# Why Combining Text and Visualization Could Improve Bayesian Reasoning: A Cognitive Load Perspective

Melanie Bancilhon
Washington University in St. Louis
St. Louis, United States
mbancilhon@wustl.edu

AJ Wright
Washington University in St. Louis
St. Louis, United States
ajwright@wustl.edu

Sunwoo Ha
Washington University in St. Louis
St. Louis, United States
sha@wustl.edu

Jordan Crouser
Smith College
Northampton, United States
jcrouser@smith.edu

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

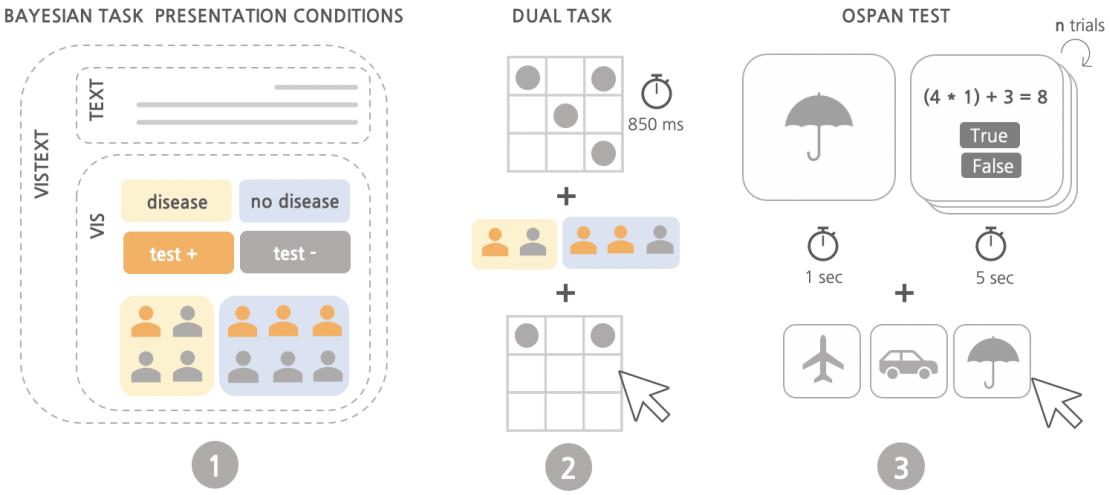


**Figure 1: An illustrative overview of our experimental design, where we used cognitive load theories to study the impact of presentation format on Bayesian reasoning. 1) Users were shown a Bayesian problem via: text-only (*text*), visualization-only (*vis*), or combined text and visualization (*vistext*) 2) We experimentally manipulated cognitive resources using a dual-task framework, asking participants to keep a 4-dot pattern in memory while completing the Bayesian task 3) We measured working memory capacity with an operation span (OSPAN ) task designed by [18], testing the ability to remember sequences of four, five, and six images while completing interspersed math problems.**

## ABSTRACT

Investigations into using visualization to improve Bayesian reasoning and advance risk communication have produced mixed results, suggesting that cognitive ability might affect how users perform with different presentation formats. Our work examines the cognitive load elicited when solving Bayesian problems using icon arrays, text, and a juxtaposition of text and icon arrays. We used a three-pronged approach to capture a nuanced picture of cognitive demand and measure differences in working memory capacity, performance under divided attention using a dual-task paradigm, and subjective ratings of self-reported effort. We found that individuals with low working memory capacity made fewer errors and experienced less subjective workload when the problem contained an icon array compared to text alone, showing that visualization improves accuracy while exerting less cognitive demand. We believe these findings can considerably impact accessible risk communication, especially for individuals with low working memory capacity.

## CCS CONCEPTS

• **Bayesian Reasoning** → **Visualization**; *Decision-Making*; • **Cognitive Load** → Multimedia; Evaluation.

## KEYWORDS

Decision-making, Bayesian reasoning, Perception and Cognitive Load

**ACM Reference Format:**
Melanie Bancilhon, AJ Wright, Sunwoo Ha, Jordan Crouser, and Alvitta Ottley. 2023. Why Combining Text and Visualization Could Improve Bayesian

*CHI '23, April 23–28, 2023, Hamburg*
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Reasoning: A Cognitive Load Perspective. In *Proceedings of reCHInnecting (CHI '23)*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Scholars have long studied the impact of multimedia formats on comprehension and performance in various settings. In psychology, for example, studies suggest that combining a diagram and text description provides more learning benefits than showing one or the other separately (e.g.,[13, 14]). Similarly, in education, scholars advocate for multimedia representations over singular formats [37]. However, the guidelines are not as clear-cut for visualization, even though combining text and visualization is ubiquitous in mass media storytelling, education, and health communication.

One area in visualization research where the efficacy of combining text and visualization is fraught with uncertainty is the communication of conditional probabilities. Conditional probabilities or Bayesian reasoning is necessary to communicate crucial statistical information to a broad audience, especially in medical decision-making. In particular, health officials need to express how often a test reports that a person has a virus when they do not (false positive). Additionally, patients need to understand their chance of having the disease given a positive test (true positive) to make informed decisions about risks and potential treatment. Still, extensive research shows that understanding conditional probabilities is challenging for novices and experts alike, even with multimedia representations [26, 32, 38, 73, 75].

One of the most important guidelines proposed to improve Bayesian reasoning accuracy is to show information in the form of natural frequency formats (e.g., *8 out of 10*) instead of percentages (e.g., *80%*) [32, 45, 56]. However, further investigations examining whether including visualization can improve Bayesian reasoning have produced mixed results. Early studies found that adding visualizations such as icon arrays to text formats can prompt faster and more accurate responses than text-only formats (e.g., [11]). More recent crowdsourced studies found that supplementing textual information with Euler diagrams increased accuracy only when numerical data were removed from the textual description [49], suggesting a potential conflict when presenting numbers and visualization together. Researchers have examined interaction techniques that link the text to the visualization but found no measurable benefit compared to static multimedia formats [51]. Other studies have shown that spatial ability is a mediating factor for accuracy and advocate for considering individual differences in visualization evaluation [56].

The research on Bayesian reasoning presentation extends beyond the visualization community and is more expansive than the few papers we have highlighted here. Yet, despite the extensive research, our knowledge is limited, partly due to over-reliance on coarse performance measures such as reasoning accuracy. We propose that other factors, such as the cognitive load elicited by different presentation formats, might provide an additional window into the mechanisms underlying how people use text and visualization to support Bayesian reasoning. Cognitive load is a measure of the effect that a particular task has upon the user's cognitive system [57]. It can impact user experience under various conditions, such as making decisions under stress or emotional burden [50], under divided attention [15, 16, 67], or with limited mental resources [2, 33].

We evaluate the cognitive load elicited by the icon array (*visualization-only*), text (*text-only*), and a combination of icon array and text (*combined*) using three different methods: a working memory capacity test, a dual task, and self-reported effort. We posit that measuring working memory capacity will provide insight into individual differences in users' cognitive abilities. Additionally, by burdening cognitive resources, the dual-task paradigm is a more direct method of measuring the impact of format on cognitive load and simulates real-world conditions where attention is divided. Finally, we captured perceived effort via a NASA-TLX questionnaire. These three methods together provide a comprehensive view of cognitive load.

By observing individual differences in working memory capacity, we found that individuals with low working memory made significantly fewer errors when using *visualization-only* compared to *text-only* formats. Furthermore, NASA-TLX scores show that users with low working memory capacity reported experiencing less temporal and physical demand using *visualization-only* and *combined* formats compared to text alone. Low working memory users also reported feeling less frustrated when using *combined* compared to *text-only*. Together, these provide supportive evidence that visualization elicits less cognitive load compared to text alone.

In summary, this paper documents the following contributions to the study of visualization-supported Bayesian reasoning:

(1) Using cognitive load, our findings offer a new perspective on the role of visualization for Bayesian reasoning. In particular, we found that **showing repeated information across text and visualization in combined formats could be beneficial**. We provide suggestive evidence that this enables people to select which formats better fit their mental model.
(2) We demonstrate that **individual differences in working memory capacity affect Bayesian reasoning** with different formats. This has implications for the use of visualization across a broad population (e.g. in medical decision-making) and adds a new dimension of complexity to the process of visualization recommendation.
(3) We demonstrate **how to use varying measures of cognitive load for visualization evaluation**, adding to the literature that calls for the diversification of evaluation measures by expanding beyond traditional performance metrics such as accuracy.

## 2 BACKGROUND

People are notoriously bad at reasoning with conditional probabilities [3, 38]. Consider, for example, the following scenario from [32]:

> *The probability of breast cancer in the population is 1% for a woman aged 40 who participate in a routine screening. If the woman has breast cancer, the probability is 80% that she will have a positive mammography. If a woman does not have breast cancer, the probability is 9.5% that she will also have 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?* _____

According to Bayes' theorem,

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \quad (1)$$

where, in our scenario, $D$ is the positive mammography and $H$ is the hypothesis that the woman in question has breast cancer. It is common for people, including experts, to be subject to *base-rate neglect*, ignoring the base rate $P(H)$ when reasoning about the true positive rate [38]. For decades, there have been efforts across various fields to devise ways to improve Bayesian reasoning by mitigating the base rate fallacy.

Several studies have shown that frequency formats (e.g., 8 out of 10 instead of 80%) can facilitate Bayesian reasoning and significantly improve accuracy [10, 23, 31, 32, 45]. Additionally, many researchers have investigated the effect of visualization on Bayesian reasoning (e.g., [20, 21, 41, 42, 69]), with the most prevalent designs being Euler diagrams [11, 39, 41, 49] and frequency grids or icon arrays [8, 30, 39, 41, 49, 55, 68, 71]. These designs represent two dominant theories behind Bayesian facilitation. Euler diagrams align with the *nested set theory*. They are useful to help the viewer reason about how subsets relate to each other[5, 45, 68, 72], while icon arrays, showing natural frequencies (i.e., 8 out of 10), align with the *ecological rationality framework* positing based on evolutionary theories that humans are better at reasoning with countable objects [23, 31]. Our work uses icon arrays because of their popularity and the well-documented success of natural frequency formats for Bayesian reasoning (e.g., [10, 23, 32, 49, 56]).

To investigate the potential benefit of visualization in Bayesian reasoning, researchers have typically compared responses to Bayesian problems presented in text format to formats that combine visualization and text. However, these studies have produced mixed findings. For example, Micallef et al. [49] found no measurable difference in accuracy between text alone and a combination of text and visualization. Still, their follow-up study demonstrated that removing the numbers from the text significantly improved Bayesian accuracy. Ottley et al. [54, 56] replicated this first study result and found no overall reliable differences in accuracy between the text alone versus a combined format. However, they found that participants with high spatial ability performed reliably better with visualization alone compared to text alone [56]. In another study, Ottley et al. [54] used eye-tracking to examine how people extract information from text-only, visual, and combined formats in Bayesian reasoning problems. They found that users easily identify information with visualization but extract information more easily from the text. Additionally, their analysis found no differences in how the study participants used each format when they saw the combined presentation. Finally, Mosca et al. [51] investigated the effect of linking the text and visualization via interaction. They found that adding interaction did not improve accuracy in Bayesian reasoning compared to static formats.

We posit that the outstanding questions on whether visual designs can improve Bayesian reasoning could be due to a lack of understanding of underlying cognitive mechanisms. Investigations by Lesage et al. [45] showed that performance in Bayesian reasoning is reliant upon available mental resources, regardless of presentation format. Although visualization researchers often seek to improve speed and accuracy measures, we know little about the impact of visualization on cognitive load. Moreover, speed and accuracy do not always correlate with cognitive load when reasoning about visualizations [36, 65]. Thus, there is a need to understand the processes that govern Bayesian reasoning with different presentation formats. In this paper, we expand the evaluation of Bayesian communication techniques by measuring cognitive load through individual differences in working memory capacity, a dual-task paradigm, and perceived cognitive load. We aim to develop a more nuanced understanding of the potential effect of presentation formats on Bayesian facilitation and provide more comprehensive visualization design guidelines.

## 2.1 Measuring Cognitive Load

Working memory consists of multiple components that can store a limited amount of information for a limited amount of time and is an essential resource in the reasoning process [24]. Cognitive load, typically defined as the amount of working memory required to process a task, is an important usability factor that indicates how easy or how hard it is to process information [57]. There exist numerous techniques to measure cognitive load, including self-reported measures (e.g. NASA-TLX), performance-based measures (e.g. dual-task paradigm, operation span tests) and physiological measures (e.g. pupillometry, fNIRS) [12, 25, 36, 40, 46, 52, 57, 59]. Several researchers have leveraged these techniques to investigate the effect of visualization design on cognitive load [1, 9, 18, 61, 70, 78], sometimes reexamining long-standing beliefs. For example, Matthews et al. highlight the importance of using several methods to cross-examine the effect of workload [48]. In their work, Borgo et al. challenged traditional notions about chart junk and showed using a dual-task paradigm that visual embellishments do not prompt higher cognitive load compared to other visualizations [9]. Peck et al. used fNIRS as well as NASA-TLX to evaluate visualization interfaces and found no difference in the cognitive load elicited by bar graphs and pie charts, contrarily to popular belief [61].

While physiological measures have proven to be effective techniques for measuring cognitive load, their high intrusiveness makes them unsuitable for real-life implementation [43]. Other measures are more accessible, facilitate longitudinal studies, and allow us to survey a diverse population. In our work, we chose to investigate the effect of presentation formats on cognitive load for Bayesian reasoning using three different methods: an operation task to observe individual differences in working memory capacity, a dual-task paradigm, and self-reported scores through a NASA-TLX questionnaire.

*2.1.1 Individual Differences Approach to Cognitive Load.* Individual differences can impact how we reason with different formats (see [47] for a comprehensive review of individual differences in visualization), and there is strong evidence that cognitive traits can influence statistical reasoning [45, 51, 56, 76]. Some researchers showed evidence that when information was presented in the form of natural frequencies, participants with high working memory capacity performed significantly better than participants with low working memory capacity [45, 76]. Castro et al. [17] have shown

that visualization designs can elicit different levels of cognitive load when reasoning about uncertainty visualizations.

A test that has shown high correlations with measures of working memory capacity is the Cognitive Reflection Test (CRT)[45]. The CRT test measures one's ability to overcome heuristics and biases and trigger analytical thinking [45]. A more direct way of measuring individual differences in working memory capacity is by using an operation span task (OSPAN ) [18]. In a typical OSPAN task, participants must simultaneously try to remember presented words in their correct order while solving simple math equations sequentially. In this paper, we use Castro et al.'s adapted online OSPAN test to measure working memory capacity [17] [1]. To complement this method, we use a dual-task paradigm, which according to Lesage et al. [45], can be used to infer a causal role for cognitive resources in the performance of Bayesian reasoning tasks.

*2.1.2 Dual-Task Paradigm.* Although it has not been prominently featured in visualization research, the *dual-task methodology* is an effective way to assess the dependency of a task on cognitive resources and has been used to evaluate workload in psychology for decades [19, 60]. In a dual-task paradigm, the user conducts two tasks simultaneously, a primary task and a secondary task. This creates *divided attention* and increases cognitive load, producing a decline in performance compared to the primary task alone. This decline is often referred to as the *dual-task cost* [58], which can be used to infer the cognitive load elicited by the task.

Several researchers have investigated the impact of formats on cognitive load using a dual-task paradigm [9, 16, 70], one reason being that it is helpful to simulate real-life conditions where attention is often divided [15, 16, 67]. Castro et al. [16] have used a dual-task method to investigate how display dimensions and screen size of mobile devices influence attention. In their study, participants controlled the movements of a blue ball by tilting the mobile device on displays of different sizes (primary task) while performing a change detection task which consisted of vocally reporting which of 4 arrows changed directions on a fixed display (secondary task). Using this methodology, they found that larger displays are more mentally demanding under divided attention. Tintarev et al. [70] investigated the effect of presentational choices for *planning* on cognitive load using a dual-task paradigm. Participants had to keep information about a list of words in memory while answering some questions about a plan, then had to recall the list of words in the correct order. The authors found no reliable differences in performance across different formats of the plan.

In our work, we quantify differences in elicited cognitive load across presentation formats using a dual-task methodology inspired by [45], consisting of remembering a pattern of four dots on a grid while conducting the primary task.

## 3 RESEARCH GOALS

We designed two complementary studies to investigate whether cognitive load can shed light on the conflicting and sometimes puzzling findings around Bayesian reasoning and visualization.

These findings collectively point to a potential relationship between cognitive resources and Bayesian facilitation — adding visualization and interaction to an already cognitively challenging task might not produce the desired effects. There is a gap in our understanding of how cognitive load affects Bayesian reasoning across different formats. Motivated by this, the current work focuses on examining the potential differences in cognitive load elicited by visualization-only, text-only, and a combination of text and visualization format in the context of Bayesian reasoning. We use the icon array for our visualization condition because it is prominently used to communicate Bayesian information, especially in the context of medical risk, supporting ecological validity.

When considering options for the experiment design, we weighed trade-offs between (1) controlling the framing and learning effects, (2) minimizing noise from individual variability, and (3) minimizing the overall length of the study. Unfortunately, no single experiment strikes the perfect balance. Thus, we present the results of two controlled user experiments. The first adopts a between-subject, 3 (*presentation format*) × 2 (*load condition*), experiment design to mitigate the learning effects that a within-subject study would introduce. The second utilizes a mixed design, with 3 (*presentation format*) between-subject and × 2 (*load condition*) within-subject protocol to better control for individual variability. Together, they tell a cohesive story about the relationship between cognitive load, Bayesian reasoning, and visualization.

## 4 EXPERIMENT 1: BETWEEN-SUBJECT STUDY DESIGN

We assigned each participant randomly to one of three presentation conditions — icon array (*vis*), text (*text*), and a combination of icon array and text (*vistext*) — making the comparison of presentation between subject. We also assigned each user randomly to either a **Single** or **Dual** task, making the comparison of these tasks also between subjects. We chose a between-subject design to keep the Bayesian problem consistent across all conditions. Prior work has shown that different Bayesian scenarios can lead to different levels of accuracy [49] [2].

### 4.1 Presentation Formats and Bayesian Task

We replicated Mosca et al.'s [51] grouped icon array design, which had the highest accuracy among their tested visualization formats. The authors designed the icon array according to Bertin's[6] guidelines, where 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 TEST POSITIVE versus TEST NEGATIVE. Participants in our the *text* condition saw the same data in textual format, and those in the *vistext* condition saw both the textual format and the icon array, vertically stacked.

We showed participants data about the prevalence of a disease in a population, as well as the test results in the form of either *vis*, *text* or *vistext*. We asked them to estimate i) the number of people who will test positive and ii) of those people, how many actually have the disease. This technique of prompting the user for the

[1] Link to OSPAN test used in this work (developed by [17]): https://bit.ly/2QHErIv

[2] Link to Experiment 1 surveys, data, and analyses: https://bit.ly/3BFwokx

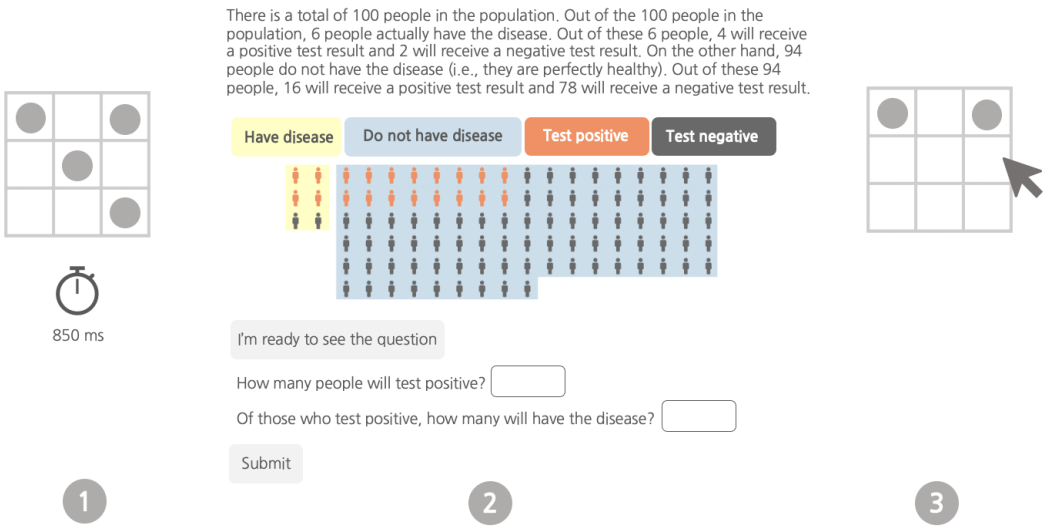


**Figure 2: An overview of the Bayesian survey for the Dual condition with the *vistext* format 1) Users were shown for 850 ms a pattern consisting of four dots on a 3x3 grid that they were asked to memorize 2) This is an example of the Bayesian task for the *vistext* condition. Users were asked to read the problem and then press a button when they were ready to answer questions 3) Once users submitted their answers to the Bayesian questions, they were asked to replicate the dot pattern on an empty 3x3 grid.**

positive count followed by the true positive count is called *probing*. *Probing* is a valid technique that evaluates Bayesian comprehension independently of mathematical skills through the retrieval of nested data (using the words "of those"). It has been shown to elicit more accurate responses compared to non-probed questions [11, 23, 56].

## 4.2 Load Conditions and Dual-Task Methodology

Participants either saw the Bayesian probability estimation task alone or along with a secondary task. Participants who were randomly assigned the Dual condition were shown a pattern consisting of four dots on a grid for 850ms and were asked to complete the Bayesian Probability Estimation task while keeping the pattern in memory. Participants were then asked to reproduce the dot pattern as accurately as possible by selecting the appropriate cells on an empty grid. Figure 1 illustrates the dual task setup, inspired by Lesage et al.'s [45] study of text-only formats and originally developed by Bethell et al. [7]. This task is appropriate as it taxes visuospatial working memory, which would possibly interfere with the primary task and cause the desired increase in cognitive load.

## 4.3 Measures of Abilities and Surveys

The survey also contained a NASA-TLX questionnaire, a spatial ability test, and a Cognitive Reflection Test (CRT). Participants then completed a working memory capacity questionnaire from [18].

**NASA-TLX.** We used the NASA-TLX [34, 35] to examine participants' subjective workload. Participants reported on the workload they believed the Bayesian task elicited on six subscales: mental demand, temporal demand, frustration, physical demand, performance, and effort.

**Working Memory Capacity Test (OSPAN).** We asked participants to remember a series of objects sequentially while answering simple True or False math problems. The test consisted of 6 sequences of 4-, 5- or 6- spans, shown two times each in a randomized order (the term *n*-span refers to the sequence occurring *n* times). In each span, participants were shown an image for 1 second and were asked to keep it in memory while answering a simple math question in under 5 seconds. This sequence is repeated *n* number of times and at the end of the span, participants have to recall the images shown in the correct order. This version of the OSPAN has been designed by Castro et al. [16].

**Cognitive Reflection Test.** The Cognitive Reflection Test (CRT) has been shown to be a valid measure of cognitive load [45]. In our work, we use a version of the CRT test that contains 3 questions. It tests for the ability to switch from Type 1 (intuitive) to Type 2 (strategic) reasoning. Since the latter requires using working memory [38], researchers posit that someone who is able to perform the switch has a high working memory capacity [59].

**Spatial Ability Test.** A spatial ability test measures an individual's capacity to process visual and spatial information. In this study, we used the paper folding test (VZ-2) from Ekstrom, French, and Hardon [27] consisting of two sessions of 3 minutes and 10 questions each. This test has been used as a standard technique to evaluate Bayesian reasoning performance across spatial ability in other studies [39, 49].

**VisText Usage Report.** We asked participants in the *vistext* condition what percentage of the visualization and the text they utilized to answer the Bayesian questions. They reported their

preferred method by selecting the appropriate value on a scale ranging from *only text* to *only visualization*

## 4.4 Hypotheses

**H1** We hypothesize that performance on the Bayesian reasoning task depends on available cognitive resources. Therefore, the **Single** condition will result in more accurate reasoning than the **Dual** condition.

**H2** Since available cognitive resources are mediated by working memory capacity, we expect that individuals with high working memory capacity will be more accurate than their low working memory counterparts, especially in the **Dual** condition.

**H3** Prior work that examined the impact of text-only, icon array, and the juxtaposition of text and icon array on Bayesian reasoning found no significant difference in accuracy between the three presentation formats [49, 56]. Therefore, we anticipate no significant difference in Bayesian reasoning accuracy across *vis*, *text*, and *vistext* in the **Single** condition.

The detailed analysis for pre-registered hypotheses H4a - H4d can be found in the supplementary material.

## 4.5 Participants

We recruited users via Amazon's Mechanical Turk that were from the United States, were English-speaking, and had a HIT acceptance rate of 100%.

**Payment.** All participants were paid in accordance with minimum wage laws, on average receiving $4.84 and taking 25.2 minutes to complete both surveys. In the Bayesian task, participants won a bonus of $0.50 for each question answered. Participants in the **Dual** condition were assigned an additional task, increasing the amount of time spent on the task, and thus received an additional $0.25 for each dot correctly remembered (i.e. up to $1 additional bonus compared to the *single* condition). The allocated bonus per dot remembered also served as an incentivization to remember the pattern.

We conducted a statistical power analysis using the software G*Power on a mixed ANOVA and determined that the target sample size needed for a statistical power of 95% is 251. We recruited 450 participants due to the typically high number of exclusions in Mechanical Turk studies. Users were asked to complete two separate surveys: the Bayesian survey and the OSPAN survey. Our pre-registered exclusion criteria [3], determined based on prior work, required that users i) take the surveys only once ii) complete both surveys iii) score above chance in the math portion of the OSPAN test iv) score above 10% in the memory portion of the OSPAN test v) score over 2 standard deviations from the mean in the dot pattern task. After excluding data that did not fit the exclusion criteria, 316 participants remained. After preliminary data analysis, we noticed some additional fraudulent and invalid responses that we had not anticipated prior to the pre-registration. We decided to exclude users who entered more than 4 dots in the dot pattern recall test, thus biasing their odds of getting the correct pattern (n=13). We also excluded participants whose answers to the Bayesian questions were less than or equal to 0 (n=4), which shows a lack of attention and leads to an invalid ERROR value upon data processing (see section 4.6). We conducted a post hoc sensitivity analysis that showed that the addition of the two exclusion criteria did not affect the study results (see supplementary material). After these non-pre-registered exclusions, 299 participants remained, of which 104 were assigned *text*, 100 were assigned *vis*, and 95 saw the *vistext* (129 in the **Dual** condition and 170 in the **Single** ).

[3]Link to Experiment 1 pre-registration: https://bit.ly/3xtC1zX

## 4.6 Data Collection

The independent variables for this experiment are:

- **3 presentation formats:** { *text*, *vis*, *vistext*}
- **2 load conditions:** { **Single** , **Dual** }

To measure Bayesian performance we calculated the true positive rate from the participant's response as described in subsection 4.1. Our dependent variables were:

- **EXACT** $\in \{0, 1\}$, binary value for whether the response was exact.
- **BIAS** is the $log_{10}$ ratio of the response and the ground truth.
- **ERROR** is the absolute value of bias.

While EXACT evaluates verbatim comprehension, BIAS and ERROR are proxies for gist (approximate) comprehension, which is more prominently used for reasoning and decision-making [4, 29, 59, 63, 64].

The covariates and other computed measures were:

- **OSPAN** $\in [0...30]$, measures general cognitive capacity.
- **WMC** $\in \{low, high\}$, based on a median split of OSPAN scores.
- **NASA-TLX** $\in [0...20]$, measures combined subjective workload.
- **Spatial Score** $\in [-4...20]$, is the spatial ability test score.
- **Spatial Level** $\in \{low, high\}$, from a median split of spatial scores.
- **CRT** $\in \{0, 1, 2, 3\}$, is the cognitive reflection test score.
- **Text-Vis Usage** $\in [1...20]$, maps 0 to using primarily text and 20 to mostly visualization for those in the *vistext* condition.

## 4.7 Attrition Analysis

There has been a growing body of work about the issue of high attrition rate in online studies [44, 62, 77]. According to research by Zhou et al. [77], studies that are cognitively taxing should be concerned if dropout rates are 20% or above. The authors also highlight the importance of checking for selective attrition by making sure

**Table 1: Experiment 1 condition-wise dropout rates for the Bayesian Task**

| Load condition | Dropout rates |
|---|---|
| **Single** : Participants conducted the Bayesian Task | 3.98% |
| **Dual** : Participants conducted the Bayesian Task and a secondary recall task | 7.18% |

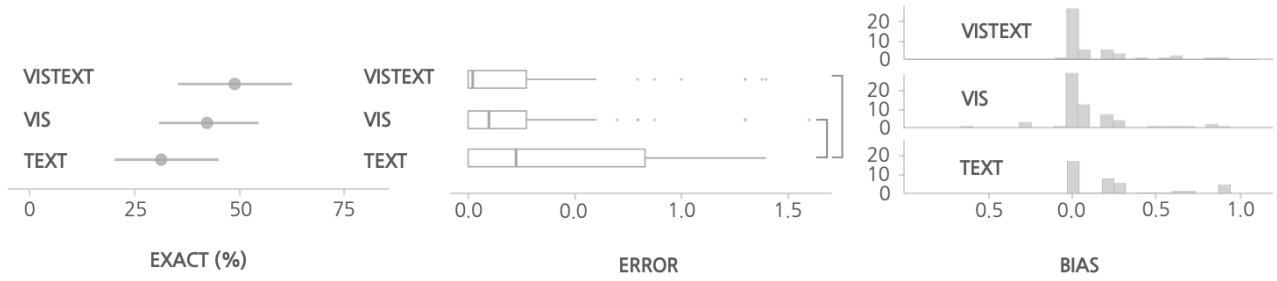


**Figure 3: Single task EXACT (95% CI), ERROR and BIAS across presentation formats. ] indicates a significant difference between the two formats ($\alpha = 0.0167$). We found significant differences in ERROR between *vis*and *text*, and *vistext* and *text*.**

the dropout rates are not significantly different across experimental conditions. To provide transparency and encourage practices that improve internal validity, we conducted an attrition rate analysis as recommended by Zhou et al. [77].

Our experiment consists of two surveys, a Bayesian Task implemented by the authors and an OSPAN test from [18] on Qualtrics. We conducted an attrition analysis for the Bayesian task, where participants were assigned either a single task (**Single** ) or a dual task (**Dual** ). We adapted our methodology from Zhou et al. [77] and only took into account participants who consented to the study and discarded fraudulent responses where participants took the survey more than once by using their recorded IP addresses. Table 1 shows the condition-wise dropout rates, computed according to [77] by dividing the number of participants who were assigned to a given condition and completed the entirety of their task [4] by the number of those who were assigned to the same condition who at least gave their consent and only took the task once. We observe low dropout rates for both conditions that have no significant difference ($\chi^2(2) = 1.88, p = 0.1704, d = 0.1406$).

## 4.8 Findings

Out of 299 participants, 104 were assigned *text*, 100 were assigned *vis*, and 95 saw the *vistext*. Each participant completed a single Bayesian problem depicting the *disease* scenario in subsection 4.1. Further, 170 were assigned to the single task (**Single** ) condition and 129 were assigned the dual-task condition with an added load (**Dual** ).

### 4.8.1 *Single Task: Establishing a baseline.*

We begin our analysis by inspecting how participants performed under the single task (**Single** ) condition and testing whether format influence performance. The existing literature has produced mixed results on the effect of visualization on reasoning accuracy [49, 54, 56], and our **H3** posits no significant difference in Bayesian reasoning accuracy.

*BIAS .* We conducted an exploratory analysis by examining how much participants' responses deviated from the EXACT answer and the effect of format on their discrepancy. We observe an overall median BIAS of .10 for the single task condition with varying median BIAS of 0.22 for *text*, 0.00 for *vis*, and 0.00 for *vistext*. From Figure 3, we can observe that participants' BIAS are not normally distributed. Thus, we use non-parametric tests for our analysis. Additionally, participants in the *vis* and *vistext* conditions were marginally more likely to produce the EXACT answer (BIAS = 0) than those who used *text*. When we ran a 3-way Kruskal-Wallis test with presentation format as a between-subject factor we found a significant difference in BIAS across the three conditions ($H(2) = 12.87, p = .0016, \eta^2(H) = 0.065$). Follow-up Mann-Whitney Wilcoxon tests with an adjusted alpha $\alpha = 0.0167$ revealed significant differences in BIAS between *vis* and *text* ($W = 2379.5, p = 0.0006, \eta^2(H) = 0.092$) as well as *vistext* and *text* ($W = 1772.5, p = 0.0087, \eta^2(H) = 0.057$).

*EXACT.* We examined our first measure of accuracy, EXACT , to investigate whether the presentation format influences the proportion of correct answers. Overall, 40.9% of participants correctly answered both Bayesian questions for the single task condition, with *text*, *vis*, and *vistext* yielding 31.5%, 43.08%, and 49.02% exact answers respectively. Our omnibus proportion z-test shows no significant effect of presentation format on accuracy ($\chi^2(2) = 3.4899, p = .1795$). Thus, *the proportion of successful exact reasoning did not depend on presentation format.*

*ERROR .* For a more fine-grained measure of accuracy, we examined ERROR to assess how far participants' responses deviated from the EXACT answer and whether presentation format mediated this effect. The median ERROR was 0.097 overall and 0.097 in the *vis* condition, 0.22 in the *text* condition, and 0.021 *vistext* condition. A Kruskal-Wallis non-parametric test revealed significant differences in ERROR between conditions ($H(2) = 8.43, p = 0.0148, \eta^2(H) = 0.037$). Post-hoc Mann-Whitney Wilcoxon tests with an adjusted alpha ($\alpha = .0167$) revealed significant differences between *vis* & *text* ($W = 2227.5, p = .0093, \eta^2(H) = 0.0047$) and *vistext* & *text* ($W = 1738.5, p = .0164, \eta^2(H) = 0.046$). We found no significant difference between *vis* and *vistext* ($W = 1610.5, p = 0.675$). These findings suggest that *reasoning with text-only led to significantly higher errors compared to other formats.*

Altogether, these findings show evidence that presentation format can impact reasoning errors. However, the observed effects

[4] our server recorded an end-of-experiment timestamp when a participant completed the entire survey

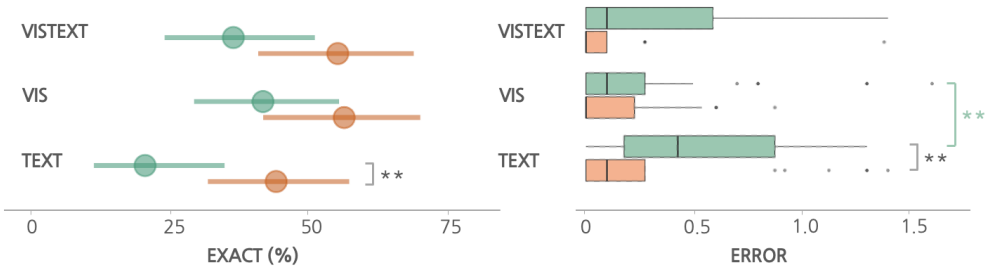


**Figure 4: Single task EXACT (95% CI) and ERROR across presentation format and working memory group. ** indicates a significant difference between groups. We found significant differences in EXACT and ERROR between Low and High working memory groups ($\alpha = 0.05$) in the condition. Among Low working memory capacity individuals, ERROR was significantly higher in *text* compared to *vis* ($\alpha = 0.0167$)**

were small and there was no significant impact on EXACT response rates. Thus, **our results only partially support H3**. More specifically, they suggest that visualization, even when combined with text, can have benefits on Bayesian accuracy compared to text alone. It is noteworthy that these results also partially contradict the visualization literature that compared Bayesian formats. On one hand, our findings are similar to Ottley et al. [54, 56] who found no difference in EXACT between *text*, *vis*, and *vistext*, but did not examine ERROR . On the other hand, our results differ from Micallef et al. [49] who examined *text* and *vistext* and found no measurable effect of these formats on EXACT or ERROR . However, the discrepancies between our results and prior work could be attributed to the type of visualization used and differences in the experiment design.

#### 4.8.2 *Individual Differences in Working Memory Capacity.*

A primary goal of this project is to examine whether cognitive resources can explain Bayesian reasoning results. Specifically, with **H2**, we hypothesized that accuracy in Bayesian reasoning will depend on available cognitive resources. To this end, we examine the effect of working memory capacity on accuracy in the Single task.

To examine whether working memory mediates accuracy in participants' EXACT and ERROR measures. We first performed a binary logistic regression to test for the effect of OSPAN on EXACT and found that correctly answering the Bayesian questions is 1.49 times more likely to occur for every 5-point increase in the working memory test (95% CI [.04, .12]). Analyzing ERROR , a generalized linear model also revealed a significant impact of OSPAN on ERROR ($t(169) = -3.326, p = 0.00108$). Thus, *the higher their working memory capacity, the more accurate participants were in their answers.*

Following prior work [18], we split participants into Low and High working memory groups based on a median split of their OSPAN scores. Figure 4 summarizes the accuracy of each working memory group across presentation formats, showing their respective proportions of EXACT answers and ERROR distribution. Overall, in the Single task, 52.29% of those in the High group produced EXACT answers compared to 34.59% in the Low group. Additionally, Low had a median ERROR of 0.176 and High had a median ERROR of 0. Consistent with the regression analysis, we show a statistically significant difference between the Low and High groups when we compared EXACT ($\chi^2(1) = 6.4762, p = 0.0109, d = 0.3980$) and ERROR (Kruskal-Wallis, $H(1) = 6.8535, p = 0.0088, \eta^2(H) = 0.0348$). Together, these results support **H2**, showing *suggestive evidence that participants' working memory mediated Bayesian reasoning accuracy.*

#### 4.8.3 *Working Memory Capacity & Presentation Formats.*

In light of our previous finding that successful reasoning might depend on cognitive resources, we conducted further analysis to examine the effect of presentation format on reasoning accuracy within the Low and High groups. Specifically, we ran separate 3-way proportion tests to compare the frequencies of EXACT answers and found no difference between presentation formats for both Low ($\chi^2(2) = 4.8401, p = 0.0889$) and High ($\chi^2(2) = 0.6504, p = 0.7224$) groups. Further, a Kruskal-Wallis test comparing ERROR for *text*, *vis*, and *vistext* within the Low group revealed a statistically significant difference between the three presentation formats ($H(2) = 10.086, p = 0.006453, \eta^2(H) = 0.0817$). We ran pairwise Mann-Whitney Wilcoxon tests with an adjusted alpha ($\alpha = 0.0167$) and found significant differences in ERROR between *text* and *vis* ( $W = 874, p = .0017, \eta^2(H) = 0.1291$), but failed to reject the null hypothesis for the *text* & *vistext* and *vis* & *vistext* comparisons. Examining the High group, a Kruskal-Wallis test found no overall significant differences between presentation formats ($H(2) = 1.7301, p = 0.4210$). These analyses suggest that *presentation choices can impact users with low working memory capacity*, with *text* eliciting significantly higher error rates compared to *vis*. However, the *high working memory capacity participants were less impacted by the format they used.*

Our final analysis here investigates how Low and High groups performed within each presentation condition. Our analysis revealed that the Low and High groups had similar proportions of EXACT ($\chi^2(1) = 1.2013, p = 0.2731$) answers and ERROR ($W = 428.5, p = 0.4381$) rates when reasoning with *vis*. The two working memory groups also did not differ in EXACT ($\chi^2(1) = 1.5875, p = 0.2077$) and ERROR when using *vistext* ($W = 230, p = 0.1000$). However, we observed a statistically significant difference in EXACT ($\chi^2(1) = 5.889, p = 0.01524, d = 0.6997$) and ERROR with the *text* condition ($W = 235.5, p = 0.02481, \eta^2(H) = 0.0776$). Thus, *text* is marginally more likely to elicit a deviation in accuracy between Low and High compared to *vis* or *vistext*.

#### 4.8.4 *Dual Task: Reasoning Under Divided Attention.*

In **H1**, we posit that if we can experimentally manipulate executive

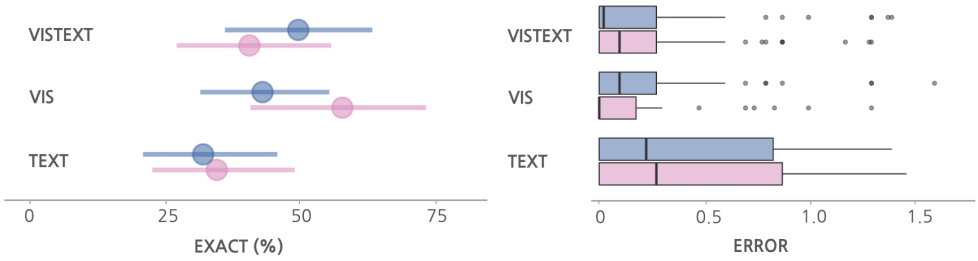


**Figure 5: EXACT (95% CI) and ERROR for Single and Dual task**

capacity by adding a secondary task, we will incur a decline in performance, known as the **dual-task cost**. As a result, formats that require high cognitive resources will have a significant dual-task cost.

*EXACT .* We observed a near-identical proportion of EXACT answers for the **Single** and **Dual** conditions. Participants in the **Dual** condition produced the EXACT answer 41.86% of the time, compared to 41.17% in the **Single** condition. We compared the proportion of EXACT answers in the **Single** and **Dual** task and found no overall significant differences ($\chi^2(2) = 0.0141, p = 0.9054$). The analysis revealed 34% of EXACT answers for *text*, 57.14% for *vis*, and 38.64% for *vistext*. A 3-sample proportion test found no significant difference in EXACT between the presentation formats in the **Dual** group ($\chi^2(2) = 4.816, p = .09$).*Thus, manipulating load had no significant effect on our participants' EXACT responses.*

*BIAS .* A Kruskal-Wallis test found no significant difference in BIAS between presentation formats in the **Dual** group ($H(2) = 4.44, p = 0.1084$). We compared overall BIAS for the **Single** and **Dual** conditions using a Mann-Whitney Wilcoxon test and found no significant difference between the two conditions ($W = 11130, p = 0.8175$).

*ERROR .* Finally, we also observed an identical overall median ERROR of 0.097 for both the **Single** and **Dual** task. An overall comparison with a Mann-Whitney Wilcoxon test found no significant difference in ERROR between **Single** and **Dual** ($W = 11232, p = 0.709$). The median ERROR was 0.27 for *text*, 0 for *vis*, and 0.097 for *vistext* in the **Dual** task. Similar to the **Single** condition, an omnibus Kruskal-Wallis test revealed an overall effect of presentation format on ERROR in the **Dual** condition ($H(2) = 7.7344, p = .0207, \eta^2(H) = 0.0455$). Follow-up Mann-Whitney Wilcoxon tests with an adjusted alpha (($\alpha = .0167$) revealed significant differences in ERROR between *text* and *vis* ($W = 1162.5, p = .0073, \eta^2(H) = 0.0747$). We found no significant difference between *text* and *vistext* ($W = 1277.5, p = 0.1674$) and *vis* and *vistext* ($W = 612.5, p = 0.1002$).

**Considering differences in working memory capacity.** In section 4.8.2, we showed evidence that working memory capacity impacts Bayesian reasoning. Here, we examine the difference in performance between the **Single** and **Dual** conditions by taking into account individual differences in working memory capacity. For individuals in the **High** group, we found no significant difference between those in the **Single** and **Dual** task conditions when examining EXACT ($\chi^2(2) = 0.4258, p = 0.5141$) and ERROR ($W = 3104, p = 0.3264$). Similarly, we found no measurable difference between the **Single** and **Dual** conditions for participants in the **Low** groups when examining EXACT($\chi^2(2) = 0.0701, p = 0.7912$) and ERROR ($W = 2467, p = 0.5253$) Taken together, the secondary task did not elicit the expected results and the evidence for **H1** is inconclusive.

Although in section 4.7 we found no significant difference in attrition rate between the **Single** and **Dual** tasks, we conducted a Kruskal-Wallis test to investigate whether the distribution of OSPAN scores varied between the two tasks after all data quality exclusions. We found a significant difference in OSPAN scores between **Single** and **Dual** ($W = 13688, p = 0.0002, \eta^2(H) = 0.0422$), with higher OSPAN scores in the **Dual** condition. This could be due to selective attrition or bias in our sample. Participants in the **Dual** condition had significantly higher OSPAN scores compared to the **Single** task. This could also explain why we did not observe a significant decline in performance between the two tasks. We will consider this confounding factor in our interpretation of Experiment 1's results.

#### 4.8.5 *NASA-TLX Self-Reported Effort.*

When looking at self-reported effort in the **Single** task, we found an overall significant difference in perceived **frustration** across presentation formats (Kruskal-Wallis, $H(2) = 11.72, p = 0.003, \eta^2(H) = 0.0582$). We conducted separate Mann-Whitney Wilcoxon tests with an adjusted alpha $\alpha = 0.0167$ for pairwise comparisons that revealed a significant difference in **frustration** between *vistext* and *text* ($W = 1918.5, p = 0.0005, \eta^2(H) = 0.1078$).

Since working memory capacity is likely to affect reported NASA-TLX scores, we observed differences between presentation formats for each working memory group separately. We found no significant difference between presentation formats across any of the NASA-TLX subscales in the **High** working memory group. Within the **Low** group, we conducted separate Kruskal-Wallis tests and found significant differences in accuracy between presentation formats in the following:

- *temporal demand:* ($H(2) = 10.305, p = 0.006, \eta^2(H) = 0.0839$)
- *physical demand:* ($H(2) = 8.95, p = 0.0114, \eta^2(H) = -0.0045$)
- *frustration:* ($H(2) = 10.825, p = 0.004, \eta^2(H) = 0.089$)

As a follow-up, we conducted Mann-Whitney Wilcoxon tests with an adjusted alpha ($\alpha = 0.0167$) within the **Low** group and found significant differences between *vistext* and *text* in the following:

- *temporal demand* ($W = 644, p = 0.004, \eta^2(H) = 0.1262$),
- *physical demand* ($W = 638, p = 0.005, \eta^2(H) = 0.1175$)

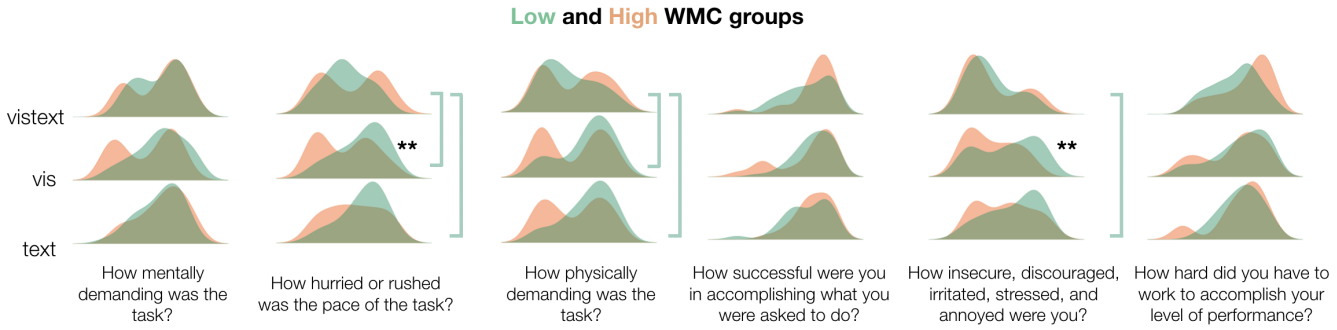


**Figure 6: Distribution of NASA-TLX scores in the Single task for the Low and High WMC groups. ** indicates a statistically significant difference between working memory groups ($\alpha < 0.05$) and ] indicates differences across formats for the corresponding WMC group ($\alpha < 0.0167$).**

- *frustration* ($W = 672.5, p = 0.0009, \eta^2(H) = 0.1712$)

Within the Low group, we also found differences between *vis* and *vistext* in the following:

- *temporal demand* ($W = 894.5, p = 0.006, \eta^2(H) = 0.0904$),
- *physical demand* ($W = 878, p = 0.011, \eta^2(H) = 0.08296$)

We investigated differences in reported NASA-TLX scores across High and Low groups for each presentation format. In the *vis* condition, we found differences in the following:

- *temporal demand* ($W = 307, p = 0.01525, \eta^2(H) = 0.0673$)
- *frustration* ($W = 332.5, p = 0.0379, \eta^2(H) = 0.0525$)

Finally, we found no differences in reported scores between working memory groups in the *text* and the *vis* conditions.

#### 4.8.6 *Additional Analyses.*

*Spatial Ability.* We conducted a generalized linear model with a logit link and found that spatial ability score had a significant impact on EXACT ($z(298) = 4.670, p = 3.00e{-}06$). We also examined the effect of spatial ability score on ERROR through a generalized linear model and found significant effects ($t(298) = -4.003, p = 7.91e-05$). These findings replicate prior work showing that spatial ability mediates Bayesian reasoning[51, 56].

*Completion Time.* Kruskal-Wallis tests revealed no significant effect of presentation format ($H(2) = 1.2454, p = 0.5365$) or load condition ($H(2) = 2.03, p = 0.1546$) on the completion time of the Bayesian task. Moreover, we found no significant difference in completion time between the Low and High working memory groups ($H(2) = 0.0797, p = 0.7778$).

*Cognitive Reflection Test.* Our CRT results largely replicated the OSPAN findings. We found an overall significant impact of CRT score on EXACT ($\chi^2(3) = 20.502, p = 0.0001, \eta^2(H) = 0.5246$). Further, the Kruskal-Wallis test shows a statistically significant effect of CRT on BIAS ($H(3) = 20.566, p = 0.0001, \eta^2(H) = 0.0595$) and ERROR ($H(3) = 24.986, p = 1.555e-05, \eta^2(H) = 0.0745$), showing evidence that *individuals with a higher CRT score were significantly more likely to enter the exact answers and made smaller reasoning errors.*

## 5 EXPERIMENT 2: MIXED DESIGN STUDY

Experiment 1 used a between-subject design to control for learning effects and ensured consistency by comparing responses to the same Bayesian problem. However, we found no significant effect of the dual task on accuracy. This could be due to the differences in working memory capacity between the two groups, or high individual variability due to the study design. We conducted a second mixed design study[5] to 1) control for individual variability in the single and dual tasks and 2) test whether the lack of replication is due to population or methodological differences.

We made the following changes to the experiment design to reduce the overall difficulty of the task and better control for individual variability.

- **Improve Study Preparation with a Practice Round**: We added a pre-study trial to familiarize participants with the task and study structure. Participants saw and attempted a sample Bayesian reasoning task before continuing to the main task.
- **Control Individual Variability**: We used a mixed factorial design with the load condition (Single , Dual ) as a within-subject factor and presentation format (*vis*, *text*, *vistext*) as a between-subject factor.
- **Remove CRT Test**: We removed the Cognitive Reflection Test from the survey as we found in Experiment 1 that it is positively correlated with the OSPAN test ($r(297) = 0.27, p = 2.051e^{-06}$), which is more widely recognized [22, 28, 53, 74]. This shortened the survey.

### 5.1 Task & Procedures

Our tasks were similar to Experiment 1 except that in the Bayesian survey, the users conducted both a single and dual task in no particular order where the Bayesian problems were presented using two scenarios, a *cab* and a *class* scenario. After solving the Bayesian problems, users completed a NASA-TLX and a spatial ability test. In this experiment, we did not conduct the Cognitive Reflection

[5]Link to Experiment 2 surveys, data, and analyses: https://bit.ly/3LfgXCZ

**Table 2: Scenarios use in the Bayesian task in Experiment 2**

| Scenario | Description |
|---|---|
| *cab* | There is a total of 100 witnesses to the car accident. Out of the 100 witnesses, 15 claimed that the car which caused the accident was a cab. Out of these 15 witnesses, 12 claimed the car was blue and 3 claimed the car was green. On the other hand, 85 witnesses claimed that the car which caused the accident was not a cab. Out of these 85 witnesses, 3 claimed the car was blue and 82 claimed the car was green. |
| *class* | There is a total of 100 college freshmen in the population. Out of these 100 freshmen, 30 are enrolled in an introductory entrepreneurship course. Out of these 30 freshmen, 20 plan on going into business after graduation, and 10 do not. On the other hand, 70 freshmen are not enrolled in an introductory entrepreneurship course. Out of these 70 freshmen, 10 plan on going into business after graduation, and 60 do not. |

Test. Similarly to Experiment 1, users also completed an OSPAN test.

**Practice Round**. In the practice round, users practiced the dot pattern recall task, the **Single** Bayesian task, as well as both tasks together as part of the **Dual** condition.

**Payment.** Participants received a base pay of $2 and could win a total bonus of up to $2.5, comprising of $0.5 for each correct Bayesian question and $0.5 for a correctly reproduced dot pattern. Participants received an average bonus of $1.51 and completed the Bayesian and OSPAN surveys in an average time of 26.8 minutes.

## 5.2 Experimental Design

Similarly to Experiment 1, we assigned each user randomly to one of three presentation conditions (*vis*, *text* or *vistext*), making the comparison of presentation between subjects. Each user completed both the **Single** and **Dual** tasks and saw either the *cab* or *car* scenario, making load condition a within-subject condition.

## 5.3 Presentation Conditions

Our presentation formats remained a between-factor condition and were the same as Experiment 1: *text*, *vis*, and *vistext*. We utilized the *disease* scenario for the pre-task tutorial, and each participant saw two Bayesian problems narrating two different scenarios: *cab* and *class* [32, 49, 54]. The *cab* scenario involves eye-witness testimonies of a hit-and-run scenario, while the *class* scenario presents the career prospects of college students. We randomly assigned one scenario to the **Single** task and the other to the **Dual** condition, and the order of the conditions was counterbalanced.

## 5.4 Participants

As per our pre-registration [6], we conducted a power analysis based on a three-way mixed ANOVA and determined the ideal sample size to be 168. We recruited 240 participants via Amazon's Mechanical Turk to account for a 30-40% exclusion rate. Participants were English-speaking from the United States and had a HIT acceptance rate of 100%. After excluding 88 participants based on the same pre-registered criteria determined in Experiment 1 (see section 4.5), 152 participants remained (*text*= 46, *vis*=55, *vistext*=51).

*5.4.1 Attrition Rate.* Using the same methodology as Experiment 1, we conducted an attrition rate analysis for Experiment 2. Table 3 shows the condition-wise dropout rate, showing participants who first saw the **Dual** task or the **Single** task. We found no significant difference in dropout rate between the two conditions ($\chi^2(2) = 1.6824, p = 0.1946, d = 0.1293$). When looking at performance in the OSPAN test after exclusions, we found no significant difference in scores. This suggests that the population who completed the experiment was consistent across both conditions ($H(2) = 0.34157, p = 0.5589, \eta^2(H) = -0.0048$).

**Table 3: Experiment 2 condition-wise dropout rates for the Bayesian Task**

| Task Order | Dropout rates |
|---|---|
| **Single** , **Dual** : Participants saw the dual task followed by the single task | 11.23% |
| **Dual** , **Single** : Participants saw the single task followed by the dual task | 15.67% |

## 5.5 Results

In this experiment, our aim is to uncover differences in dual-task costs elicited by each presentation format through a mixed-design study. First, we establish a baseline for accuracy in the single task and compare our findings to Experiment 1. Then, we examine and compare the decline in performance elicited by the dual task **(dual-task cost)** between presentation formats.

*5.5.1 Single Task.*

*EXACT .* Overall, 38.2% of participants correctly answered both Bayesian questions in the **Single** task, with *text*, *vis*, and *vistext* leading 21.7%, 47.3% and 42.3% EXACT answers respectively. Contrarily to Experiment 1, our analysis shows a significant effect of presentation format on EXACT ($\chi^2(2) = 7.58, p = .023, d = 0.4584$). Follow-up pairwise 2-sample proportion tests with an adjusted alpha ($\alpha = 0.0167$) revealed a significant difference in EXACT between *vis* and *text* ($\chi^2(2) = 7.1195, p = 0.0076, d = 0.5537$).

*BIAS .* We found no significant difference in BIAS between the presentation formats (Kruskal-Wallis, $H(2) = 1.0826, p = 0.582, \eta^2(H) = -0.0063$).

[6]Link to Experiment 2 pre-registration: https://bit.ly/3qIzJJu

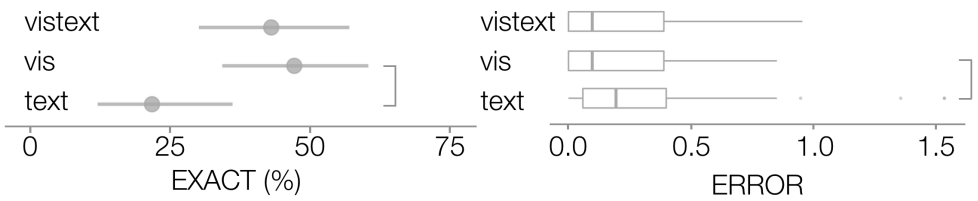


**Figure 7: Experiment 2. Single task EXACT , and ERROR across presentation formats. ] indicates a significant difference between the two formats ($\alpha = 0.0167$).**

*ERROR .* The median ERROR was 0.097 overall and 0.097 in the *vis* condition, 0.194 in the *text* condition and 0.076 in the *vistext* condition. A Kruskal-Wallis non-parametric test found a significant difference in ERROR between the presentation formats ($H(2) = 7.61, p = 0.0223, \eta^2(H) = 0.0376$). Post-hoc Mann-Whitney Wilcoxon tests with an adjusted alpha ($\alpha = 0.0167$) revealed a significant difference between *vis* and *text* ($W = 1619.5, p = 0.01329, \eta^2(H) = 0.05182$). Overall, the general trends are in line with Experiment 1 and demonstrate that *participants were the least accurate with text compared to visualization*. However, the differences are more pronounced in Experiment 2.

Prior work has shown that different Bayesian scenarios can have a different impact on accuracy [49]. We found no significant difference in accuracy between the *class* and *cab* scenarios when looking at EXACT ($\chi^2(2) = 0.0251, p = 0.8741, d = 0.0257$), BIAS ($W = 2870.5, p = 0.9727, \eta^2(H) = -0.0067$) or ERROR ($W = 2952.5, p = 0.7844, \eta^2(H) = -0.00616$).

### 5.5.2 *Single vs Dual Task.*

*DUAL-TASK.* We found that in the Dual task, the mean number of dots recalled was 3.56 ($\sigma = 0.77$). 40.8% of participants correctly answers both Bayesian questions, with 26.1% of EXACT answers for *text*, 41.8% for *vis* and 51.9% for *vistext*. Overall differences in EXACT between presentation formats were significantly different ($\chi^2(2) = 6.8197, p = 0.0331, d = 0.4335$). Follow-up pairwise comparisons revealed a significant difference in EXACT between *text* and *vistext* ($\chi^2(2) = 6.8003, p = 0.0091, d = 0.5461$), *suggesting that the combination of visualization and text leads to fewer errors than text alone under divided attention*. We found no significant difference in BIAS (Kruskal-Wallis, $H(2) = 2.3795, p = 0.3043$) or ERROR (Kruskal-Wallis, $H(2) = 4.451, p = 0.108$) between presentation formats.

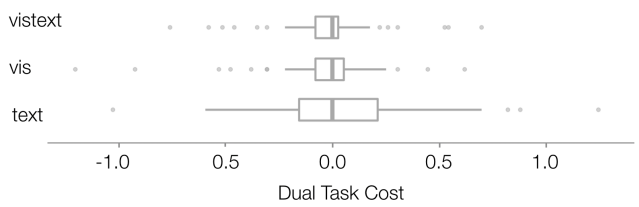


**Figure 8: Dual-task cost across presentation formats.**

*DUAL-TASK COST.* For each participant, we observed the decline in performance in the dual task compared to the single task by computing the difference in ERROR , a measure known as the **dual-task cost**. By comparing dual-task costs across presentation formats, we can infer differences in cognitive load. We conducted a Kruskal-Wallis test and found no overall difference in dual-task costs between presentation formats ($H(2) = 1.0314, p = 0.5971, \eta^2(H) = -0.0065$).

*CALIBRATING DUAL-TASK COST.* In section 4.8.2, we showed evidence that working memory capacity impacts Bayesian reasoning. To this end, we observe the effect of dual-task cost for High and Low working memory capacity groups. We found that dual-task costs were not significantly different between presentation formats within the High (Kruskal-Wallis, $H(2) = 1.4324, p = .4886, \eta^2(H) = -0.0089$) or Low (Kruskal-Wallis, $H(2) = .14215, p = .9314, \eta^2(H) = -0.0226$) group. *Therefore, we conclude that even when considering individual differences in working memory capacity, the effect of the dual-task was consistent across presentation formats.*

## 6 DISCUSSION

Our work leveraged cognitive theory to understand the conflicting findings about the effect of combining text and visualization in the context of Bayesian communication. Analyzing general trends in accuracy alone seldom paints the complete picture in an evaluation study, as there is ample research on the impact of individual differences on Bayesian reasoning and beyond [18, 51, 56, 76]. Our results suggest that combining visualization and text does not increase cognitive load and in some cases improves subjective workload. We present our main takeaways from these studies.

We analyzed accuracy to compare our findings to the prior work and to provide context for our cognitive load results. At a high level, our produced results are similar to prior studies in the visualization community [49, 51, 54, 56]. *Presentation format alone had little to no effect on Bayesian reasoning, but the inclusion of visualization improved Bayesian reasoning.* In Experiment 1, we found that users' proportion of correct answers in the baseline Single condition was not significantly different across the three formats, replicating Ottley et al.'s findings [56]. While user error rates were significantly lower using visualization-only compared to text-only, the effect size was small for the statistical test. In Experiment 2, we saw a significantly greater proportion of correct answers with the combined presentation format than with text alone, with a small effect. Still, although our accuracy analysis does a good job of uncovering differences, it does not explain the phenomena.

## 6.1 Implications of Cognitive Load

We leveraged three different but complementary techniques for evaluating cognitive load: a working memory capacity test, self-reported effort, and a dual-task. Our investigations into working memory capacity were influenced by prior work, primarily focusing on text-only formats, and showed a positive correlation between working memory capacity and reasoning performance [45, 76]. In our work, we found that the effect of working memory capacity held, with high working memory individuals generally outperforming their low working memory counterparts. This effect was especially salient in the text-only condition. These findings help contextualize our findings on Bayesian task accuracy, which suggest that visualization and multimedia formats may be superior to text-only.

We also saw that participants with low working memory capacity performed better when using visualization alone than text alone. In line with Castro et al.'s work [17], this difference in performance in the low working memory group is indirect evidence that text-only elicited more cognitive load than visualization-only. Expanding this argument, we can deduce that the combination of text and visualization did not elicit more cognitive load than visualization-only. Given the prior findings that removing numbers from the text in the combined presentation positively affects reasoning performance[49], we expected to find evidence that combining text and visualization increases cognitive load, but our data does not support this notion. These findings have practical implications for visualization recommendation and accessibility. Visualizations can benefit populations with lower cognitive abilities and be beneficial in situations of high cognitive burden.

Notably, in Experiment 1, participants with low working memory capacity reported experiencing significantly lower frustration and temporal demand when using the combination format compared to text alone. This finding further supports the use of the multimedia format. However, it is noteworthy that Experiment 1 also showed no significant difference in the accuracy rates between the combination format and text for individuals with low working memory, highlighting the deficiency of analyzing accuracy alone. In general, our findings somewhat support the notion of a multimedia effect. In particular, there may be some benefit to having both text and visualization available to facilitate reasoning, especially for people with low working memory capacity. One potential explanation might be that visualization allows the viewer to offload items from memory, but the text is familiar and easy to process. This hypothesis corroborates the results of prior work that captured eye-gaze data as people solved Bayesian tasks [54]. Their results suggest that visualization makes it easy to identify relevant information, but the text may be easier to process compared to the visual format [54]. Another plausible explanation for our results is that participants with low working memory might prefer the flexibility of the combined format, which enables them to choose the format that best aligns with their mental model or preference. Further investigation is needed to better understand this phenomenon.

## 6.2 On The Failure of the Dual-Task Paradigm

The dual-task paradigm did not reveal differences in cognitive load across formats, even when accounting for individual differences in working memory capacity. Specifically, asking participants to hold a dot pattern in memory did not influence their reasoning performance. We hypothesized in Experiment 1 that this effect could be due to individual variability in the between-subject design. However, the within-subject Experiment 2 revealed similar findings, which contradicts H3 and prior work [45] possibly due to differences in experimental design. For example, Lesage et al.[45] performed a laboratory experiment with 179 first-year psychology students who participated in the previous study for course credits. Our study used a more diverse crowdsourced study population. Another possible explanation is that the secondary task was too easy, or our study participants may have written down the pattern instead of holding it in memory. Alternatively, the observed disparity may also be due to differences in the demographic makeup of our study populations.

Several researchers have developed guidelines for choosing an adequate secondary task, which includes considerations for task difficulty and similarity [58, 66]. However, it can be challenging to strike the perfect balance between the primary and secondary task as the latter has to be hard enough to increase cognitive load but not to the point of cognitive overload. Although the exact reason for the failed replication is unknown, we encourage researchers to consider modifying the dual-task methodology when conducting crowdsourced evaluations. One alternative study design could be to calibrate the secondary task's difficulty based on participants' abilities. For example, Castro et al. [18] used calibration in a study investigating the impact of divided attention on driving. Their participants performed a pre-test to identify the level of difficulty that elicited a 75% accuracy on the secondary task. For our choice of secondary task, one option would be to calibrate the size of the dot pattern to memorize based on participants' performance. Alternatively, Borgo et al.'s study on the impact of visual embellishment on engagement and working memory used a word selection secondary task [9], where users identified fruits among a crawling list of words. Researchers could calibrate the secondary task by tailoring the crawl speed of the words to each participant. Further, Borgo et al.'s [9] dual-task setup would be less susceptible to violations of the study tasks since it does not involve a recall task.

# 7 CONCLUSION

Our work expands the understanding of the relationship between working memory capacity and Bayesian reasoning by examining and comparing three presentation formats. By examining more granular accuracy measures, we showed that visualization-only and combination formats lead to less error in Bayesian reasoning than text-only formats. Moreover, we showed that working memory capacity mediates Bayesian reasoning accuracy, particularly in the text format. Finally, we showed that users with low working memory capacity are more accurate when using visualization alone compared to text alone. We discuss how these findings can impact visualization design guidelines, especially for low working memory capacity users. To this end, we argue for more diversified evaluation metrics and encourage the visualization community to leverage and apply existing research in cognitive science and related fields to better understand how people perceive and reason with visualizations.

## ACKNOWLEDGMENTS

The authors wish to thank Lace Padilla for her insights into the use of cognitive methods and Shayan Monadjemi for his valuable feedback on the manuscript. This material is based upon work supported by the National Science Foundation under grant number 2142977.

## REFERENCES

[1] Erik W Anderson, Kristin C Potter, Laura E Matzen, Jason F Shepherd, Gilbert A Preston, and Cláudio T Silva. 2011. A user study of visualization effectiveness using EEG and cognitive load. In *Computer graphics forum*, Vol. 30. Wiley Online Library, 791–800.

[2] Amy L Atkinson, Alan D Baddeley, and Richard J Allen. 2018. Remember some or remember all? Ageing and strategy effects in visual working memory. *Quarterly Journal of Experimental Psychology* 71, 7 (2018), 1561–1573.

[3] Laurel C Austin. 2019. Physician and nonphysician estimates of positive predictive value in diagnostic v. mass screening mammography: an examination of bayesian reasoning. *Medical Decision Making* 39, 2 (2019), 108–118.

[4] Melanie Bancilhon and Alvitta Ottley. 2020. Did You Get The Gist Of It? Understanding How Visualization Impacts Decision-Making. *arXiv preprint arXiv:2010.04096* (2020).

[5] Aron K Barbey and Steven A Sloman. 2007. Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences* 30, 3 (2007), 241–254.

[6] Jacques Bertin. 1983. *Semiology of graphics; diagrams networks maps.* Technical Report.

[7] Charles E Bethell-Fox and Roger N Shepard. 1988. Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance* 14, 1 (1988), 12.

[8] 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.

[9] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W Grant, Irene Reppa, Luciano Floridi, and Min Chen. 2012. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2759–2768.

[10] Gary L Brase. 2008. Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review* 15, 2 (2008), 284–289.

[11] Gary L Brase. 2009. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23, 3 (2009), 369–381.

[12] Roland Brunken, Jan L Plass, and Detlev Leutner. 2003. Direct measurement of cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 53–61.

[13] Tad T Brunyé, Holly A Taylor, and David N Rapp. 2008. Repetition and dual coding in procedural multimedia presentations. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 22, 7 (2008), 877–895.

[14] Tad T Brunyé, Holly A Taylor, David N Rapp, and Alexander B Spiro. 2006. Learning procedures: The role of working memory in multimedia learning experiences. *Applied Cognitive Psychology* 20, 7 (2006), 917–940.

[15] Alan D Castel, Jay Pratt, and Fergus IM Craik. 2003. The role of spatial working memory in inhibition of return: Evidence from divided attention tasks. *Perception & Psychophysics* 65, 6 (2003), 970–981.

[16] Spencer Castro. 2017. How handheld mobile device size and hand location may affect divided attention. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 1370–1374.

[17] Spencer C Castro, Helia Hosseinpour, P Samuel Quinan, and Lace Padilla. 2021. Examining Effort in 1D Uncertainty Communication Using Individual Differences in Working Memory and NASA-TLX. (2021).

[18] Spencer C Castro, David L Strayer, Dora Matzke, and Andrew Heathcote. 2019. Cognitive workload measurement and modeling under divided attention. *Journal of experimental psychology: human perception and performance* 45, 6 (2019), 826.

[19] WD Chiles. 1982. Workload, task, and situational factors as modifiers of complex human performance. *Human performance and productivity: Stress and performance effectiveness* (1982), 11–56.

[20] W Cole and J Davidson. 1989. Graphic representation can lead to fast and accurate bayesian reasoning. In *Proceedings. Symposium on Computer Applications in Medical Care*. 227–231.

[21] William G Cole. 1989. Understanding Bayesian reasoning via graphical displays. *ACM SIGCHI Bulletin* 20, SI (1989), 381–386.

[22] Andrew RA Conway, Michael J Kane, Michael F Bunting, D Zach Hambrick, Oliver Wilhelm, and Randall W Engle. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review* 12, 5 (2005), 769–786.

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

[24] Nelson Cowan, J Scott Saults, and Lara D Nugent. 1997. The role of absolute and relative amounts of time in forgetting within immediate memory: The case of tone-pitch comparisons. *Psychonomic Bulletin & Review* 4, 3 (1997), 393–397.

[25] Dick De Waard and KA Brookhuis. 1996. The measurement of drivers' mental workload. (1996).

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

[27] Ruth B Ekstrom and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests, 1976.* Educational testing service.

[28] Randall W Engle, Michael J Kane, Stephen W Tuholski, et al. 1999. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. *Models of working memory: Mechanisms of active maintenance and executive control* 4 (1999), 102–134.

[29] Jonathan R.G. Etnel, Jasmin M. de Groot, Moad El Jabri, Anouk Mesch, Nathalie A. Nobel, Ad J.J.C. Bogers, and Johanna J.M. Takkenberg. 2020. Do risk visualizations improve the understanding of numerical risks? A randomized, investigator-blinded general population survey. *International Journal of Medical Informatics* 135 (2020), 104005. https://doi.org/10.1016/j.ijmedinf.2019.104005

[30] 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.

[31] Gerd Gigerenzer. 1994. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability*. Wiley, 129–161.

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

[33] Amanda L Gilchrist, Nelson Cowan, and Moshe Naveh-Benjamin. 2008. Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller chunks in older adults. *Memory* 16, 7 (2008), 773–787.

[34] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[35] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[36] Weidong Huang, Peter Eades, and Seok-Hee Hong. 2009. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization* 8, 3 (2009), 139–152.

[37] Nabil Issa, Mary Schuller, Susan Santacaterina, Michael Shapiro, Edward Wang, Richard E Mayer, and Debra A DaRosa. 2011. Applying multimedia design principles enhances learning in medical education. *Medical education* 45, 8 (2011), 818–826.

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

[39] Vince Kellen, Susy Chan, and Xiaowen Fang. 2007. Facilitating conditional probability problems with visuals. In *International Conference on Human-Computer Interaction*. Springer, 63–71.

[40] David O Kennedy and Andrew B Scholey. 2000. Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology* 149, 1 (2000), 63–71.

[41] 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.

[42] Azam Khan, Simon Breslav, and Kasper Hornbæk. 2018. Interactive instruction in bayesian inference. *Human–Computer Interaction* 33, 3 (2018), 207–233.

[43] M Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction* 30, 5 (2014), 343–368.

[44] Robert Kraut, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen, and Mick Couper. 2004. Psychological research online: report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American psychologist* 59, 2 (2004), 105.

[45] Elise Lesage, Gorka Navarrete, and Wim De Neys. 2013. Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning* 19, 1 (2013), 27–53.

[46] Ottmar V Lipp and David L Neumann. 2004. Attentional blink reflex modulation in a continuous performance task is modality specific. *Psychophysiology* 41, 3 (2004), 417–425.

[47] Zhengliang Liu, R Jordan Crouser, and Alvitta Ottley. 2020. Survey on individual differences in visualization. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 693–712.

[48] Gerald Matthews, Lauren E Reinerman-Jones, Daniel J Barber, and Julian Abich IV. 2015. The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human factors* 57, 1 (2015), 125–143.

[49] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2536–2545.

[50] Tim P Moran. 2016. Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological bulletin* 142, 8 (2016), 831.

[51] Ab Mosca, Alvitta Ottley, and Remco Chang. 2021. Does interaction improve bayesian reasoning with visualization?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[52] Lambertus JM Mulder. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology* 34, 2-3 (1992), 205–236.

[53] Frederick L Oswald, Samuel T McAbee, Thomas S Redick, and David Z Hambrick. 2015. The development of a short domain-general measure of working memory capacity. *Behavior research methods* 47, 4 (2015), 1343–1355.

[54] Alvitta Ottley, Aleksandra Kaszowska, R. Jordan Crouser, and Evan M. Peck. 2019. The Curious Case of Combining Text and Visualization. In *EuroVis 2019 - Short Papers*, Jimmy Johansson, Filip Sadlo, and G. Elisabeta Marai (Eds.). The Eurographics Association. https://doi.org/10.2312/evs.20191181

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

[56] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, and R. Taylor, H. A.... & Chang. 2015. Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 529–538.

[57] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.

[58] Lace MK Padilla, Spencer C Castro, P Samuel Quinan, Ian T Ruginski, and Sarah H Creem-Regehr. 2019. Toward objective evaluation of working memory in visualizations: A case study using pupillometry and a dual-task paradigm. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 332–342.

[59] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. 2018. Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications* 3, 1 (2018), 1–25. https://doi.org/10.1186/s41235-018-0120-9

[60] Harold Pashler. 2016. *Attention*. Psychology Press.

[61] Evan M M Peck, Beste F Yuksel, Alvitta Ottley, Robert JK Jacob, and Remco Chang. 2013. Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 473–482.

[62] Ulf-Dietrich Reips. 2000. The Web experiment method: Advantages, disadvantages, and solutions. (2000), 89–117.

[63] Valerie F Reyna. 2004. How people make decisions that involve risk: A dual-processes approach. *Current directions in psychological science* 13, 2 (2004), 60–66.

[64] Valerie F Reyna. 2008. A theory of medical decision making and health: fuzzy trace theory. *Medical decision making* 28, 6 (2008), 850–865.

[65] Maria Riveiro, Tove Helldin, Göran Falkman, and Mikael Lebram. 2014. Effects of visualizing uncertainty on decision-making in a target identification scenario. *Computers & graphics* 41 (2014), 84–98.

[66] MS Sanders. 1987. McCormick—Human Factors in Engineering and Design.

[67] Valerio Santangelo and Emiliano Macaluso. 2013. The contribution of working memory to divided attention. *Human Brain Mapping* 34, 1 (2013), 158–175.

[68] Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of experimental psychology: general* 130, 3 (2001), 380.

[69] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing uncertainty about the future. *science* 333, 6048 (2011), 1393–1400.

[70] Nava Tintarev and Judith Masthoff. 2016. Effects of individual differences in working memory on plan presentational choices. *Frontiers in psychology* 7 (2016), 1793.

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

[72] Amos Tversky and Daniel Kahneman. 1981. *Evidential impact of base rates*. Technical Report. Stanford Univ Ca Dept Of Psychology.

[73] Amos Tversky and Daniel Kahneman. 2013. Judgment under uncertainty: Heuristics and biases. In *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I*. World Scientific, 261–268.

[74] Nash Unsworth, Richard P Heitz, Josef C Schrock, and Randall W Engle. 2005. An automated version of the operation span task. *Behavior research methods* 37, 3 (2005), 498–505.

[75] Patrick Weber, Karin Binder, and Stefan Krauss. 2018. Why can only 24% solve Bayesian reasoning problems in natural frequencies: frequency phobia in spite of probability blindness. *Frontiers in psychology* 9 (2018), 1833.

[76] Lin Yin, Zifu Shi, Zixiang Liao, Ting Tang, Yuntian Xie, and Shun Peng. 2020. The Effects of Working Memory and Probability Format on Bayesian Reasoning. *Frontiers in Psychology* 11 (2020), 863. https://doi.org/10.3389/fpsyg.2020.00863

[77] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology* 111, 4 (2016), 493.

[78] Bin Zhu and Stephanie A Watts. 2010. Visualization of network concepts: The impact of working memory capacity differences. *Information Systems Research* 21, 2 (2010), 327–344.