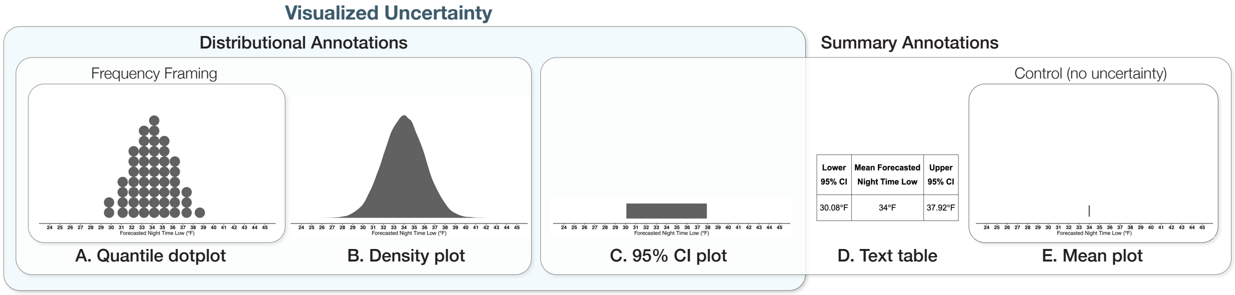
Examining Effort in 1D Uncertainty Communication Using Individual Differences in Working Memory and NASA-TLX

Spencer C. Castro, Helia Hosseinpour, P. Samuel Quinan, and Lace Padilla



Lower 95% CI	Mean Forecasted Night Time Low	Upper 95% CI
30.08°F	34°F	37.92°F

Fig. 1. Example stimuli and categorization used in the current study. Quantile dotplots (A) and density plots (B) visualize distributional information. Quantile dotplots also use frequency framing. Interval plots (C) showing 95% confidence intervals and text tables (D) showing 95% confidence intervals and a mean convey summary statistics. The mean plot (E) was used as a control, because it does not display uncertainty.

Abstract— As uncertainty visualizations for general audiences become increasingly common, designers must understand the full impact of uncertainty communication techniques on viewers' decision processes. Prior work demonstrates mixed performance outcomes with respect to how individuals make decisions using various visual and textual depictions of uncertainty. Part of the inconsistency across findings may be due to an over-reliance on task accuracy, which cannot, on its own, provide a comprehensive understanding of how uncertainty visualization techniques support reasoning processes. In this work, we advance the debate surrounding the efficacy of modern 1D uncertainty visualizations by conducting converging quantitative and qualitative analyses of both the effort and strategies used by individuals when provided with quantile dotplots, density plots, interval plots, mean plots, and textual descriptions of uncertainty. We utilize two approaches for examining effort across uncertainty communication techniques: a measure of individual differences in working-memory capacity known as an operation span (OSPAN) task and self-reports of perceived workload via the NASA-TLX. The results reveal that both visualization methods and working-memory capacity impact participants' decisions. Specifically, quantile dotplots and density plots (i.e., distributional annotations) result in more accurate judgments than interval plots, textual descriptions of uncertainty, and mean plots (i.e., summary annotations). Additionally, participants' open-ended responses suggest that individuals viewing distributional annotations are more likely to employ a strategy that explicitly incorporates uncertainty into their judgments than those viewing summary annotations. When comparing quantile dotplots to density plots, this work finds that both methods are equally effective for low-working-memory individuals. However, for individuals with high-working-memory capacity, quantile dotplots evoke more accurate responses with less perceived effort. Given these results, we advocate for the inclusion of converging behavioral and subjective workload metrics in addition to accuracy performance to further disambiguate meaningful differences among visualization techniques.

Index Terms—Uncertainty Visualization, Working Memory, Individual Differences, Online OSPAN, Effort, Workload, NASA-TLX

1 Introduction

As scientific communication of uncertainty information intended for a broad audience becomes increasingly common, researchers frequently utilize visualizations to communicate that uncertainty [92]. Although the term uncertainty has varying meanings across domains (for an expanded discussion, see [84]), in this paper, we focus on quantified uncertainty, such as in distributions. Unfortunately, reasoning with uncertainty presents difficulties for many people [52]. Correctly interpreting even common uncertainty visualizations, such as confidence intervals, is challenging for novices and trained experts alike [5].

With the rising interest in uncertainty communication, researchers have conducted a growing number of empirical studies to identify which uncertainty visualization techniques best support decisions with uncertain data (for reviews, see [40, 70]). This research provides evidence both that uncertainty visualizations support numerous types of judgments (e.g., [16, 17, 39, 50, 54, 83]) and that uncertainty visualizations can produce systematic biases (e.g., [5, 43, 72, 74]). Recently, however, prominent scholars in the field have questioned the utility of uncertainty visualizations compared to textual expressions of uncertainty [43]. As criticisms, they cite studies that find no difference between textual and visualized expressions of uncertainty (e.g., [42, 59, 62]), differences that are ameliorated with a longer time to complete the task [14], and biases that uncertainty visualizations produce (e.g., [5, 43, 62, 69, 72, 85]). Given that situations exist where the advantages of uncertainty visualizations become less clear and that some uncertainty visualizations produce biases, these researchers postulate that textual expressions of uncertainty may be preferable in certain cases [43].

The debate regarding the utility of uncertainty visualizations compared to text, however, remains unsettled. Some studies that have compared text to uncertainty visualizations predate the development of more efficacious distributional uncertainty visualization techniques,

- *S. Castro, is with the University of California Merced in Management of Complex Systems.*
- *P. S. Quinan is with the University of Utah School of Computing.*
- *H. Hosseinpour and L. Padilla are with the University of California Merced in Cognitive and Information Sciences.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

such as quantile dotplots [50], in which dots represent discrete probabilities in a distribution; hypothetical outcome plots (HOPs) [41], in which samples from the distribution are shown in an animation; value-suppressing uncertainty palettes [17], in which uncertainty is mapped onto a separate color dimension from value; and distributional ensemble plots [53], in which model runs are redrawn to more concisely represent the distribution. The few studies conducted after the development of these techniques that found no differences between visualizations and text (e.g., [59, 62]) also did not include these newer uncertainty visualization approaches in their evaluations.

Reliance solely on task accuracy, the most common evaluation metric for uncertainty visualizations (utilized in 36% of studies; [40]), may contribute to the inconsistent findings. Accuracy alone may not detect all meaningful differences among uncertainty communication techniques in simplified versions of real-world tasks. Considering additional metrics along with accuracy, such as the time to complete a task, can improve the precision of evaluations. Speed and accuracy, however, often exhibit a trade-off where an individual's performance improves when taking longer to complete a task, producing complex covariance (for a review, see [35]). Within visualization research, some have advocated for a *converging methods* approach (i.e., using multiple observable phenomena beyond speed and accuracy) to provide evidence for a visualization's utility [68, 71]. Scholars have also recommended incorporating individual differences (e.g., [31, 32, 71]), which quantify how different people vary in given abilities, such as graph literacy [66].

Two evaluation metrics less commonly found in uncertainty visualization research are the NASA Task Load Index (NASA-TLX), a standardized and psychometrically validated measure of subcomponents of *workload* (e.g., mental effort, physical effort, time pressure, performance, difficulty, and frustration [33, 34]), and *working memory*. Working memory consists of multiple subcomponents as well, each of which stores a finite amount of information in the mind for a short period of time. It is a limiting factor in the amount of mental effort that can be allocated to a given task and has been studied by researchers in psychology and cognitive science for decades (e.g., [2, 4, 18, 60, 71]). Theoretical work has recently suggested that working memory is likely the cognitive mechanism that produces reasoning errors with uncertainty visualizations (for a review, see [68]). As with other abilities (e.g., maximum running pace), individuals differ in their average capacity to utilize working memory (i.e., *working-memory capacity*) [20, 44]. Individuals with lower working-memory capacity have fewer cognitive resources available to complete demanding tasks. Thus, if participants with low-working-memory capacity show worse performance with some uncertainty communication techniques compared to others, we would have indirect evidence that the poor-performing uncertainty communication techniques require more working memory [68].

1.1 Overview

In this publication, we present an online evaluation of one-dimensional (1D) uncertainty communication techniques using the NASA-TLX (i.e., a measure of perceived workload [34, 91]) and a measure of individual differences in working-memory capacity (described in, Section 2.4.2). We employ both measures to compare the effort required by quantile dotplots [50], density plots, interval plots, visualizations with no uncertainty (showing only point estimates of means, as a control), and textual expressions of uncertainty shown in tables (see Figure 1). The stimulus types were selected to represent specific categories of uncertainty communication, which will be discussed in Section 2.

As a preview, our results demonstrate that for individuals with high-working-memory capacity, quantile dotplots evoke reliably better performance than densities, intervals, means, and text. Also, high-working-memory capacity individuals report less effort when using quantile dotplots than densities, intervals, and text. For individuals with low-working-memory capacity, quantile dotplots produce better performance than means and text, while densities perform better than intervals, means, and text. These results suggest that high- and low-working-memory individuals perform equally well with densities, but that those with high-working-memory capacity gain further advantages with quantile dotplots, performing certain tasks more accurately and more easily.

As its primary contribution, this work provides additional evidence for claims that modern distributional visualizations outperform textual descriptions of uncertainty, while also reproducing and recontextualizing recent work that challenges similar claims [43]. This work also begins to characterize differences among uncertainty communication techniques related to effort, which highlights differences in generalizability. Additionally, our approach is one of the first to consider individual differences combined with effort in the context of uncertainty communication. As a secondary contribution, we also provide a previously psychometrically tested, easy-to-use online individual-differences measure of working memory and a language-independent online version of the NASA-TLX that future visualization researchers can use to evaluate impacts on working memory and perceived effort.

2 Related Work: Uncertainty Visualization

When communicating uncertainty in 1D data, visualization practitioners commonly use graphic annotations to communicate distributional properties, such as confidence or credible intervals and distributional moments [70]. As illustrated in Figure 1, graphic annotations of uncertainty can either summarize the data via distributional moments, such as means and confidence intervals, or depict the full distribution in a more expressive way, such as with quantile dotplots or density plots.

2.1 Summary annotations

A substantial body of research shows that summary annotations produce numerous errors [5,42,43,72,74] that afflict both experts and novices [5] and can be invariant to training [7] (for a review, see [68, 70]). One theory proposes that summary annotations that use boundaries, such as error bars, create artificial conceptual categories that produce consistent errors [70]. For example, the Cone of Uncertainty produced by the National Hurricane Center depicts a 66% confidence interval around a mean forecasted hurricane path. The interval leads viewers to believe that areas inside the cone are in a perceived "danger zone" because the boundaries of cones create conceptual delineations of safe and dangerous areas [74, 83] (see also, [58]). However, the 66% confidence interval is not an inherently meaningful indication of danger. Viewers' beliefs about which areas are in the danger zone would change if the designers decided to instead plot a 65% or 95% confidence interval around the mean predicted hurricane path. Similar findings where *visual boundaries equal conceptual categories* occur with error bars [5, 37, 63] and visualizations showing the mean of a probability distribution [63].

2.2 Distributional annotations

Mounting evidence suggests that distributional uncertainty visualizations evoke improved performance compared to summary visualizations [16, 24, 39, 45, 46, 50, 74, 83], text communicating distributional information, and visualizations with no uncertainty [24]. Using distributional visualizations to convey distributional data has important benefits. Distributional visualizations are more expressive and provide a more thorough representation of the data. In contrast, summary visualizations such as confidence intervals can mask important data features, such as skewness or kurtosis. Providing viewers with a more accurate representation of the data with distributional visualizations may also require viewers to consider the uncertainty in the data rather than ignore it [41, 45]. Distributional visualizations may also help viewers avoid the categorical thinking often observed with summary statistics [7].

2.3 Frequency Framing

Beyond the usability advantages of more expressive distributional uncertainty visualizations, some also employ *frequency framing* (e.g., 1 out of 10) rather than probability framing (e.g., 10%), which has additional advantages. In empirical studies, frequency framing visualizations consistently perform as well as or better than all other tested techniques (e.g., [22–24, 26, 29, 39, 50, 88]). The three most common frequency framing visualizations are quantile dotplots [50], HOPs [41], and icon arrays that use icons to convey ratios (e.g., 1 of 10 icons) [26]. Frequency framing theory proposes that we experience probabilities as frequency in our daily lives [28]. For example, when driving to work, one might hit traffic on a particular route four out of five times and decide to take another route. Not everyone would conclude that they will experience traffic on that route 80% of the time. Scholars propose that frequency-framing visualizations such as quantile dotplots show consistently superior performance because they express uncertainty in a way that matches how most people experience probability [28]. Due to this intuitiveness, frequency-framed visualizations may require less effort when used in a visualization task [68].

When comparing modern distributional uncertainty visualizations, a winner is not always clear because some techniques show superiority only for specific experimental conditions. Whereas studies find that quantile dotplots outperform density plots most of the time [24,45,50], in some conditions quantile dotplots perform only as well as various other distributional plots, including density plots [24,45]. We argue that part of this inconsistency is due to standard metrics not being sufficiently sensitive to detect meaningful differences in effort. Part of the inconsistency could also relate to differences in population targets, which recording individual differences would help to address.

2.4 Evaluations of effort in uncertainty visualization

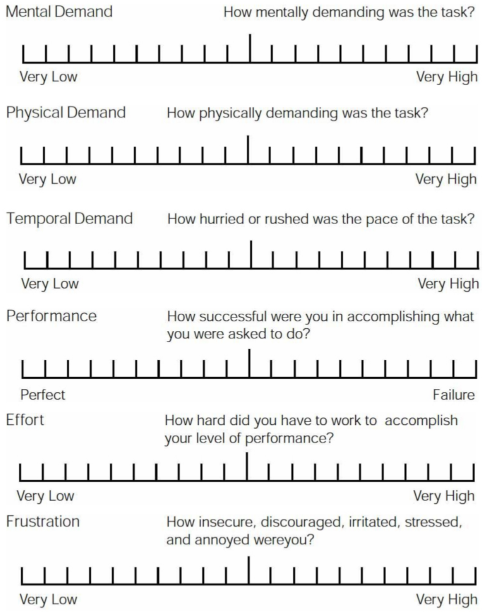
Visualization researchers, particularly those in cartography and geographic information systems mapping, have a history of interest in the effort required by uncertainty visualizations. In a review of empirical uncertainty visualization evaluations, 7.8% of studies evaluated proxies of effort (e.g., intuitiveness, effectiveness, and helpfulness) [40]. For example, seminal work examined the intuitiveness of visual encodings of uncertainty and found that participants rated fuzziness, location, value, arrangement, size, and transparency to be more intuitive for expressing uncertainty than a large set of other uncertainty encodings [55]. Intuitiveness may be inversely correlated with effort. From a Human-Computer Interaction (HCI) perspective, systems that match the real world can build an experience that feels more intuitive [65] and are therefore less effortful to use [6]. This approach seeks to minimize the use of voluntary effort and leverage learned, automated, natural responses, which do not increase subjective effort [47].

Other uncertainty research has explored potential proxies for effort. One study found that both novices and experts reported that urban growth uncertainty-geospatial visualizations were not too difficult or too complex to use [1]. Other studies asked about the effectiveness [19] and helpfulness [21] of maps with uncertainty. Although these works did not explicitly ask about effort, feedback about the difficulty and complexity of a task or the effectiveness and helpfulness of an uncertainty visualization may capture components of effort.

Uncertainty visualization researchers have also sought to measure aspects of effort more directly. In one study, researchers asked participants to report on multiple aspects of effort with questions about the ease of data lookup, ease of identifying the uncertainty, and degree of visual overload in uncertainty visualizations of volumetric data [64]. When comparing new glyph techniques to previously established methods, the researchers found some advantages for each technique but no definite superior approach [64]. One of the limitations of this study and other work that considers indirect measures of effort (e.g., [1,19,21,55,64]) is that how effectively the questions measure effort is unclear, which could be one source of the variability across findings.

2.4.1 NASA-TLX

Standardized measures of effort exist [34,71], some of which are commonly used in HCI and Human Factors (e.g., [12,71,75]). For example, the NASA-TLX has been utilized far beyond its original domain of aviation research for over 30 years (e.g., [25,30,33]). Because many different factors may contribute to workload , evaluating several of them individually is more precise than a single global evaluation of workload. The NASA-TLX has a standard set of six rating scales for workload [51] (see Figure 2). The factors that influence an experience of workload may come from the task itself, the participants' feelings about their performance, how much effort they put in, or the stress and frustration they feel. The workload evoked by different task elements may change as participants get more familiar with a task, perform easier or harder versions of the task, or move from one task to another.



Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Fig. 2. Online NASA-TLX questionnaire in which participants clicked on the mark that corresponded with their perception of each factor.

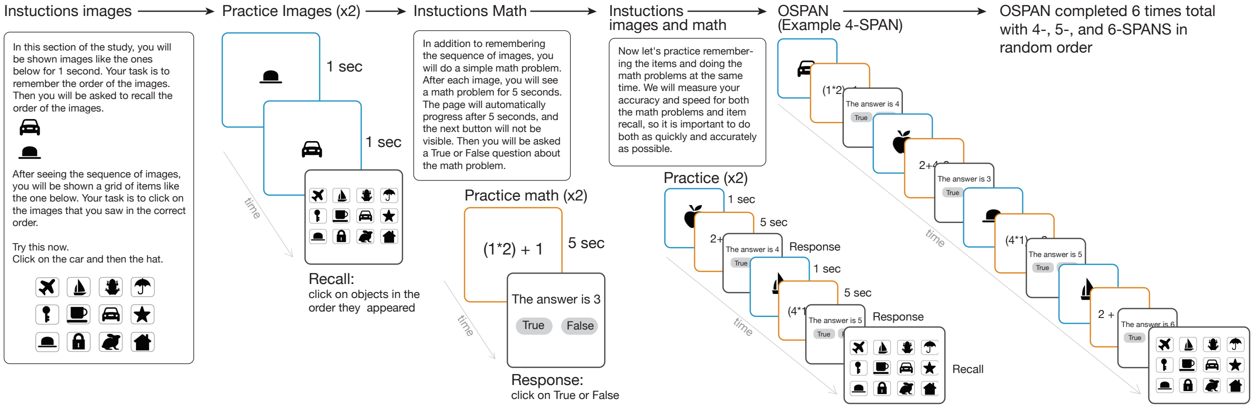
Physical components of workload are relatively easy to conceptualize and evaluate. The mental components of workload, however, can be more difficult to measure [13].

The NASA-TLX is suitable for visualization research because it can be conducted efficiently either online or in-person after a visualization task. Therefore, various HCI studies have used the NASA-TLX to examine applications that include uncertainty visualizations [9,80–82]. Researchers found that street map applications that incorporate uncertainty visualizations require lower mental demand and effort than those without uncertainty [9], but did not find differences in task accuracy. Another study found no difference in performance or workload for experts using an interactive map showing sensor data with and without uncertainty [82]. However, follow-up work found that experts experience less mental demand, physical demand, temporal pressure, required effort, and frustration than the general public when using interactive sensor maps with uncertainty [81]. This finding illustrates how different groups of people may experience different workload levels.

2.4.2 Individual differences in working memory

Working memory is a multicomponent system [18] that can influence most visualization decision-making processes [69]. Prior work has determined that working-memory capacity impacts accuracy when using 1D uncertainty communication, although the study did not compare the working memory demands of different uncertainty communication techniques [31]. Other work has compared the working memory demands of multiple visualizations that did not include uncertainty. For example, Zhu and Watts [98] demonstrated that using certain types of network diagrams increased difficulty specifically for individuals with low-working-memory capacity. This approach is noteworthy because the findings reflect real differences in users' abilities [27] that should be considered in visualization design. Finding that a visualization is easier for people with low-working-memory capacity suggests that it requires less working memory.

One of the common methods used to measure individual differences in working memory is an operation span *(OSPAN)* task [20,67] (for a pictorial OSPAN task, see Figure 3). In the original OSPAN task, participants must simultaneously try to remember sequentially presented words in their correct order while solving simple math equations [15,67,86]. OSPAN tasks require participants to hold information in their mind for a short period and measure how many items they can maintain while simultaneously retrieving and manipulating other information [2,3]. As illustrated in Figure 3, the OSPAN we used shows participants a sequence of images that they must remember and



In this section of the study, you will be shown images like the ones below for 1 second. Your task is to remember the order of the images. Then you will be asked to recall the order of the images.

After seeing the sequence of images, you will be shown a grid of items like the one below. Your task is to click on the images that you saw in the correct order.

Try this now.
Click on the car and then the hat.

In addition to remembering the sequence of images, you will do a simple math problem. After each image, you will see a math problem for 5 seconds. The page will automatically progress after 5 seconds, and the next button will not be visible. Then you will be asked a True or False question about the math problem.

Now let's practice remembering the items and doing the math problems at the same time. We will measure your accuracy and speed for both the math problems and item recall, so it is important to do both as quickly and accurately as possible.

Fig. 3. Procedure for the online OSPAN task adapted from [67] and made available in the supplemental materials[1]. In sum, participants complete 30 math problems while remembering a total of 30 images in six sequences. Participants receive an OSPAN score that combines their accuracy in the math problems with the number of correctly remembered items.

recall. At the same time, participants must also complete simple math problems to make sure that they are not rehearsing the sequence of images in their mind, which involves a different cognitive mechanism.

2.4.3 Converging measures

Converging measures matter because assumptions are associated with each outcome metric that could be compared to evaluate visualizations. For example, Hart and Staveland, in their NASA-TLX conceptual framework [34], describe three categories that relate to overall perceived workload: task imposed workload, operator behavior, and performance. Metrics of performance, such as speed, accuracy, and reliability, are observable by researchers and users. With direct feedback, visualization users can modulate their amount of effort to achieve desired goals, which is a change in user behavior. We have previously discussed how effort and accuracy are not perfectly correlated, which previous literature demonstrates [82]. Therefore, a measure of performance and a measure of effort (e.g., workload) would provide a clearer explanation of a visualization's efficacy. Finally, users' behavior could be better understood by having a measure of users' capabilities (e.g., working-memory capacity), which would improve predictions about the efficacy of visualizations for specific target audiences. With converging measures, we can collect metrics for the components that can influence the others within this conceptual framework.

3 Methods

The goal of the current work was to learn about the effort required by four uncertainty communication techniques (quantile dotplots, density plots, 95% confidence interval plots, and textual expressions of uncertainty in a table) compared to a control condition where no uncertainty information was present (mean plot). We selected quantile dotplots to represent distributional uncertainty visualizations that use frequency framing because studies consistently find they evoke better performance than other tested techniques [24, 45, 50]. We selected density plots to represent the category of distributional visualizations that do not use frequency framing since researchers have also previously tested this technique [24, 50]. Other effective distributional visualizations include violin plots [16], but we opted for densities since they are visually more similar to quantile dotplots and thus make for a more direct comparison. We selected 95% confidence intervals (95% CIs) to represent summary visualizations of uncertainty, as opposed to boxplots or other types of intervals, because numerous studies have evaluated their efficacy [5, 16, 38, 74, 83], and they are common in scientific communication. We did not include the mean in the 95% confidence interval plot because recent work has found that it can bias viewers' judgments [45]. We also used a text table to display 95% confidence intervals with the mean, representing textual expressions of uncertainty as a category. Here, we included the mean since no prior work suggests that textual means bias readers' judgments. Finally, as a control, we included a visualization of the mean, to determine whether the results were driven by visualization generally rather than uncertainty visualization.

The experiment consisted of five online conditions, each utilizing one of the five afformentioned communication techniques and taking approximately 20 minutes to complete. The online survey software Qualtrics [78] randomly assigned each participant to complete one of the five conditions. Our goal was to collect 150 participants per condition from the Prolific [77] online population. A few more participants were collected per condition to account for those who had been excluded due to failing one or more of four criteria: duplicate IP addresses (n = 4, one for density and text, one for density and dots), failing the Qualtrics fraud detection criteria [78] (n = 11), accuracy below chance (i.e., 50% or 15 of 30) on simple math True/False questions (n = 2), or remembering less than 10% (i.e., 3 of 30) of the objects in the working memory task (n = 1) (described in detail in Section 3.1.1). We interpreted below chance performance or identifying 3 of 30 objects correctly as participants not paying attention to the task. The full sample included: quantile dotplots n = 151, density plots n = 154, interval plots n = 152, mean plots n = 151, text tables n = 153; total n = 761. The analysis excluded four participants from the quantile dotplot group (n = 147 remaining), seven from the density plot group (n = 147 remaining), three from the interval group (n = 149 remaining), one from the mean group (n = 150 remaining), and three from the text group (n = 150 remaining). The analysis described in Section 3.3 includes 743 participants in total (Female = 354, Male = 362, nonbinary/third gender = 22, Prefer not to say = 1, Prefer to self-describe = 4; Age M = 31.55, SD = 11.66). All participants were paid in accordance with minimum wage laws for agreeing to participate, were at least 18 years old, and were not allowed to participate in more than one condition. Before beginning the study, participants read and agreed to an IRB-approved consent form.

3.1 Procedure and Tasks

The study consisted of three main parts:

1. After providing consent, participants completed a pictorial OSPAN adapted for online use [67] (see, Figure 3).
2. Participants completed an established uncertainty-visualization resource-allocation task where they assumed a risk manager's role [73]. Provided with information via one of the five communication techniques (see Section 3.2), participants determined how to allocate resources for preparatory risk reduction measures.
3. Participants completed the NASA-TLX [57, 91]. Subsequently, Qualtrics prompted participants to describe the strategy they used in as much detail as possible in a text box. Next, they answered a series of demographic questions and questions to determine graph literacy [66]. Finally, participants were told how they performed on the resource allocation task.

[1] Supplemental materials for this work, including the Qualtrics versions of the online OPSAN and NASA-TLX tasks, scoring scripts, links to examples of each task, each stimuli, and our analyses, can be found at: https://osf.io/6xzna/.

3.1.1 Individual-differences working-memory measure

We created an online version of a shortened OSPAN (i.e., the individual-difference measure of working memory), based on [67]. As illustrated in Figure 3, participants were shown a sequence of images to remember and recall, interspersed with simple math problems they were required to solve. After reading the instructions, participants completed two practice trials of remembering and recalling a sequence of two images, two practice math problems, and two practice OSPANs that required remembering images and solving math problems simultaneously. In the OSPANs, participants viewed a simple black-and-white silhouette of a common item for one second. During that time, the mouse disappeared, and participants could not progress to the next page. The page progressed automatically after one second. Then a math problem appeared on the screen for five seconds, after which the page, again, progressed automatically. Participants then judged whether a solution to a math problem was true or false and clicked next. The test repeated displaying an image, then a math problem, followed by a true or false question about the math problem a set number of times (twice for the practice trials and 4, 5, or 6 times for the working-memory measure). The term *n*-SPAN (e.g., 4-SPAN) refers to the sequence occurring *n* times. At the end of each OSPAN, participants recalled the order in which the images appeared by clicking on images in a 4 x 3 grid. For the individual-difference measure, participants completed 4-, 5-, and 6-SPANs two times each in random order.

Scoring of the OSPANs begins by adding together the number of times during the recall phase that participants correctly selected an item in the order it was shown. For example, in the 4-SPAN condition, participants were subsequently shown a house, an apple, an umbrella, and a frog. Participants who selected the house, then the apple, then the frog received one point for the house and one point for the apple for a total of two out of four possible points. A perfect score across all six OSPANs totaled 30 out of 30 items selected in the correct order; 96 participants achieved this score. These initial scores were then weighted by multiplying the participants' proportion of correct True-or-False math-problem responses. For example, if participants answered all of the math problems correctly, their total score would not change (i.e., it would be multiplied by one; 43 participants achieved perfect scores), but if participants answered only 50% of the math problems correctly (i.e., chance performance; zero participants received this score), their score would be cut in half. This scoring method ensures that participants who performed poorly or ignored the math problems in favor of maintaining the SPAN items in memory did not receive a strategic advantage in their working-memory score. Participants were divided into high- and low-working memory groups via a median split in line with previous working-memory research [27, 93]. The online version of this short OSPAN working-memory measure and the R script for scoring are available in the supplemental materials.

3.1.2 Uncertainty visualization task

This study employed a previously established task [73] in which participants assumed a risk manager's role and had to decide when to allocate funds from a virtual budget for risk mitigation actions in Peru. The instructions detailed that alpacas in Peru have died from cold temperatures in previous years because they cannot typically withstand temperatures below 32°F. The participants' task was to determine when to send cold weather aid to alpaca farms to reduce alpaca deaths. Researchers designed this task based on communications with colleagues at the Red Cross, who detailed their decision-making process when in 2016, Peru issued a state of emergency because tens of thousands of alpacas were dying in a cold snap [36].

In the task's hypothetical scenario, the Red Cross has a limited budget ($18,000) for 18 days. Purchasing and delivering blankets to farmers costs $1,000 per night. If participants failed to issue blankets to the farmers and the temperature dropped below 32°F, it cost $6,000 from their budget for postdisaster relief. The incentive for not simply giving blankets each night was that, for every $1,000 left in their budget at the end of the study, participants would receive an extra 10¢ bonus.

Before beginning the task, participants completed an attention check where they answered a question about the information in the instructions. If they answered the question incorrectly, the Qualtrics online survey software disqualified them from completing the study. Participants then viewed 18 nighttime temperature forecasts using one of the five stimuli types (quantile dotplots, density plots, 95% CI plots, mean plots, and text tables) shown in Figure 1. The 18 forecasts consisted of six mean temperatures (31-36°F), each shown with three levels of variance (low, medium, and high variance; shown in Figure 4). The order of the trials was randomized. For each forecast, participants had to decide if the nighttime low would drop below freezing (32°F), in which case they should issue blankets to the alpaca farmers. Participants received feedback on their decisions only at the end of the experiment (i.e., after all 18 days/trials of part 2 and the post-experiment questionnaire of part 3), when they received their payment and any bonuses.

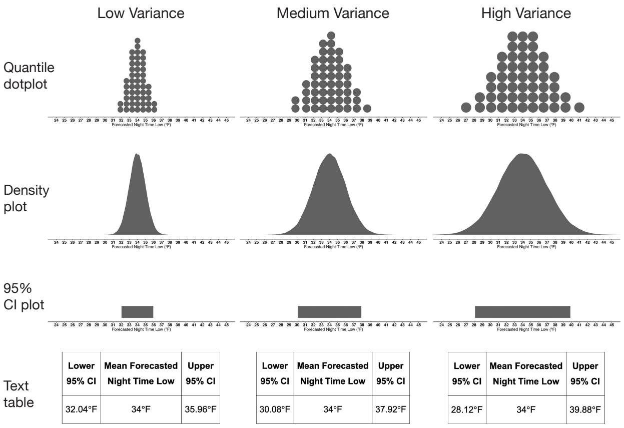
Based on the cost of issuing aid ($1,000) and the penalty for not issuing aid if the temperature drops below freezing ($6,000), the optimal strategy is to issue aid when the probability of freezing is greater than 16.6% (1,000/6,000 = 0.166). As detailed in prior work [73], for simulated forecast data with low variance (as defined by the data generation procedure described in Section 3.2), the optimal strategy is to give aid only at or below 32.97°F. For simulated forecast data with medium variance, it is optimal to only give aid at or below 33.94°F and for data with high variance at or below 34.90°F. These strategies resulted in a scoring guide where the correct answers were to issue aid for data with low variance from 31-32°F, with medium variance from 31-33°F, and with high variance from 31-34°F.

3.1.3 NASA-TLX

We used the NASA-TLX [34, 57, 91] (see Figure 2) to examine participants' workload experiences. Participants reported on the workload they believed the uncertainty visualization task evoked by clicking on each of the six scales at the point that matched their experience. Recently, researchers have advocated for independently assessing the NASA-TLX subscales rather than assessing overall workload [25]. In line with these recommendations, we report on each subscale in our results (see Section 3.3.2). Also, as with the individual-differences working-memory measure, an online version of the NASA-TLX is available in the supplemental materials.

3.2 Stimuli Generation

The stimuli in this experiment were generated within R [79] using the packages *ggplot2* v. 3.3.0 [89], *tidybayes* v. 2.0.3 [48], *ggdist* v. 2.4.0 [49], and *stats* v. 4.0.3 [79]. We simulated the forecast data using the `rnorm` function to generate 100,001 random deviates, with specified means (31, 32, 33, 34, 35, 36) and standard deviations (1, 2, 3). This simulation resulted in 18 data sets with normal distributions. As noted in Section 3.1.1, for each communication technique, 18 stimuli were generated using these 18 simulated data sets. We used functions within ggdist to visualize the probability density information for quantile dotplots, density plots, and 95% CI plots. For the quantile dotplots, each dot represents a one out of 50 chance that the nighttime low will



Lower 95% CI	Mean Forecasted Night Time Low	Upper 95% CI
32.04°F	34°F	35.96°F

Lower 95% CI	Mean Forecasted Night Time Low	Upper 95% CI
30.08°F	34°F	37.92°F

Lower 95% CI	Mean Forecasted Night Time Low	Upper 95% CI
28.12°F	34°F	39.88°F

Fig. 4. Example low-, medium-, and high-variance stimuli from the study.

be a particular temperature. Studies have determined that viewers can effectively interpret quantile dotplots with 50 dots [50]. For the mean plot, we used ggplot2 to display each mean temperature. We created the text tables in Adobe Illustrator, using the mean of each data set along with the upper and lower confidence intervals. The full stimulus set is available in the supplemental materials.

Along with each stimulus, the trials included a brief description of the corresponding communication technique. For the quantile dotplots, *"Each dot in the forecast represents a 1 out of 50 chance."* For the density plots, *"The tallest point on the curve represents the most likely nighttime low."* For the 95% CI plots, *"The forecasters are 95% confident that the nighttime low will be in the range shown by the grey bar."* For the mean plots, *"The vertical line shows the mean forecasted nighttime low temperature."* Finally, for text tables, *"Forecasters are 95% confident that the night time low will be between the upper and lower confidence intervals (95% CIs)."*

3.3 Analysis

In the following analyses, we examined the effects of uncertainty communication techniques on both accuracy and perceived effort in a resource allocation task for individuals with high- and low-working-memory capacities. We analyzed both accuracy (see Section 3.3.1) and perceived effort (see Section 3.3.2) in multiple Bayesian regression models using R and the *tidyverse* v. 1.2.1 [90], *brms* v. 2.13.0 [10], and *tidybayes* v. 2.0.3 [48] packages.

Bayesian modeling is appropriate for mixed designs and has the crucial advantage of yielding posterior distributions. Researchers can visualize posterior distributions using distributional visualization techniques instead of the error bars typical in frequentist statistics. To interpret Bayesian results, readers should consider both the credible intervals and posterior distributions. Any 95% credible intervals that do not include zero provide reliable evidence that a difference exists between the comparison groups. In this paper, we do not report "significant" effects common with frequentist statistics because scholars have called into question the validity of this characterization of findings [87]. Instead, we present degrees of evidence of reliable effects. Readers can visually see the effect's size by noting the distance of the confidence intervals from zero and the relative differences between posterior distributions. Posterior distributions show each condition's relative effects after controlling for all other items in the model. Brackets (e.g., []) in the analysis contain 95% Bayesian credible intervals, and all measures of central tendency (*M*) are medians of the posterior distribution.

Our accuracy model was a Bayesian multilevel logistic regression that included accuracy (coded as 0 = incorrect, 1 = correct) as the response variable. We included the fixed effects of communication technique (with quantile dotplots as the referent), working-memory group, the interaction between communication technique and working-memory group, variance in the forecast data (variance), forecast mean temperature, and graph literacy. We included random slopes for variance and temperature and a random intercept for each participant. We believed it was essential to control for the effects of variance, temperature, and graph literacy as covariates, but they were not central to our research question. We thus centered variance, temperature, and graph literacy to interpret the other effects in relationship to the average effects of these covariates. In the following sections, we report only on the effects of visualization communication technique and working-memory group, but note that these effects are above and beyond the effects of variance, temperature, and graph literacy (detailed in supplemental materials). The results produced by this model were in the form of log-odds that we converted to probabilities for easier interpretation.

For perceived effort, we conducted a sequence of Bayesian linear regression models with each NASA-TLX factor as the outcome measure. We included the communication techniques, working-memory group, the interaction between communication techniques and working-memory group, and centered graph literacy as predictors in each model. As in the prior model, graph literacy was considered a covariate.

We used qualitative analysis of self-reported strategies to identify ways participants' resource allocation judgments varied across the working-memory groups and communication techniques. Our motivation for collecting self-reported strategies was to understand the viewers' considerations when making their judgments and how the visualizations' designs impacted their interpretations of the data. One of the investigators read all 743 open-ended responses and coded each for four predefined strategies as detailed in Section 3.3.3.

3.3.1 Accuracy

The accuracy for each visualization communication technique and working-memory group is shown in Figure 5. The left column of Figure 5 shows 95% credible intervals for the comparisons between high- and low-working-memory groups viewing each communication technique. The right column depicts posterior distributions that show the accuracy for each group as the probability of a correct response. Figure 5 is also broken down by data variance — low variance (*M* = 86.26%, [82.14, 90.07]), medium variance (*M* = 82.69%, [78.15, 87.33]), and high variance (*M* = 78.49%, [72.97, 83.94]) — to show the relative increase in accuracy when less certainty is associated with the data. We found no reliable interaction between data variance and visualization technique (i.e., the effect of variance was the same for each communication technique), but we included variance in the figure for thoroughness.

Across the working-memory groups, participants viewing quantile dotplots (*M* = 86.16%, [84.48, 87.76]) reliably performed more accurately than participants viewing intervals (*M* = 82.63%, [80.83, 84.53]), means (*M* = 79.16%, [76.97, 81.09]), or text (*M* = 80.93%, [78.97, 82.81]). Participants viewing densities (*M* = 84.94%, [83.20, 86.65]) performed more accurately than participants viewing means or text. We did not find reliable evidence for a difference between quantile dotplots and density plots. This finding replicates prior work showing that distributional uncertainty visualizations outperform summary annotations of uncertainty, text of summary statistics, and visualizations with no uncertainty [24]. Intervals were also slightly more accurate than mean plots. Otherwise, we found no evidence for a difference in accuracy between interval plots, text tables, and mean plots, which also replicates prior work that reports no difference between these uncertainty communication techniques and text (for a review, see [43]). Together, these findings highlight the importance of testing a wide range of uncertainty visualizations, including modern distributional techniques, to avoid presenting uncertainty visualization findings in overly broad strokes.

There was also an effect of working-memory group for quantile dotplots and mean plots (seen in Figure 5 where the 95% credible intervals do not include 0, denoted with **). These effects provide evidence that individuals with higher working-memory capacity had better performance than those with lower working-memory capacity when viewing either quantile dotplots and mean plots.

The results also revealed a reliable interaction between uncertainty communication technique and working-memory group (log odds: *b* = 0.37, [.01, .74]). We computed the same model as previously described to breakdown the interaction but only with participants in the high- or low-working-memory groups. The credible intervals for the between-visualization comparisons are shown in Figure 6 (left for high-working memory and right for low-working memory). For participants with high-working-memory capacity, quantile dotplots (*M* = 88.84%, [86.10, 90.37]) produced reliably better performance than all the other techniques (see Figure 6a): densities (*M* = 84.84%, [82.24, 87.21]), intervals (*M* = 84.58%, [81.97, 87.06]), means (*M* = 81.36%, [78.62, 84.17]), and text (*M* = 81.71%, [78.95, 84.44]). This finding adds nuance to the efficacy of quantile dotplots by suggesting that the people who benefit the most from this technique have sufficiently high-working-memory capacity to capitalize on the additional frequency information.

For individuals with low-working-memory capacity, quantile dotplots (*M* = 84.63%, [82.19, 88.05]) and densities (*M* = 85.85%, [83.45, 88.04]) produced better performance than means (*M* = 78.07%, [75.03, 81.02]) and text (*M* = 80.85%, [78.00, 83.46]). Densities also produced better performance than intervals (*M* = 81.47%, [78.77, 84.08]). If a visualization designer is interested in using uncertainty visualizations useful for individuals with low-working memory, this finding suggests that either quantile dotplots or density plots would be equally appropriate. However, given that quantile dotplots also meaningfully improve

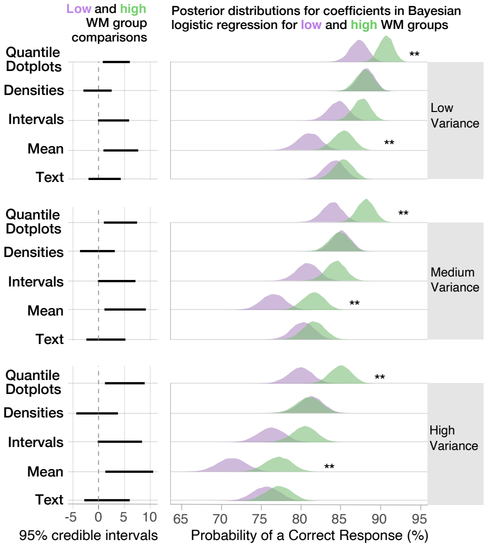
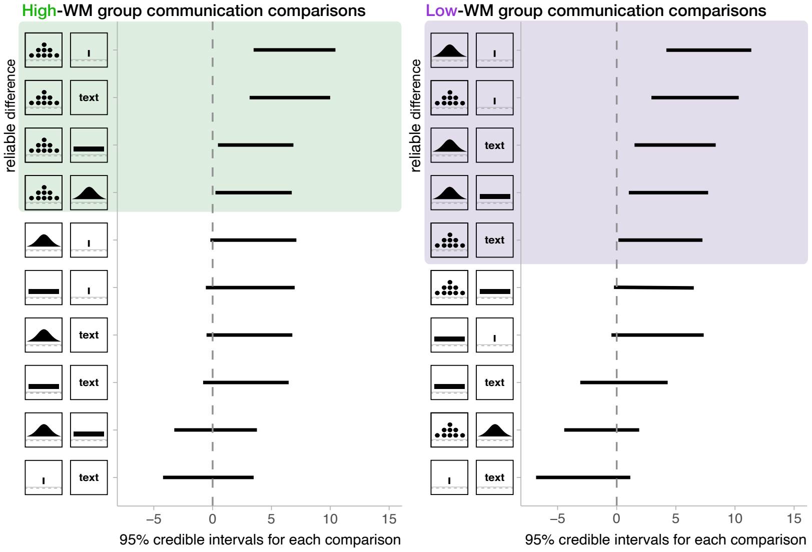


Fig. 5. 95% credible intervals and posterior distributions for multilevel Bayesian logistic regression modeling accuracy for each communication type and variance level (n.b., ** indicates a reliable difference between high- and low-working-memory groups).



(a) High-Working-Memory Group

(b) Low-Working-Memory Group

Fig. 6. 95% credible intervals for comparisons between each visualization type ordered from most to least reliable, for participants with (a) high-working memory and (b) low-working memory.

performance for individuals with high-working-memory capacity, our work indicates that quantile dotplots are the most efficacious uncertainty communication technique out of those we tested.

3.3.2 Perceived Workload

The main effects for each subscale of the NASA-TLX for each visualization communication technique and working-memory group are shown in Figure 7. As indicated by the 95% credible intervals that do not include zero (also denoted by **), we found a relatively consistent working memory effect for each visualization and each subscale of the NASA-TLX (c.f., temporal demand). This finding provides evidence that, on average, individuals with lower working-memory capacity perceived the resource allocation task as more effortful than those with high-working-memory capacity.

Specifically, the results revealed an effect of working memory for **perceived effort** (b = -2.44, [-3.74, -1.20]) and **mental demand** (b = -1.86, [-3.03,-0.64]). Participants with lower working-memory capacity reported more **effort** (M = 12.14, [10.76,13.70]) and **mental demand** (M = 10.59, [9.33, 12.62]) compared to those with high-working memory (**effort**: M = 10.58, [8.17, 13.02]; **mental demand**: M = 9.51, [7.83, 11.33]). Participants in the low-working-memory group (M = 8.42, [7.24, 10.19]) also reported worse **performance** (b = -1.59, [-2.74, -0.46]) than those in the high-working-memory group (M = 7.08, [5.71, 8.60]). Further, participants in the low-working-memory group (M = 7.80, [6.69,8.92]) reported higher **frustration** (b = -2.96, [-4.18, -1.72]) than those in the high-working-memory group (M = 5.48, [4.06, 6.92]).

When examining the average effects for both high- and low-working-memory groups, our results demonstrated that participants' **perceived effort** was reliably lower (b = 1.57, [0.20, 2.94]; b = 1.34, [0.20, 2.51]) for quantile dotplots (M = 10.39, [8.17, 12.30]) compared to intervals (M = 11.71, [10.11, 13.40]) or text (M = 12.75, [11.41, 14.02]). This finding is noteworthy in combination with the finding that participants had greater accuracy when using quantile dotplots compared to text for both working-memory groups. Participants also found viewing means (M = 10.59, [8.93, 12.15]) required less **effort** than text (b = 1.87, [0.72, 2.99]). This result is also notable because means performed quantitatively the worst. Given the lack of uncertainty information, it seems reasonable that the mean plots are both easy to interpret (i.e., do not require much effort) and do not provide the information necessary for participants to successfully perform the task.

Participants also reported that viewing quantile dotplots (M = 9.72, [8.40, 11.08]), density plots (M = 9.89, [8.51, 11.22]) and interval plots (M = 10.32, [8.98, 11.66]) was less **mentally demanding** than viewing textual expressions of uncertainty (M = 11.20, [9.91, 12.52]; respectively: b = -1.47, [-2.54, -0.40]; b = -1.48, [-2.83, -0.15]; b = -1.31, [-2.67, -0.01]). This finding provides additional evidence that certain visualizations meaningfully offload cognition onto the visual system, making judgments with visualizations less mentally demanding than those made with text.

In terms of **perceived performance** (see Figure 2, noting the reverse coding where low numbers indicate perfect performance and high numbers denote failure), participants viewing quantile dotplots (M = 7.17, [5.77, 8.52]) and means (M = 7.36, [5.97, 8.70]) reported feeling more successful in accomplishing their task (b = -1.65, [-2.88, -0.41]; b = -1.17, [-2.14, -0.21]) than participants viewing text (M = 8.53, [7.16, 9.86]). Although this finding makes sense for quantile dotplots, why participants viewing mean plots were so confident in their performance is unclear. Without uncertainty information, enough information may not be available to evaluate performance accurately. Conversely, for quantile dotplots, the frequency framing increases the viewer's confidence in the decision if they are able to more readily understand the additional information. At no point did any participants receive feedback on any task before completing the NASA-TLX.

Densities and intervals do not show consistent differences between working-memory groups in the NASA-TLX or in accuracy. Given that density plots had greater accuracy than intervals and working-memory capacity had no significant impact on their accuracy or perceived mental demand, density plots are an ideal visualization choice for viewers across working-memory capacities. However, visualization designers could further advantage high-working-memory-capacity individuals with superior performance and less effort by utilizing quantile dotplots and expect similar performance and effort to densities with low-working-memory viewers.

3.3.3 Self Reported Strategies

To examine participants' strategies, we first reviewed previously collected responses from an earlier pilot study run on Amazon's Mechanical Turk. Although the open-ended responses were sparse, we observed three unique strategies: 1) uncertainty-aware strategies which used words like variability, spread, and uncertainty; 2) strategies that used deterministic rules, such as the confidence intervals including 32°F; and 3) feeling deeply for the alpacas and making risk-averse budget allocation judgments to protect them (i.e., always issue aid).

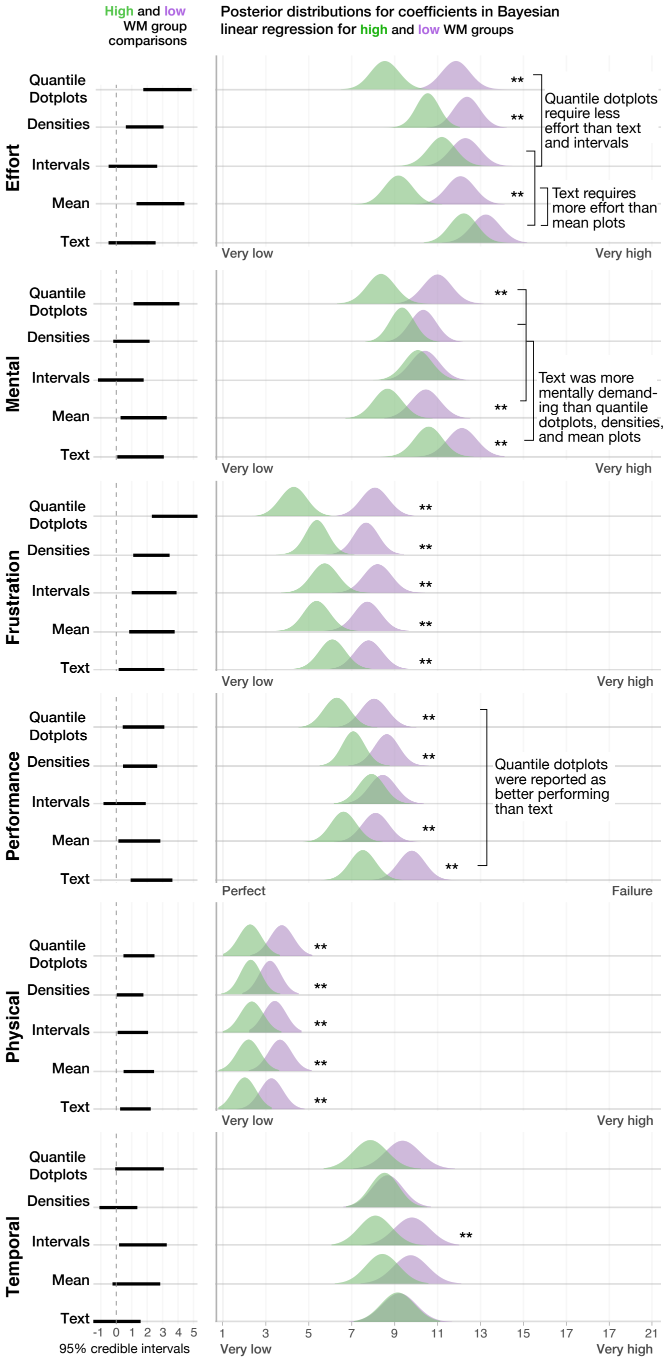


Fig. 7. 95% credible intervals and posterior distributions for multilevel Bayesian linear regression modeling NASA-TLX for each communication type and working-memory group (n.b., ** indicates a reliable difference between high- and low-working-memory groups). Annotations report reliable perceived effects for differences between visualization types (averaged across working-memory group).

We conducted the current full study using Prolific, and we noticed a drastic improvement in the quality of the open-ended responses (average character count = 168). Encouraged by the data's quality, we created a four-strategy coding scheme that used the three previously observed strategies and a code for an indeterminate strategy. The last was added because we wanted to know if some communication techniques evoked greater confusion, leading to no distinguishable strategy.

The **uncertainty strategy** code was given to participants who mentioned using some variation of the visualization's spread or standard deviation or wrote about probability, chance, or any other synonym for uncertainty. For example, a participant in the density plots group wrote, *"Since the graph looked like a bell curve, I eyeballed the standard deviation for most of it. If the temperature was around 2-3 SD, I would give the blankets. However, if it was further out, I wouldn't do so."*

We used the **deterministic strategy** code when participants reported a thought process that was less about the uncertainty in the forecast and more about particular data points, such as the mean or the freezing temperature (32°F). A deterministic strategy highlights when participants tend to ignore uncertainty in their interpretation. An example of the deterministic strategy from a participant in the density plots group was,*"If the temperature hit 32 or was below that I issued blankets, if it was over 32 or barely touched it, I didn't give them blankets."*

We selected the **alpaca-preservation strategy** because we observed that animal lovers might have a bias toward saving the alpacas. Such a strategy is not entirely surprising. We selected this task hoping to maximize participants' incentives to do their best by using a context that some people may care about and paying extra based on performance. An example of the alpaca-preservation strategy from a participant in the density plots group was, *"Why would I want to save money to line my own pockets? The alpacas take the higher priority."*

The **indeterminate** code was used when participants did not show signs of using any particular strategy. Some participant responses simply fail to include useful strategy information. An example of an indeterminate-strategy response from a participant in the density plots group was,*"I used the graphs."*

Notably, some responses used a mix of strategies. For example, the following participant in the density plots group used both the uncertainty and the alpaca-preservation strategies,*"As a Buddhist, I do not want animals to suffer, thus I mostly chose to give them blankets, if there was any chance that they might be harmed or freeze. It took a little reading and practice to get things familiar. I considered the overall percentage, and if 5 percent chance existed, they could be arm, or higher, they definitely got blankets."* In cases of mixed strategies, participants received codes for all applicable strategies.

Table 1 provides the summary of the codes for participants' self-reported strategies based on visualization type and working-memory group. Key findings from the table include:

- High-working-memory participants who viewed densities and quantile dotplots were more likely than all other groups to employ a strategy that incorporated uncertainty.
- Participants viewing mean plots and text tables were most likely to use a deterministic strategy.
- Whereas participants who viewed mean plots and text tables were the least likely to use a strategy that incorporated uncertainty, those with high-working-memory capacity were more likely than those with low-working-memory to incorporate uncertainty into their judgments.
- Those with low-working-memory who viewed mean plots were more likely than those with high working-memory to report no strategy.

In sum, distributional visualizations caused participants to report considering the uncertainty more than summary uncertainty communications or visualizations with no uncertainty. Further, textual expressions of uncertainty more often led to participants using strategies associated with ignoring the uncertainty.

4 General Discussion

This work presented an empirical evaluation of four 1D uncertainty communication techniques and a control condition with no uncertainty. Our goal was to use converging measures of individual differences in working memory, workload, and accuracy to provide a more nuanced picture of the efficacy of 1D uncertainty communication techniques. Considering these metrics together, this work found that both quantile dotplots and density plots showed superior performance to interval plots, textual descriptions of uncertainty, and visualizations with no uncertainty in a resource allocation task. Further, the qualitative analysis demonstrated that distributional visualizations tended to evoke strategies that included uncertainty more often, which may have contributed to these techniques' superior performance. When comparing quantile dotplots to density plots, the analysis revealed that both methods work

Stimuli	Working Memory Group	Uncertainty Strategy (% use)	Deterministic Strategy (% use)	Alpaca-preservation Strategy (% use)	Indeterminate Strategy (% use)
Quantile dotplot	High	61.6%	42.5%	5.5%	4.1%
	Low	44.6%	48.6%	5.4%	12.2%
	Average	53.1%	45.6%	5.4%	8.2%
Density plot	High	63%	35.6%	9.6%	2.7%
	Low	52.7%	37.8%	6.8%	9.5%
	Average	57.8%	36.7%	8.2%	6.1%
95% CI plot	High	45.9%	39.2%	14.9%	6.8%
	Low	52%	41.3%	10.7%	9.3%
	Average	49%	40.3%	12.8%	8.1%
Mean plot	High	29.3%	66.7%	8%	6.7%
	Low	17.3%	57.3%	9.3%	18.7%
	Average	23.3%	62%	8.7%	12.7%
Text table	High	46.7%	61.3%	8%	2.7%
	Low	37.3%	61.3%	4%	5.3%
	Average	42%	61.3%	6%	4%

Table 1. Percentages of uncertainty, deterministic, alpaca-preservation, and indeterminate strategy reported by users in high- and low-working-memory groups per visualization type.

equally well for low-working-memory individuals. However, for individuals with high-working-memory capacity, quantile dotplots evoked more accurate responses with less perceived effort.

Our work alludes to the possibility that the frequency framing of quantile dotplots is useful in a resource allocation task but requires a sizable amount of working memory, which does not fully align with assertions from prior work about frequency framing [28]. Our results also suggest that people with an abundance of working memory can allocate some of it to capitalize on the frequency information in quantile dotplots to make superior judgments while still reporting lower effort. We conclude that the benefits of quantile dotplots are driven by the additional information they provide. For people with less working memory, the extra effort of using the frequency information may cause them to ignore it and simply use the quantile dotplots like density plots.

We also found consistently poor performance for summary annotations compared to distributional visualizations. The summary annotations produced worse accuracy and evoked strategies associated with ignoring the uncertainty. Further, participants reported that textual expressions of uncertainty were the most effort-inducing and mentally demanding. In response to the debate about uncertainty visualization's efficacy compared to textual descriptions of uncertainty, we found that textual expressions of uncertainty performed consistently worse than quantile dotplots and density plots. We argue that these results provide converging evidence for how well-designed distributional visualizations of uncertainty can offload cognition during a task, whereas summary annotations of the same information do not share this advantage.

One recommendation from this work is to consider selecting an uncertainty visualization based on the target audience's unique characteristics. For example, previous research establishes a correspondence between working-memory deficits and children with attention deficit hyperactivity disorder, dyslexia, and dyscalculia [56]. Older adults also tend to show working-memory decline with normal cognitive aging [11]. Either density plots or quantile dotplots may be appropriate for these groups with known working-memory deficits. In contrast, some expert groups have high levels of working memory [61] and may experience outsized benefits from quantile dotplots. Designers will need to remain cognizant of potential tradeoffs between selective advantages for small groups and uniform experiences in larger populations.

4.1 Limitations and caveats

This work demonstrates how converging measures that include considerations of effort can produce a more representative picture of relative differences in visualization efficacy, but it also has several limitations. For one, there are many other potential converging measures that we did not test (for a review, see [40]). In particular, perceptual measures such as just-noticeable differences may offer insights into why various techniques show improved performance and require less effort [45]. Another limitation of this work is that we selected the communication techniques that we believed were representative of important categories of uncertainty communication. We want to emphasize that, although our results may allude to differences between categories, the results are reliable only for the particular communication techniques we tested. More work is needed to determine if these findings generalize to other communication techniques within each category, across different data types, or across different tasks and scenarios. Additional limitations result from our chosen method of data collection. Studies conducted exclusively online present a number of challenges for researchers, most notably determining the quality of participant engagement in cognitively demanding tasks. However, researchers have demonstrated that Prolific's quality control metrics significantly improve data quality over Amazon Mechanical Turk [76], and others have shown that the use of Amazon Mechanical Turk results in similar data to in-person samples on several of the most widely used psychological tasks [8]. Finally, an individual differences approach also has drawbacks, such as requiring large numbers of participants to examine the difference between groups, which may limit its utility in future research.

Due to the interdisciplinary and applied nature of this work, other caveats persist. We use a very general definition of working memory that does not account for a large body of research that defines working memory subcomponents and their relationships to attention (see [3, 18, 47]). We also operationalized working-memory capacity in a similarly general way and did not discuss different individual measures that examine various subcomponents of capacity limits (see [20, 44, 67]). As a result, this work cannot speak to the exact relationship between working-memory capacities and uncertainty communication. Additional work is needed to examine the effects of different working-memory components on decision-making with uncertainty visualizations.

5 Conclusions

Whereas prior studies demonstrated mixed findings concerning how individuals make decisions with various visual and textual depictions of uncertainty, this work provides converging evidence that quantile dotplots and density plots (i.e., distributional annotations) consistently outperformed interval plots, mean plots, and textual descriptions of uncertainty (i.e., summary annotations) in a resource allocation task. In particular, quantile dotplots performed as well as or better than all other techniques tested in all of the experimental conditions. Using two approaches for examining effort across uncertainty communication techniques, we found that both visualization methods and working-memory capacity impact participants' decisions. Given these results, we advocate for the inclusion of converging behavioral and subjective workload metrics in addition to accuracy performance to further disambiguate meaningful differences among visualization techniques.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation (# 2028374).

REFERENCES

[1] J. C. Aerts, K. C. Clarke, and A. D. Keuper. Testing popular visualization techniques for representing model uncertainty. *Cartography and geographic information science*, 30(3):249–261, 2003.

[2] A. Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[3] A. Baddeley. The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1413):1345–1350, 2001.

[4] A. D. Baddeley and G. Hitch. Working memory. *Psychology of learning and motivation*, 8:47–89, 1974.

[5] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005.

[6] T. Betsch and A. Glöckner. Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry*, 21(4):279–294, 2010.

[7] A. P. Boone, P. Gunalp, and M. Hegarty. Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of experimental psychology: applied*, 24(3):275, 2018.

[8] M. D. Buhrmester, S. Talaifar, and S. D. Gosling. An evaluation of amazon's mechanical turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2):149–154, 2018.

[9] S. Burigat and L. Chittaro. Pedestrian navigation with degraded gps signal: investigating the effects of visualizing position uncertainty. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pp. 221–230, 2011.

[10] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01

[11] S. Cansino, F. Torres-Trejo, C. Estrada-Manilla, M. Pérez-Loyda, L. Ramírez-Barajas, M. Hernández-Ladrón-deGuevara, A. Nava-Chaparro, and S. Ruiz-Velasco. Predictors of working memory maintenance and decline in older adults. *Archives of gerontology and geriatrics*, 89:104074, 2020.

[12] S. Castro, J. Cooper, and D. Strayer. Validating two assessment strategies for visual and cognitive load in a simulated driving task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, pp. 1899–1903. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[13] S. C. Castro, D. L. Strayer, D. Matzke, and A. Heathcote. Cognitive workload measurement and modeling under divided attention. *Journal of experimental psychology: human perception and performance*, 45(6):826, 2019.

[14] L. Cheong, S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham. Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty. *International Journal of Geographical Information Science*, 30(7):1377–1404, 2016.

[15] A. R. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle. Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, 12(5):769–786, 2005.

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

[17] M. Correll, D. Moritz, and J. Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2018.

[18] N. Cowan. The many faces of working memory and short-term storage. *Psychonomic bulletin & review*, 24(4):1158–1170, 2017.

[19] S. Deitrick. Evaluating implicit visualization of uncertainty for public policy decision support. In *Proc. AutoCarto*, 2012.

[20] R. W. Engle, M. J. Kane, and S. W. Tuholski. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. 1999.

[21] B. J. Evans. Dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences*, 23(4):409–422, 1997.

[22] A. Fagerlin, C. Wang, and P. A. Ubel. Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Medical decision making*, 25(4):398–405, 2005.

[23] D. Feldman-Stewart, M. D. Brundage, and V. Zotov. Further insight into the perception of quantitative information: judgments of gist in treatment decisions. *Medical Decision Making*, 27(1):34–43, 2007.

[24] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.

[25] E. Galy, J. Paxion, and C. Berthelon. Measuring mental workload with the nasa-tlx needs to examine each dimension rather than relying on the global score: an example with driving. *Ergonomics*, 61(4):517–527, 2018.

[26] R. Garcia-Retamero and M. Galesic. Communicating treatment risk reduction to people with low numeracy skills: a cross-cultural comparison. *American journal of public health*, 99(12):2196–2202, 2009.

[27] M. P. Gerrie and M. Garry. Individual differences in working memory capacity affect false memories for missing aspects of events. *Memory*, 15(5):561–571, 2007.

[28] G. Gigerenzer. The psychology of good judgment: frequency formats and simple algorithms. *Medical decision making*, 16(3):273–280, 1996.

[29] M. Greis, A. Joshi, K. Singer, A. Schmidt, and T. Machulla. Uncertainty visualization influences how humans aggregate discrepant information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.

[30] R. A. Grier. How high is high? a meta-analysis of nasa-tlx global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, pp. 1727–1731. SAGE Publications Sage CA: Los Angeles, CA, 2015.

[31] M. A. Grounds, S. Joslyn, and K. Otsuka. Probabilistic interval forecasts: An individual differences approach to understanding forecast communication. *Advances in Meteorology*, 2017, 2017.

[32] M. A. Grounds and S. L. Joslyn. Communicating weather forecast uncertainty: Do individual differences matter? *Journal of experimental psychology: applied*, 24(1):18, 2018.

[33] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[34] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[35] R. P. Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, 8:150, 2014.

[36] R. Hersher. Tens of thousands of alpacas die in peruvian cold snap. Jul 2016.

[37] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

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

[39] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE transactions on visualization and computer graphics*, 24(1):446–456, 2017.

[40] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2018.

[41] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11), 2015.

[42] S. Joslyn, L. Nemec, and S. Savelli. The benefits and challenges of predictive interval forecasts and verification graphics for end users. *Weather, Climate, and Society*, 5(2):133–147, 2013.

[43] S. Joslyn and S. Savelli. Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Frontiers in Computer Science*, 2:58, 2021. doi: 10.3389/fcomp.2020.590232

[44] M. A. Just and P. A. Carpenter. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122, 1992.

[45] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[46] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics*, 25(1):892–902, 2018.

[47] S. Kaplan and M. G. Berman. Directed attention as a common resource for executive functioning and self-regulation. *Perspectives on psychological*

science, 5(1):43–57, 2010.

[48] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2020. R package version 2.3.1. doi: 10.5281/zenodo.1308151

[49] M. Kay. *ggdist: Visualizations of Distributions and Uncertainty*, 2021. R package version 2.4.0. doi: 10.5281/zenodo.3879620

[50] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103, 2016.

[51] Q. T. Le, A. Pedro, C. R. Lim, H. T. Park, C. S. Park, and H. K. Kim. A framework for using mobile based virtual reality and augmented reality for experiential construction safety education. *International Journal of Engineering Education*, 31(3):713–725, 2015.

[52] I. M. Lipkus, G. Samsa, and B. K. Rimer. General performance on a numeracy scale among highly educated samples. *Medical decision making*, 21(1):37–44, 2001.

[53] L. Liu, L. Padilla, S. H. Creem-Regehr, and D. H. House. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE transactions on visualization and computer graphics*, 25(1):882–891, 2018.

[54] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. In *Proc. Eurographics Conf. Visualization*, pp. 20151115–127, 2015.

[55] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.

[56] C. Maehler and K. Schuchardt. Working memory in children with specific learning disorders and/or attention deficits. *Learning and Individual Differences*, 49:341–347, 2016.

[57] R. D. McKendrick and E. Cherry. A deeper look at the nasa tlx and where it falls short. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, pp. 44–48. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[58] G. McKenzie, M. Hegarty, T. Barrett, and M. Goodchild. Assessing the effectiveness of different visualizations for judgments of positional uncertainty. *International Journal of Geographical Information Science*, 30(2):221–239, 2016.

[59] S. M. Miran, C. Ling, A. Gerard, and L. Rothfusz. Effect of providing the uncertainty information about a tornado occurrence on the weather recipients' cognition and protective action: Probabilistic hazard information versus deterministic warnings. *Risk Analysis*, 39(7):1533–1545, 2019.

[60] D. B. Mitchell and R. R. Hunt. How much "effort" should be devoted to memory? *Memory & Cognition*, 17(3):337–348, 1989.

[61] C. D. Moore, M. X. Cohen, and C. Ranganath. Neural mechanisms of expert skills in visual working memory. *Journal of Neuroscience*, 26(43):11187–11196, 2006.

[62] K. J. Mulder, M. Lickiss, A. Black, A. J. Charlton-Perez, R. McCloy, and J. S. Young. Designing environmental uncertainty information for experts and non-experts: Does data presentation affect users' decisions and interpretations? *Meteorological Applications*, 27(1):e1821, 2020.

[63] G. E. Newman and B. J. Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic bulletin & review*, 19(4):601–607, 2012.

[64] T. S. Newman and W. Lee. On visualizing uncertainty in volumetric data: techniques and their evaluation. *Journal of Visual Languages & Computing*, 15(6):463–491, 2004.

[65] J. Nielsen. How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1:1–8, 1995.

[66] Y. Okan, E. Janssen, M. Galesic, and E. A. Waters. Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making*, 39(3):183–195, 2019.

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

[68] L. Padilla, S. C. Castro, and H. Hosseinpour. A review of uncertainty visualization errors: Working memory as an explanatory theory. *The Psychology of Learning and Motivation*, p. 275, 2021.

[69] L. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: A cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):29, 2018.

[70] L. Padilla, M. Kay, and J. Hullman. *Uncertainty Visualization*, pp. 1–18. Wiley StatsRef: Statistics Reference Online, 2021. doi: 10.1002/9781118445112.stat08296

[71] L. M. Padilla, S. C. Castro, P. S. Quinan, I. T. Ruginski, and S. H. Creem-Regehr. 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):332–342, 2019.

[72] L. M. Padilla, S. H. Creem-Regehr, and W. Thompson. The powerful influence of marks: Visual and knowledge-driven processing in hurricane track displays. *Journal of experimental psychology: applied*, 26(1):1, 2020.

[73] L. M. Padilla, M. Powell, M. Kay, and J. Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 11, 2020.

[74] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive research: principles and implications*, 2(1):1–16, 2017.

[75] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, 116(2):220, 1994.

[76] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[77] *Prolific*. (2014), Prolific, Ltd. [Online]. Available: `https://www.prolific.co`.

[78] *Fraud detection*. Qualtrics, LLC (2014). Qualtrics [software]. Utah, USA: Qualtrics.

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[80] C. Ranasinghe, N. Schiestel, and C. Kray. Visualising location uncertainty to support navigation under degraded gps signals: A comparison study. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–11, 2019.

[81] M. Riveiro. Visually supported reasoning under uncertain conditions: Effects of domain expertise on air traffic risk assessment. *Spatial Cognition & Computation*, 16(2):133–153, 2016.

[82] M. Riveiro, T. Helldin, G. Falkman, and M. Lebram. Effects of visualizing uncertainty on decision-making in a target identification scenario. *Computers & graphics*, 41:84–98, 2014.

[83] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016.

[84] D. Spiegelhalter. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4:31–60, 2017.

[85] S. Tak, A. Toet, and J. van Erp. The perception of visual uncertaintyrepresentation by non-experts. *IEEE transactions on visualization and computer graphics*, 20(6):935–943, 2013.

[86] N. Unsworth, R. P. Heitz, J. C. Schrock, and R. W. Engle. An automated version of the operation span task. *Behavior research methods*, 37(3):498–505, 2005.

[87] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond "p¡ 0.05", 2019.

[88] E. A. Waters, A. Fagerlin, and B. J. Zikmund-Fisher. Overcoming the many pitfalls of communicating risk. In *Handbook of health decision science*, pp. 265–277. Springer, 2016.

[89] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[90] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686

[91] J. Yucheng, B. Cardoso, and K. Verbert. How do different levels of user control affect cognitive load and acceptance of recommendations. In *In Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making (RecSys 2017)*, pp. 35–42. CEUR-WS, 2017.

[92] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker. Mapping the landscape of covid-19 crisis visualizations. *To appear in CHI 2021*, 2021.

[93] K. Zinke, M. Zeintl, A. Eschen, C. Herzog, and M. Kliegel. Potentials and limits of plasticity induced by working memory training in old-old age. *Gerontology*, 58(1):79–87, 2012.