test result** that says they have skin cancer. How many of these people do you expect to **actually** have skin cancer? Your answer: ____ **out of** ____
Personalization Principle (Directed to reader.)	You know the following information: • Of **1000** people tested for skin cancer, **10** people will **actually** have skin cancer. • Of the **10** people with skin cancer, **8** people got a correct **test result** that says they have cancer. • Of the **990** people without skin cancer, **95** people incorrectly get a **test result** that says they have cancer. ~~Here is a new sample of people who~~ If you got a **test result** that says ~~they~~ you have skin cancer, ~~How many of these people do~~ what are the chances that you ~~expect to~~ **actually** have skin cancer? ~~Your answer~~ The chances are: ____ **out of** ____
Result	You know the following information: • Of **1000** people tested for skin cancer, **10** people will **actually** have skin cancer. • Of the **10** people with skin cancer, **8** people got a correct **test result** that says they have cancer. • Of the **990** people without skin cancer, **95** people incorrectly get a **test result** that says they have cancer. If you got a **test result** that says you have skin cancer, what are the chances that you **actually** have skin cancer? The chances are: ____ **out of** ____

Figure 3: Application of Principles of Instruction and Principles of Multimedia Learning to the Mammography Problem text. Yellow highlighting indicates changes between steps.

4.1. Hypothesis

We hypothesize that changing the problem wording, by incorporating guidelines from instructional psychology, would make the Mammography Problem easier to understand and lead to more participants answering it correctly. See Figure 4 and Figure 5 for the screen designs. Note that each of the four segments, or steps, are revealed after the participant presses a "Next Step" button.

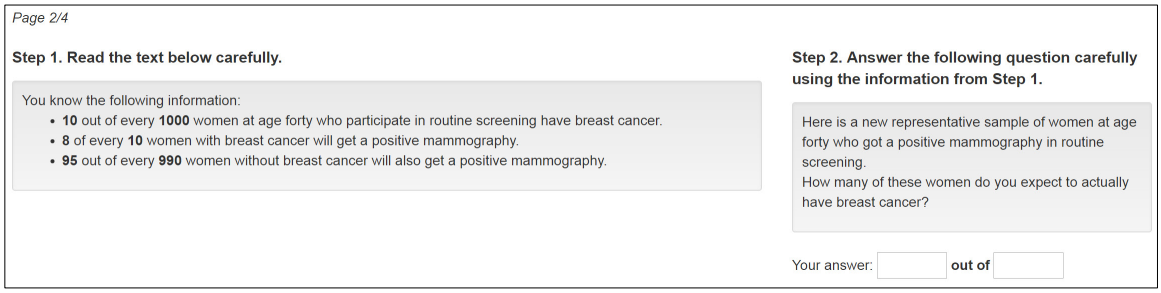

Figure 4: Experiment 1 screen for the TEXT-ONLY condition (text in step 1 and 2 specific to this condition).

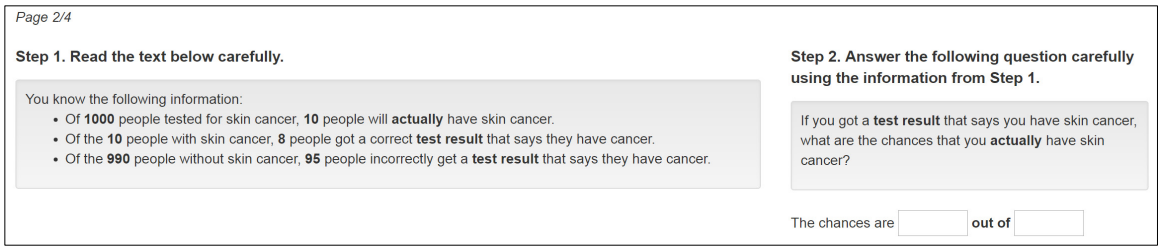

Figure 5: Experiment 1 screen for NEW TEXT-ONLY condition (text in step 1 and 2 specific to this condition).

4.2. Results

A total of 183 participants completed the experiment and answered the catch question correctly. The TEXT-ONLY replication condition achieved the same performance of 5% as in previous work (Breslav et al., 2014; Khan et al., 2015) (see Figure 2). However, the effect of applying the Principles of Instruction to the text is surprisingly high. The NEW TEXT-ONLY condition more than tripled performance with 16% correct responses. Chi-square test with Yates' continuity correction shows the difference is statistically significant ($\chi^2(1, N{=}183) = 4.7$, $p < 0.05$, $\phi = 0.18$, the odds ratio is 0.3). However, the problem is still very difficult for participants as, overall, 84% of participants who received NEW TEXT-ONLY and 95% of participants who received TEXT-ONLY still answered incorrectly. Participants did not take a statistically significant different amount of time across conditions ($F(1, 181) = 0.24$, $p = 0.65$).

4.3. Discussion: Experiment 1

The large overall performance improvement in NEW TEXT-ONLY, going from 5% to 16% correct, together with a moderate effect size, indicates the importance of carefully designing the text in word problems. While previous work has also examined the effect of specific phrasing and improving the problem explanation, we cannot directly compare their outcomes. For example, Ottley (Ottley et al., 2016) also changed which numbers were provided and asked participants to answer two questions instead of one in the traditional "out of" form. Evans also carefully controlled text manipulations (J. S. Evans et al., 2000) but not on the Mammography Problem text nor those numeric values.

This surprisingly large performance improvement indicates that inherent text difficulty is a critical aspect of Bayesian inference which has not previously been controlled for. By reducing language comprehension as a confounding factor, we can better test for statistical understanding. These results motivated the second experiment to probe the Multimedia Principle by replicating and measuring the effect of adding a static visualization.

5. EXPERIMENT 2: DIAGRAM

The Multimedia Principle states that pictures and text together is better than text alone. Previous work that tested the addition of visualization to the Mammography Problem text (Micallef et al., 2012) found this principle to hold and so, we replicated this scenario, but with additional principles applied: Segmentation, with some degree of Signaling and Spatial Contiguity. In this way, we could also produce a meaningful comparative measure of the same condition using the new transformed text, given that the transformation was successful in the first experiment.

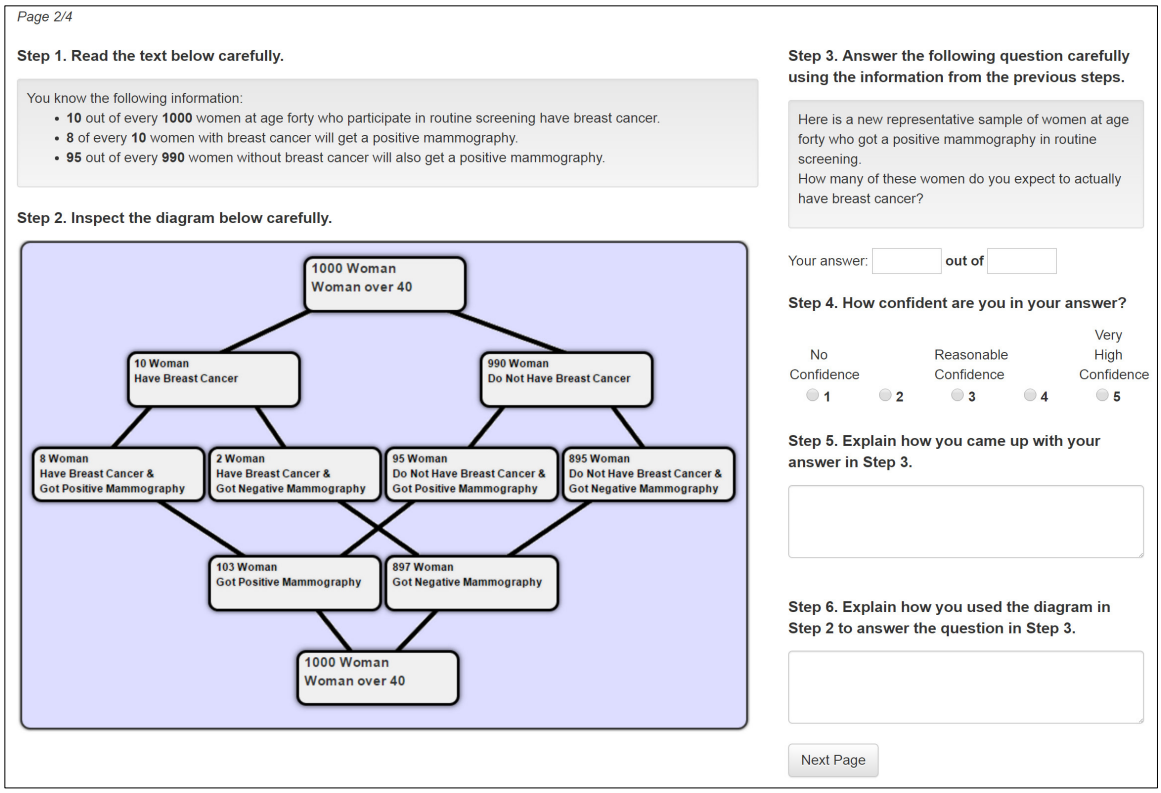

Figure 6: Experiment 2 screen for Text-diagram condition.

Experiment 2 consisted of two conditions, each containing different versions of the text. The first condition, TEXT-DIAGRAM, used text taken directly from Micallef et al. (Micallef et al., 2012) and the double-tree diagram taken from Khan et al. (Khan et al., 2015). The double-tree fully captures the double branching structure of a Bayesian problem, where the false-positive/true-positive and false-negative/true-negative symmetry of the problem is directly represented. Notably different from previous work is that we did not use color in the diagram except to subtly differentiate foreground from background. Previous work had strongly colored nodes in the double-tree to convey set containment or disbursement relationships. The second condition, NEW TEXT-DIAGRAM, used the transformed version of the text (see Figure 6 and Figure 7).

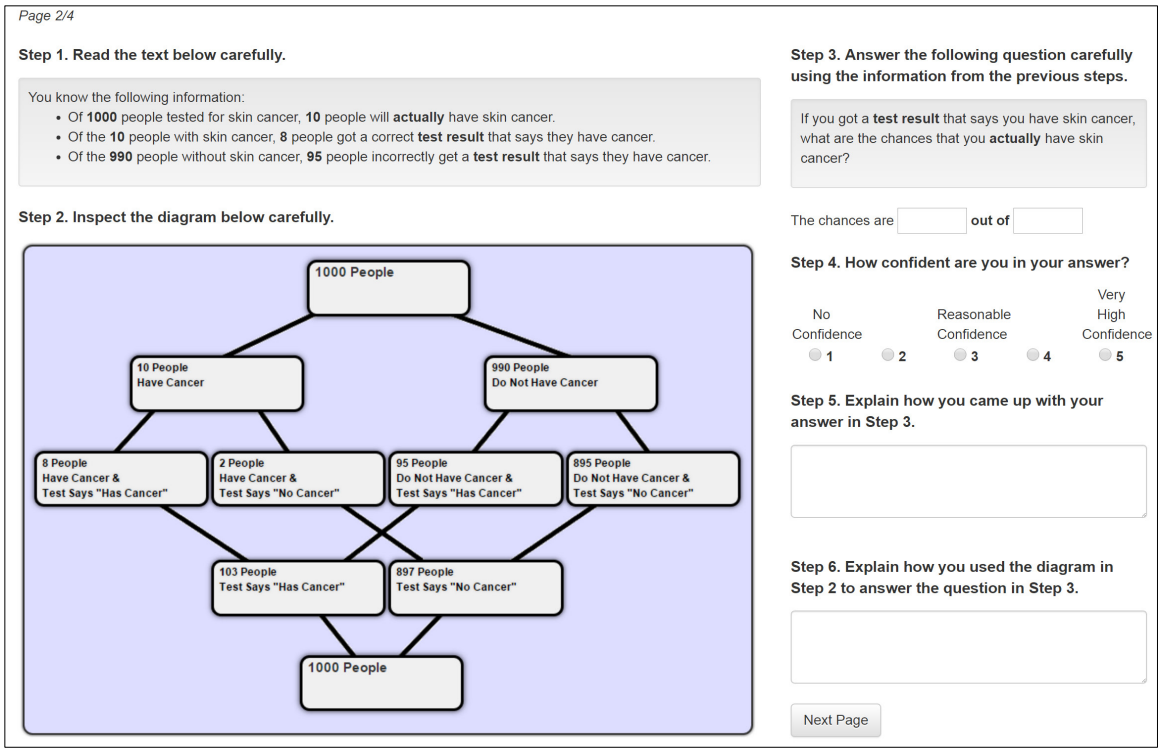

Figure 7: Experiment 2 screen for New Text-diagram condition.

5.1. Hypotheses

We hypothesize that the NEW TEXT-DIAGRAM condition will outperform the TEXT-DIAGRAM condition. In Experiment 1, when the Principles of Instruction were applied, performance increased by 11%. Based on that result, we expect a similar increase here.

5.2. Results

A total of 197 participants completed the experiment and answered the catch question correctly. The TEXT-DIAGRAM condition achieved 32% correct answers (see Figure 2). This is notably higher than previous work (20%) (Khan et al., 2015), which may be due to cohort makeup, the different node coloring, or the use of the Segmenting, Signaling, or Spatial Contiguity principles. Even more impressive is the 51% performance in the NEW TEXT-DIAGRAM condition, performing better than expected in a seemingly multiplicative way. The NEW TEXT-DIAGRAM condition was significantly better performing than the TEXT-DIAGRAM condition ($\chi^2(1, N=197) = 6.3$, $p < 0.05$, $\phi = 0.19$, the odds ratio is 0.46). Participants did not take a significant different amount of time across conditions ($F(1, 195) = 2.93$, $p = 0.09$).

5.3. Discussion: Experiment 2

Our hypothesis was confirmed that the performance of the new text diagram condition was 51% compared to 32% using the original text in a diagram. This is an even greater difference than expected. This strongly suggests that the original expression of the text prevented understanding by using technical jargon, not being gender-neutral, and not being personalized. Furthermore, from participant comments, it was clear that people really

understood the problem when answering correctly in their rephrasing into a set-subset format, described below, and were not simply copying the values into the frequency format answer boxes without understanding the nested set relations.

A complicating factor in interpreting the effect of the diagram is the additional information it presents to participants. That is, the problem text presents six numeric values whereas the diagram includes the complete set of nine numeric values. The effect of adding a complete set of numeric values to the text without a diagram was measured in previous work to improve performance from 4% to 17.2% (Khan et al., 2015), improving accurcy four fold. When a diagram was added, performance improved slightly to 20%. Our comparable condition achieved 32%, suggesting that the way the principles of instruction were applied to the overall presentation of the problem (Segmenting, Signaling, or Spatial Contiguity principles) also played a role in improving performance. However, the improvement from 32% to 51% almost doubled the benefit of the new text, suggesting that the transformation of the text is as important as the combined effect of providing complete information and a diagram.

6. EXPERIMENT 3: INTERACTIVE FOCUS

As we established that the new text is beneficial to participants, we explored adding more Principles of Instruction in this experiment. We added a simple interactive Pre-Training step, called Interactive Focus resulting in two conditions: INTERACTIVE FOCUS (see Figure 8) and NEW TEXT-DIAGRAM + INTERACTIVE FOCUS (Figure 9). Note, as with previous experiments, each of the steps were revealed one at a time after the participant pressed a "Next Step" button.

The motivation behind this new Pre-training step (see Figure 10) is to encourage participants to form a mental model of *set inclusion form* (J. S. Evans et al., 2000). We found that set inclusion format was used in the open ended comments that participants used in explanatory comments in the previous experiments (40 comments overall). For example, participants would transpose the values (set-subset) and comment "Out of 10 people, 8 people would have cancer". In contrast, the traditional frequency format (subset-set) asks participants to answer the problem by filling in blanks in the form "____out of _____", for example "8 out of 10". We refer to the subset-set order as frequency format, but we call the set-subset order *focus format* to reflect the focusing of the larger context value down to the subset value, also called Two-step Frequency (Girotto & Gonzalez, 2001).

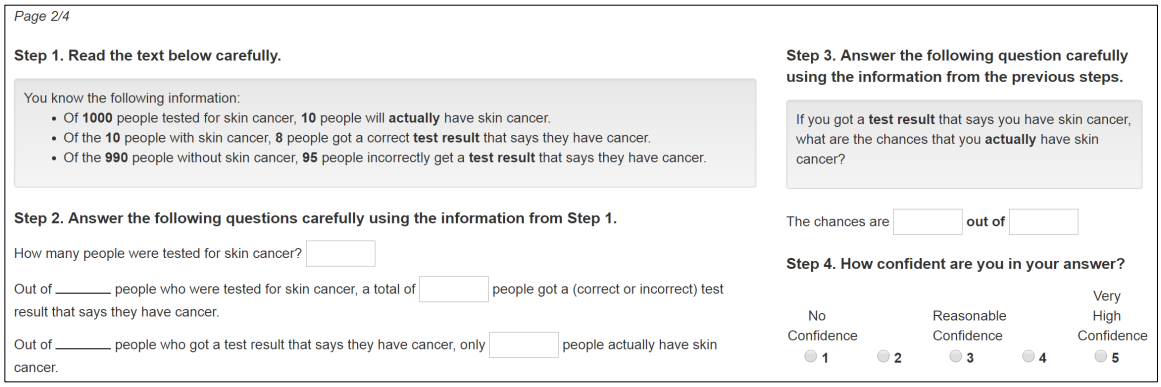

Figure 8: Experiment 3 screen for Interactive Focus condition.

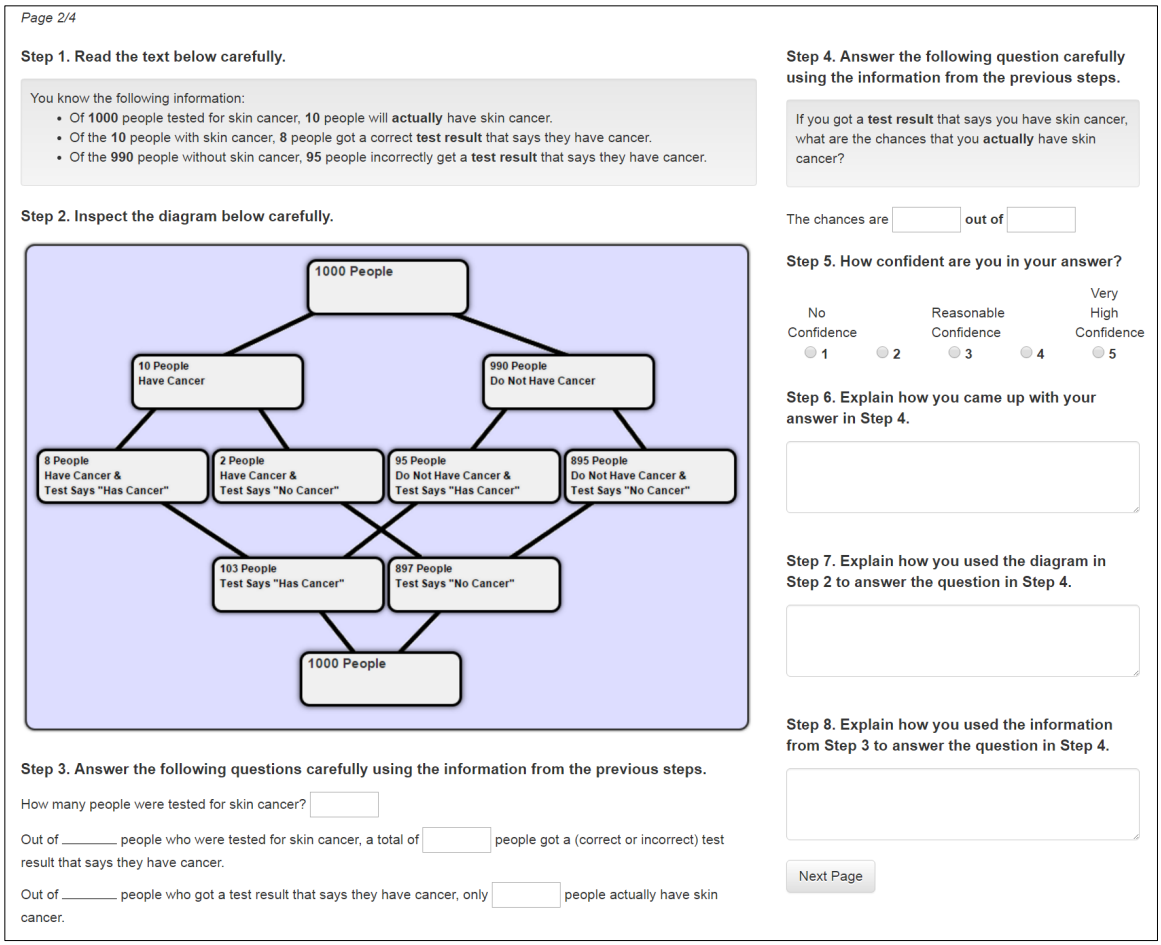

Figure 9: Experiment 3 screen for New Text-diagram + Interactive Focus condition.

To stress the set inclusion principle, we incorporated the Signaling Principle and used highlighting as an indicator that the subset of one statement becomes the containing set of the next statement, forming a chain. While this could be achieved in a single run-on sentence, we felt that Segmenting could apply here as well and we broke down the task into three statements, including two focus format statements. As text is entered into one of the text entry boxes, the related field in the next statement is immediately updated and a yellow background color briefly flashes to guide the participants' attention toward understanding the nested set relations. In Figure 10, we show an example where a participant is typing "1000" in the first field, updating the value of the first blank entry of the next sentence. On every update (after every keystroke), the background of the field is immediately changed to yellow, fading back to white over a transition period of 0.4 seconds.

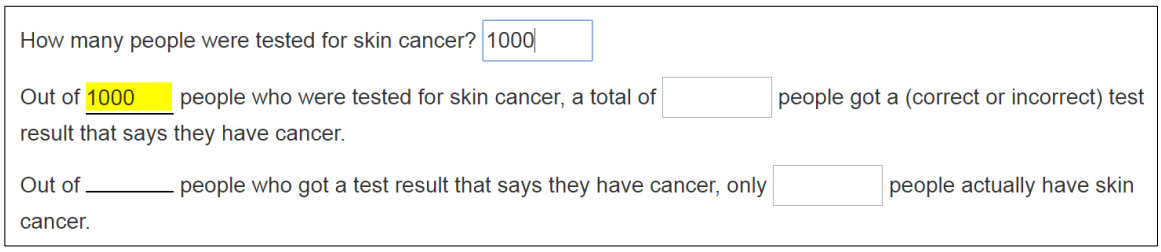

Figure 10: Interactive Focus step of the experiment.

The interactive focus task follows the Self-explanation Principle as a "scaffolded self-explanation prompt" (Wylie & Chi, 2005) and, as mentioned earlier, the task is analogous to the "two-step frequency question" (Girotto & Gonzalez, 2001). In this work, however, we posed the *problem* part (Figure 9, Step 1 and 3) in focus format but used the frequency format in the *question* part (Figure 9, Step 4), again, for meaningful comparison to previous work which used this form of the question.

6.1. Hypotheses

We hypothesize that adding the static double-tree diagram, in a NEW TEXT-DIAGRAM + INTERACTIVE FOCUS condition (Figure 9), will improve performance beyond the INTERACTIVE FOCUS condition.

6.2. Results

A total of 188 participants complited the experiment and answered the catch question correctly. However, the hypothesis was not confirmed. Surprisingly, performance was slightly lower in the NEW TEXT-DIAGRAM + INTERACTIVE FOCUS condition than in the INTERACTIVE FOCUS condition, from 37% to 34%. The difference was not significant ($\chi^2(1, N=188) = 0.04$, p = 0.8, $\phi = 0.03$, the odds ratio is 0.89), but is surprising, since although a double-tree diagram was added, performance was not increased.

In Figure 11, the three text fields in the interactive focus task are labelled Focus 1, 2, and 3 (as they occur in reading order). Exact Focus is the percentage of participants who had all three fields correct. Exact Ratio is the percentage of participants who had both values correct in Step 4 for their final answer. Exact All means that participants had both focus and ratio correct.

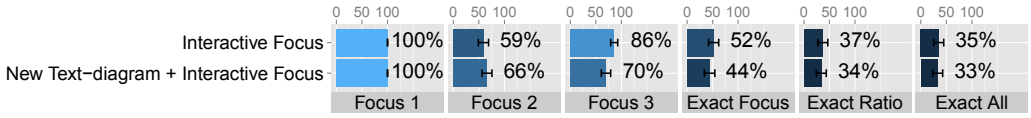

Figure 11: Exact Answers for the Ratio in Focus Study.

All participants correctly entered the population size in Focus 1. Surprisingly, participants were correct in Focus 3 more often than in Focus 2 even though the three values form a chain of subsets. In both conditions, Exact Focus was higher than Exact Ratio. This indicates that the interactive focus task was helpful to most, but not all, participants in answering the final question. As the final question was asked in frequency format –not focus format– some participants may have been confused by this. On the other hand, using formats in the problem that are different from the format of the answer may indeed be an improved way to measure understanding.

6.3. Discussion: Experiment 3

Doubling the participant performance, from 16% to 37%, by re-arranging the text, and asking for three values in the INTERACTIVE FOCUS task, indicates the high value of both the focus format and the principles of instruction. Considering that this intervention is neither graphical nor informative (no new information is provided), it is surprisingly effective. The same calculations must be made as in the NEW TEXT-ONLY condition and some participants

commented that "doing the math" was effortful. This seems to corroborate the importance of forming a helpful set inclusion mental model (J. S. Evans et al., 2000).

In contrast, an informative and visual addition to the NEW TEXT-DIAGRAM + INTERACTIVE FOCUS condition slightly reduced performance from 37% to 34%. Adding the static double-tree diagram to the interactive focus task eliminates the need for any computation or estimation as a complete set of data is embedded in the diagram. However, this extra information seems to come at the cost of interpreting the diagram which, overall, doesn't increase the performance. While this is consistent with the Coherence Principle, which states that "presenting more material results in less understanding", we were surprised by the low performance on this condition.

The most dramatic outcome is the performance cost of the combined double-tree diagram and focus task compared to the double-tree diagram alone; from 51% to 34%. It is interesting that when both representations are used, performance is reduced significantly. Prior work has shown that redundancy through additional on-screen text caused students to perform worse on tests of retention and transfer (Mayer et al., 2001).

Although the focus task seems to help participants form a useful mental model, the diagram alone helps participants in a much more pronounced way. This seems to indicate that more than one mental model is effective in the Mammography Problem class or that the cost of cognitive overload is much higher than the benefit of the mental model. More research would be needed to elucidate these interesting implications.

7. EXPERIMENT 4: INTERACTIVE DIAGRAM

The purpose of the final experiment was to investigate the effect of increased interactivity on participant performance. Specifically, we add a "control of object" task (the Pre-training Principle) (Paas et al., 2007; Wang et al., 2011) to the NEW TEXT-DIAGRAM condition in Experiment 2 and NEW TEXT-DIAGRAM + INTERACTIVE FOCUS condition from Experiment 3. In both cases, we simply replace the static diagram with an interactive diagram.

We explore the benefits, or costs, of interactivity in Pre-training through an interactive version of the double-tree diagram. The first condition of this experiment, Interactive Diagram, is very similar to New Text-Diagram condition (see Figure 5) in Experiment 2, except for the dragging task that asks participants to drag labeled boxes into the diagram. The dragging task works as follows: as a participant hovers the mouse over the labeled boxes, the cursor becomes an open hand icon (Figure 12a), and becomes a closed hand icon once the dragging begins. As the participant drags the boxes over empty white rectangles, they change color to gray, indicating a possible drop zone (see Figure 12). If the participant releases the box onto a valid drop zone, the box snaps into place, however, if the drop zone is not valid for a given box, the box will animate back to the original location. Each participant received the same arrangement of nodes (as seen in Figure 12). Note that there are 8 possible valid arrangements of the nodes and any one of them could be constructed. After the first valid box has been placed, four arrangements are still possible. After the second valid box has been placed, the arrangement becomes unique and all the remaining boxes have unique valid positions. In order for the participant to move on to the next step, participants had to successfully place all the nodes into the diagram. Please see the Supplementary Materials for a video of the dragging task in more detail.

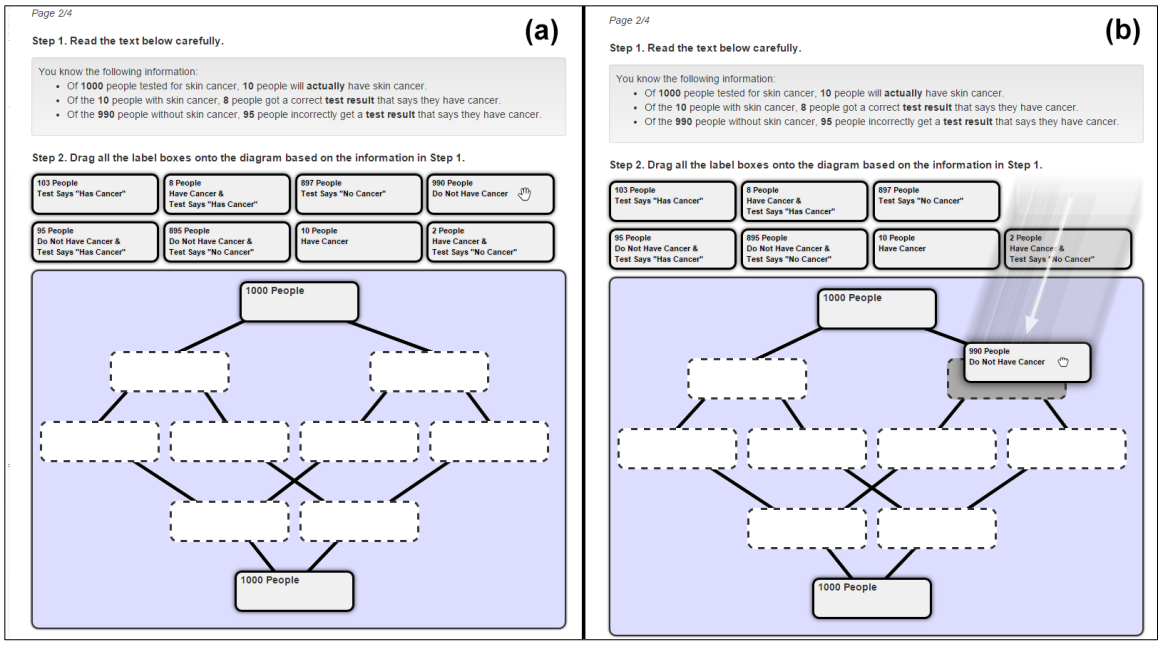

Figure 12: Interactive Diagram. (a) Initial state. (b) As a participant drags a box into the tree diagram, the empty rectangle is highlighted in gray. Note, the gray motion path and the white arrow were added to better illustrate the dragging motion in this figure but did not appear to participants.

The second condition of the experiment, INTERACTIVE DIAGRAM + FOCUS, was the same as the INTERACTIVE DIAGRAM, condition except it also had the INTERACTIVE FOCUS task as Step 3. In Experiment 3, a specific participant commented that "The diagram was somewhat hard for me to follow. I would have preferred to rearrange it in a manner that made more sense to me." For participants who shared this impression, the interactive version of the diagram may be beneficial.

7.1. Hypotheses

Motivated by the Pre-training Principal and the Interactivity Principle, we expect that adding increased interactivity, will lead to more participants answering the problem correctly, as they will have a better undersanding of all the different parts of the problem, and their relationship to each other, before attempting to answer the question.

7.2. Results

A total of 181 participants completed the experiment and answered the catch question correctly. The addition of the focus task to the INTERACTIVE DIAGRAM condition did not make any material difference; the INTERACTIVE DIAGRAM condition was very similar to INTERACTIVE DIAGRAM + FOCUS, ($\chi^2(1, N=181) = 0$, p = 1, $\phi = 0.001$, the odds ratio is 1.002).

7.3. Dragging Mistakes & Box Arrangment

We recorded all the dragging events that participants performed, to move the labelled boxes onto the diagram, in order to analyze the strategies that participants may have taken to complete the diagram. We also recorded when participants dragged a box onto a valid

node target and the drag operation was valid, or when participants made a mistake and attempted to drop a box at an incorrect node target. Overall, in the two conditions of the experiment (N=181), participants made an average of 4 mistakes (M=4.3, SD=4.6). Interestingly, there was a correlation between mistakes and getting the problem correct (Spearman $\rho = -0.2$, $p < 0.01$) which may indicate that participants who applied some level of effort to understand the interactive diagram task were more likely to answer the problem correctly.

Also, there was a significant difference in which boxes resulted in mistakes, see Figure 13. For example, when dragging a box that had the "897" number in the label, a total of 267 mistakes were made, while only 49 mistakes were made when dragging the "990" box. Both round numbers and single digit numbers were the easiest for participants to place. As previous work has not accounted for the complexity of the numbers themselves, this may also be an interesting area for future investigation. Incidentally, one participant commented: "I think I'm paying more attention to the round and/or easier numbers in the diagram than to "103" or "897"." Further investigation into a round number bias in this class of problem may be beneficial.

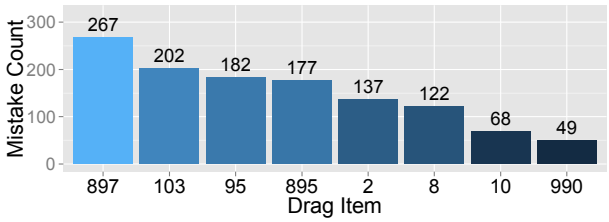

Figure 13: Dragging mistake count for different boxes (181 participants).

There are eight possible valid arrangements of the boxes in the double-tree diagram, and participants were able to arrange nodes in any of those valid arrangements. The most often arrangement used (35%) was the same one that we used in the previous experiments as the static diagram (Figure 6). As a post-hoc hypothesis, we wondered if different arrangements make a difference in terms of participant's answering the question correctly, but found no correlations (Spearman $\rho = 0.05$, $p = 0.5$).

7.4. Discussion: Experiment 4

Our hypothesis was not confirmed, as adding interaction to the diagram did not help more participants to answer the question correctly. In fact, it appears to have made the problem more difficult, possible due to cognitive overload (Mayer et al., 2001). This is a surprising result as we expected interaction to be beneficial. This may only have occurred due to the high level of difficulty of the Mammography Problem in the first place, narrowing the band of usable cognitive effort, and calling for future investigation of the effects of interaction on cognitive load.

Also interesting is the correlation between dragging mistakes and incorrect answers. While it is not clear if the mistakes are due to poor understanding of the problem, these mistakes could serve as an opportunity to hint to participants which parts of the question they need to focus on to better understand the scenario.

It may be that a higher degree of Pre-training is needed before conveying the whole problem. That is, a greater focus on the components may be needed, including the concepts, subsets, and nested-set inclusion.

8. DISCUSSION

In summary, four experiments were performed to evaluate the application of the Principles of Instruction to the text of the Mammography Problem. The results overview section reported Chi-square tests and effect sizes between conditions in the experiments (Figure 2) showing moderate effect sizes when the problem text was improved and when the static double-tree diagram was added to both new and original text. Figure 14 shows Chi-square tests and effect sizes across experiments to better convey the effect of the interventions. It is important to note that doing so may increase the possibility of observing effects of chance, since different experiments were performed at different times. Nevertheless, given a large degree of variation between all the conditions ($\chi^2(7, N=749) = 58.8$, $p < 0.001$), doing a further post-hoc analysis may be useful.

Sec.	Exp.	Condition	N	Exact	Exact (%)	χ^2 (df=1)	φ	Odds Ratio
8.1	1	Text-only	92	5	5%	20.4***	0.34	8.3
	2	Text-diagram	98	32	32%			
8.2	1	New Text-only	91	15	16%	23.4***	0.36	5.3
	2	New Text-diagram	99	50	51%			
8.3	1	New Text-only	91	15	16%	8.8*	0.23	3.0
	3	Interactive Focus	95	35	37%			
8.4	2	New Text-diagram	99	50	51%	4.7*	0.17	0.5
	3	New Text-diagram + Interactive Focus	93	32	34%			
8.4	2	New Text-diagram	99	50	51%	4.3*	0.15	0.5
	4	Interactive Diagram	89	31	35%			

Figure 14: Overview of post-hoc tests. Column Sec. gives reference to the section the itmes are discussed. Column N shows total number of participants, column Exact shows the number of participants that answered correctly and column Exact (%) shows the number as a bargraph as a percentage of N. Error bars in the Exact (%) column represent 95% confidence interval. Column χ^2 shows Chi-square test with Yates' continuity correction, * indicated statistical significance with $p < 0.001$, * indicated statistical significance with $p < 0.05$. Column ϕ shows phi coefficients.**

8.1. Text-diagram versus Text-only

Even with the original text of the problem, a large effect size was found in adding the diagram to the problem text ($\chi^2(1, N=191) = 20.4$, $p < 0.001$, $\phi = 0.34$, the odds ratio is 8.3). Notably, in the text-diagram condition, participant performance was 12 percentage points higher in our run of this condition compared to previous work (Khan et al., 2015). This may be due to the Principles of Instruction used in the condition design or graphical differences in the diagram design.

8.2. New Text diagram versus New Text-only

The largest effect size is found when adding the double-tree diagram to the new text condition ($\chi^2(1, N=189) = 23.4$, $p < 0.001$, $\phi = 0.36$, the odds ratio is 5.3) indicating a good balance between the burden and the benefit of interpreting the diagram. Clearly,

transforming the text, both in the problem text and within the diagram, was the most beneficial to participants. Even though previous work (Khan et al., 2015; Micallef et al., 2012) indicated that visualization did not have a reliable effect on performance, the large effect size in this work clearly indicates that the addition of visualization can dramatically help participants in understanding and correctly answering the problem when the text is clarified.

8.3. Interactive Focus versus New Text-only

INTERACTIVE FOCUS applies the principles of Pre-training, Interactivity, Self-explanation, Segmenting, and Signaling. As this adds low-level interactivity (C. Evans & Gibbons, 2007), we expected that participants will be more engaged and have improved learning. The addition of the interactive focus task, using the new text, had a moderate effect size compared to NEW TEXT-ONLY ($\chi^2(1, N=186) = 8.8$, $p < 0.05$, $\phi = 0.23$, the odds ratio is 3.0). This indicates that simple interaction tasks can have a pronounced positive effect on performance. In this case, the rate of correct answers was more than doubled.

8.4. Interactive conditions versus New Text diagram

Surprisingly, adding dragging interaction to the static tree diagram did not help participants answer the question correctly. Participants in both the INTERACTIVE DIAGRAM (35%) and the INTERACTIVE DIAGRAM + FOCUS (34%) conditions performed worse than the NEW TEXT-DIAGRAM (51%) condition. The difference is statistically significant in both cases (see Figure 14). However, while some forms of interaction can be beneficial, others may create additional cognitive overload in understanding the problem. The hypothesis of cognitive overload as a cause may be supported by two factors. First, the interactive tasks were quite different in nature from each other yet posed the same cost in performance. Second, the performance achieved in these conditions matches the performance in other conditions (approximately 35%) indicating some broader cognitive limit may be at play. In future work, we intend to directly map these results to relevant parts of the Cognitive Theory of Multimedia Learning (Mayer, 2005) and design further experiments to elucidate the role of the theory as applied to instruction for decision making.

8.5. Principles of Multimedia Instruction

We applied the instructional principles of coherence, personalization, signaling, segmenting, multimedia, spatial contiguity, and pre-training. All but one of these principles seemed to contribute to improved performance. Only Pre-training was not effective, as used in the INTERACTIVE DIAGRAM task. This may be due to several factors. First, the original application of pre-training was using narration which, according to the theory, takes advantage of dual-channel processing, unloading the visual channel and employing the auditory channel for pre-training. Second, we may not have sufficiently separated the parts from the whole. That is, we did not introduce each object and concept separately before introducing the entire problem. This direction may be beneficial to investigate in future work. Finally, there may be many other ways to explore pre-training in Bayesian inference and the development of a design space may help to better explore this area in a systematic fashion.

9. CONCLUSION

Bayesian inference problems are important in understanding the reasoning process in critical decision making. In addition to the use of frequency format and augmenting the problem with visualization, we have proposed the application of instructional science principles to the classic Mammography Problem. Considering the maturity of the Mammography Problem, the large improvements found in participant performance indicate the high value of considering the Principles of Instruction, Self-explanation, and Interaction in the design and understanding of Bayesian inference problems. We also showed the benefits of the focus format in the problem text and in an interactive task, and showed value in augmenting the problem with a double-tree visualization.

The use of the Principles of Instruction also creates a foundational link between cognition and Bayesian inference reasoning. While some work has examined cognitive effort in Bayesian inference (Ayal & Beyth-Marom, 2014; Girotto & Gonzalez, 2001), more work is needed to explain general participant performance outcomes. Additional investigation borrowing from instructional science may provide insights as has happened in the work presented here.

We have just scratched the surface in term of exploring the wide range of interactive tasks that may be explored. This work shows that benefits are available but also that performance costs are quickly added and that care must be taken to avoid overloading the participants. However, adding complexity until participants do get overloaded is also informative and indicative of underlying cognitive processes. The use of interactivity has also exposed measurable challenges in the numeric values themselves, through dragging mistakes. This may be another source of inherent problem difficulty that has not previously been discussed.

Some conditions provided significant benefits to participants while others did not. Finding the right balance between the burden and the benefit of a treatment will require a more refined understanding of the critical factors than we have today. Investigations into this class of problem have, so far, focused only on the benefits of certain interventions and have ignored any associated burdens. Ironically, our results indicate that a more Bayesian approach to investigating Bayesian inference problems is needed.

REFERENCES

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, *9*(3), 226–242. Retrieved from http://journal.sjdm.org/12/12714/jdm12714.html

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*(2), 284–289. doi:10.3758/PBR.15.2.284

Breslav, S., Khan, A., & Hornbæk, K. (2014). Mimic: visual analytics of online micro-interactions. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14* (pp. 245–252). New York, New York, USA: ACM Press. doi:10.1145/2598153.2598168

Calvillo, D., DeLeeuw, K., & Revlin, R. (2006). Deduction with Euler Circles: Diagrams that Hurt. In *Diagrams* (pp. 199–203). doi:10.1007/11783183_27

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, *299*(18), 999–1001. doi:10.1056/NEJM197811022991808

Cole, W. G. (1989). Understanding Bayesian reasoning via graphical displays. In *Proceedings of the SIGCHI conference on on Human Factors in Computing Systems* (pp. 381–386). New York, New York, USA: ACM Press. doi:10.1145/67449.67522

Dix, A., & Ellis, G. (1998). Starting simple: adding value to static visualisation through simple interaction. *Proceedings of the Working Conference on Advanced Visual Interfaces*, 124–134. Retrieved from http://dl.acm.org/citation.cfm?id=948514

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, *26*(5), 1024–1033. doi:10.1016/j.chb.2010.03.003

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 2399). New York, New York, USA: ACM Press. doi:10.1145/1753326.1753688

Eddy, D. (1982). Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. *Judgment Under Uncertainty: Heuristics and Biases*, 249–267. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Probabilistic+reasoning+in+clinical+medicine:+Problems+and+opportunities#0

Evans, C., & Gibbons, N. J. (2007). The interactivity effect in multimedia learning. *Computers & Education*, *49*(4), 1147–1160. doi:10.1016/j.compedu.2006.01.008

Evans, J. S., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. a. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*(3), 197–213. doi:10.1016/S0010-0277(00)00098-6

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. doi:10.1037//0033-295X.102.4.684

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247–276. doi:10.1016/S0010-0277(00)00133-5

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of Visualization in the Mammography Problem. *International Journal of Human-Computer Studies*. doi:10.1016/j.ijhcs.2015.07.001

Khan, A., Matejka, J., Fitzmaurice, G., & Kurtenbach, G. (2005). Spotlight: Directing Users' Attention on Large Displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '05*, 791–798. doi:10.1145/1054972.1055082

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, *5*, 1144. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4191302&tool=pmcentrez&rendertype=abstract

Mayer, R. E. (2005). Cognitive Theory of Multimedia Learning. In R. E. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning* (pp. 31–48). New York, New York, USA: Cambridge University Press. Retrieved from http://etec.ctlt.ubc.ca/510wiki/Cognitive_Theory_of_Multimedia_Learning

Mayer, R. E. (2008). Applying the science of learning: evidence-based principles for the design of multimedia instruction. *The American Psychologist*, *63*(8), 760–9. doi:10.1037/0003-066X.63.8.760

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*(1), 187–198. doi:10.1037//0022-0663.93.1.187

Micallef, L., Dragicevic, P., & Fekete, J.-D. (2012). Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2536–2545. doi:10.1109/TVCG.2012.199

Ottley, A., Metevier, B., Han, P., & Chang, R. (2012). Visually Communicating Bayesian Statistics to Laypersons, 1–11. Retrieved from http://www.cs.tufts.edu/tech_reports/reports/2012-02/report.pdf

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

Paas, F., Van Gerven, P. W. M., & Wouters, P. (2007). Instructional efficiency of animation: effects of interactivity through mental reconstruction of static key frames. *Applied Cognitive Psychology*, *21*(6), 783–793. doi:10.1002/acp.1349

Reed, S. K. (2006). Cognitive Architectures for Multimedia Learning. *Educational Psychologist*, *41*(2), 87–98. doi:10.1207/s15326985ep4102_2

Tsai, J., Miller, S., & Kirlik, a. (2011). Interactive Visualizations to Improve Bayesian Reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 385–389. doi:10.1177/1071181311551079

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. Retrieved from http://www.sciencemag.org/content/211/4481/453.short

Wang, P.-Y., Vaughn, B. K., & Liu, M. (2011). The impact of animation interactivity on novices' learning of introductory statistics. *Computers & Education*, *56*(1), 300–311. Retrieved from http://www.sciencedirect.com/science/article/pii/S0360131510002034

Wylie, R., & Chi, M. (2005). The Self-Explanation Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (pp. 413–432). Cambridge Press.

Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1224–31. doi:10.1109/TVCG.2007.70515