Variable Biases: A Study of Scientists' Interpretation of Plot Types Commonly Used in Scientific Communication

Laura E. Matzen[1],* Kristin M. Divis[1], Michael J. Haass[1] & Deborah A. Cronin[2]

[1]Sandia National Laboratories, [2]University of California, Davis

ABSTRACT

In scientific communication, there are visualization conventions that are widely used to convey uncertainty, such as representing the variability of a dataset with error bars. Yet prior research indicates that scientists frequently misinterpret error bars. In this study, we compared bar charts with error bars to four alternative visualizations: dot, box, violin, and density plots. Our goal was to determine whether these other plot types would produce fewer biases in interpretation relative to bar plots. Scientists who have experience generating and interpreting statistical graphs used plots to assess whether the difference between two datasets was statistically significant. Our results replicated the patterns of biases that have been observed in prior studies of error bar interpretation. However, we found that our participants still had the best overall performance for bar plots with error bars, because they were most familiar with this type of plot.

Keywords: Data visualization, statistical graphs, variability.

Index Terms: Human-centered computing ~ Visualization ~ Empirical studies in visualization

1 INTRODUCTION

Data visualizations can be highly effective tools for communicating information to viewers, enabling rapid comprehension of trends, outliers, patterns, or comparisons. However, there is often uncertainty in datasets, and it is difficult to develop visualizations that effectively convey information about variability or uncertainty [18,26]. Uncertainty can come from many different sources, including variability in the phenomenon being studied, uncertainty caused by the way in which the data was collected, and uncertainty introduced by data processing or modeling [4]. Even when the variability or uncertainty in a dataset can be characterized, visualizing that information can be difficult. Adding more information to a visualization can make it cluttered or confusing [4,13,18,30]. There is also no clear consensus on how to evaluate visualizations of uncertainty [18]. Finally, humans are notoriously bad at comprehending uncertainty [14,15] and have trouble interpreting visual representations of variability or uncertainty [7,25,28]. Even if a designer has done an excellent job of characterizing and visualizing the uncertainty in a dataset, the viewers bring visual-spatial biases [26] and cognitive biases [19,23,34] to bear when interpreting visualizations and they may still come away with the wrong conclusion. This problem can persist even when people have relevant domain expertise or repeated exposures to the same types of visualizations [9,24].

* lematze@sandia.gov

One way to address this problem is to test the impact of different representations of uncertainty on human reasoning. Although few studies have done direct comparisons of different types of uncertainty visualizations, there is a growing body of research in this area, exploring the intersection of data visualization and cognitive psychology. Research in this area will benefit the field of visual communication by developing a scientific understanding of visual-spatial biases, human comprehension of uncertainty, and how these factors impact viewers' interpretations of data visualizations.

One of the aspects of visual communication that has received the most attention from cognition researchers is the depiction of data in scientific publications. Many of the visualizations that are most commonly used in scientific publications (as well as newspapers, magazines, and other types of publications) present a summary of a dataset along with a representation of error or variability. One example is the "cone of uncertainty" used in visual representations of hurricane forecasts. The cone represents uncertainty in the weather models that are predicting the hurricane's path. Although this convention is widely used, it is often misinterpreted by viewers [5,6,27,33].

Statistical graphs, such as bar graphs with error bars, are also widely used in publications. Several studies have found that there are design tradeoffs for statistical graphs because the visual encodings chosen for the mean and error in a dataset changes viewers' interpretation of the data. In the case of bar graphs, viewers consistently demonstrate "within-the-bar bias," interpreting values within the bar as being more likely than values outside of the bar [9,24]. Even scientists who have extensive experience with creating, publishing, and interpreting bar graphs frequently misinterpret error bars [3].

Other visualizations that are commonly used to depict a summary of a dataset at a glance include scatterplots, box plots, violin plots, and density plots. Each of these methods has pros and cons from the perspective of human interpretation. Scatterplots avoid many potential sources of cognitive bias by displaying the data directly. However, this type of visualization is not feasible for large datasets and puts the onus on the viewer to extract information such as the central tendency of the dataset. Box plots explicitly encode the median, upper and lower quartiles, upper and lower extremes, and the outliers in a dataset [22]. However, data sets with very different distributions can produce identical box plots [8]. Violin plots attempt to address this drawback by encoding the distribution of the data in the width of the bar [16]. Past research has indicated that violin plots can mitigate some of the biases that have been observed for non-experts interpreting bar graphs, such as within-the-bar bias [9].

Despite the known drawbacks of bar charts and many available alternatives, they are still very widely used in publications. In the present study, we sought to expand the prior research on how experts interpret visual summaries of datasets. We compared

participants' interpretations of visualizations that represented the central tendency and variability of datasets in different ways. Our participants represented the intended audience of most scientific publications: professional scientists who have extensive experience with creating and interpreting statistical graphs. We presented them with plots comparing two data sets and asked them to judge whether there was a statistically significant difference between the two data sets. The same underlying datasets were presented in six different ways. One version showed all of the data points in the form of a jittered dot plot. The other five visualizations showed different summary representations of the data. These included two types of plots that are widely used in scientific papers: bar plots with error bars showing the standard error of the mean and bar plots with error bars showing the 95% confidence intervals. We also included three types of visualizations that may reduce cognitive biases, but are less commonly used: violin plots, box plots, and density plots.

We used eye tracking to lend insight into which parts of each visualization the participants used to make their decisions and to test whether this changed based on the difficulty of the task (i.e., the magnitude of the difference between the two datasets). We predicted that participants would be best at judging whether the two data sets were significantly different when the p-value was either very high or very low, with poorer performance for p-values that were close to 0.05. In addition, we predicted that participants would have better performance for visualizations that were familiar to them and widely used in their domains of research. Finally, we predicted that participants would perform best when using visualizations that were a good fit for the task [1,12,17,20]. In this case, the bar plots and box plots explicitly encode the central tendency and the variability of the data sets, making them a good fit for judging whether the difference between two data sets is statistically significant. The other types of visualizations allowed participants to infer this information but did not mark it explicitly.

2 METHOD

2.1 Participants

Fifteen participants were recruited from the University of Illinois community (5 males, mean age = 31.87 years, stdev = 10.84 years) and compensated $20 for their time. All participants were required to have training in statistics and authorship on at least one paper containing graphs with visual representations of variability that had been published in a scientific journal. The participants who met these criteria were graduate students, faculty, and research staff. Participants were tested for color vision deficiencies (24 plate Ishihara Test; Ishihara, 1972) and near vision acuity. The data from one participant was dropped and replaced due to colorblindness.

2.2 Materials

All of the stimuli were created in R Software [31] from simulated data using the ggplot2 software package [35]. Twenty unique pairs of datasets were created by crossing number of data points per group (n=25 or n=100), variance in the data (low or high), and p-values for an unpaired, two-tailed t-test of the difference between the two variables ($p < .001$, $.01 < p < .03$, $.04 < p < .05$, $.09 < p < .12$, and $.40 < p < .60$) . Each data set contained two independent variables (labeled A and B) drawn from Gaussian distributions. Each dataset was plotted 6 ways: a bar plot with error bars showing the standard error of the mean (SE), a bar plot with error bars showing 95% confidence intervals (CI), a jittered dot plot, a violin plot, a box plot (with 1.5 * Inter-Quartile Range error bars), and an overlaid density plot, leading to 120 different plots. The study used a within-subjects design where every participant saw all 120 stimuli. Examples of each type of plot using the same underlying data are shown in Figure 1.

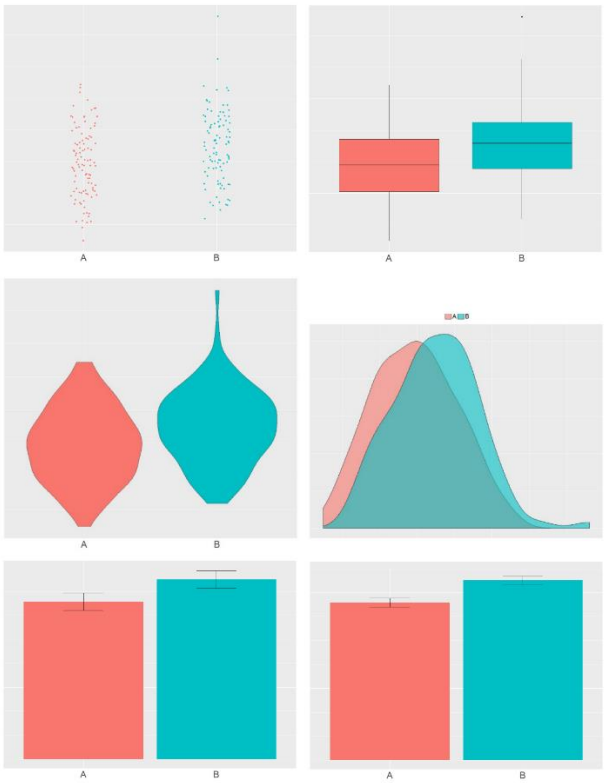

Figure 1. Example stimuli for two datasets with high N, high standard deviation, and a p-value < .001 when compared with a t-test. From top left: jittered dot plot, box plot, violin plot, density plot, bar graph with error bars showing the 95% confidence interval, and bar plot with error bars showing the standard error of the mean. When participants saw the box plots and bar plots, those plots were always accompanied by text indicating what type of error bar was used.

Prior to completing the main task, participants answered a survey about their experience with statistics and statistical graphs. They were asked to rate their statistical knowledge on a 1-5 scale where 1 was poor, 3 was moderate, and 5 was strong. To ensure that they had the requisite knowledge, they were asked to explain what it means when a t-test has a p-value of 0.01. They were also asked about the commonly accepted value for statistical significance in their field of research and what types of visualizations they use most frequently when presenting or publishing their own work.

After completing the main task, participants completed a second survey in which they were asked about their familiarity with the different types of visualizations that were used in the study.

2.3 Procedure

Participants completed the experiment individually in a dark room, seated at a nominal viewing distance of 0.8 meters from a computer monitor (.932 x .523 meters; 1920 x 1080 pixels). Their eye movements were recorded with a Smart Eye Pro eye tracker. Prior to completing each task, participants underwent the standard Smart Eye camera setup procedure and 9-point calibration. All stimuli were presented in the center of the screen on a white background. Each image was 1000 pixels in height, with a variable width to maintain the aspect ratios of the stimuli. For all

tasks, each stimulus was preceded by a fixation cross that was presented in the center of the screen for 1000 milliseconds (ms). There was a 500 ms interstimulus interval between trials.

In the main task, participants viewed each of the 120 stimuli in random order. The task was self-paced unless the participant did not advance within 10 seconds (at which time the program automatically advanced). After viewing each stimulus, the participant responded to two questions using mouse clicks:

1) Is the difference between Groups A and B statistically significant ($p < .05$)?

2) How confident are you in your response?

The response to the second question was given using a Likert scale from 1 to 5, with 1 being not at all confident, 3 being moderately confident, and 5 being very confident.

3 Behavioral Results

3.1 Survey

Participants had an average of 8.7 years of experience in research requiring statistical inference (stdev = 5.9 years). Most participants primarily worked in a field related to psychology or cognitive neuroscience (13 participants); one each was in biology or engineering. The average rating on the statistical knowledge question was 3.6 (stdev = 0.8). Every participant adequately described the meaning of a p-value of 0.01 in a t-test (as determined by two independent raters). When asked what kinds of visualizations they use to present their own research, seven of the participants mentioned bar graphs. Scatterplots, line graphs, and violin plots were mentioned by 2-3 participants each. No other visualization type was listed by more than one participant.

On the post-task survey, three participants said that they were familiar with all the visualization types used in the task. Eight participants indicated that they were unfamiliar with the box plot with inter-quartile range error bars. Three participants were unfamiliar with the violin plots and another three participants were unfamiliar with the density plots.

3.2 Significance Judgment Task

To analyze the behavioral results from the participants' primary task, we pulled plot type, confidence, p-value range, sample size, and variance together as fixed effects in a mixed effects model with a random effect for participant using the lme4 package in R software [2]. These models were used to predict accuracy, confidence, and response times (using Sattherwaite approximations to degrees of freedom).

The average accuracy, response time, and confidence rating for each chart type and p-value range is shown in Figure 2. We found that participants were significantly faster ($t(1785) = 6.18$, $p < .001$), more accurate ($Z = 3.26$, $p = .001$), and more confident in their responses ($t(1785) = 10.68$, $p < .001$) for the bar graphs relative to the other types of visualizations. When comparing the two types of bar plots to one another, participants were slightly more accurate ($Z = 1.72$, $p = .086$) and more confident ($t(1785) = 1.88$, $p = .06$) when using the bar plots with SE error bars, but these differences did not reach statistical significance.

As predicted, participants performed best when the differences between the two datasets were very large ($p < 0.001$) or very small ($p > 0.40$). We found significant differences in accuracy between all adjacent ranges of p-values (all $Z > 4.9$, all $p < .001$), with highest accuracy for the extreme ends of the scale and the lowest accuracy for p-values close to 0.05. Participants were significantly more confident for the extreme p-value ranges than for the ranges closer to 0.05 (all $t > 4.59$, all $p < .01$ for comparisons between the highest and lowest p-value ranges and the intermediate ranges). In general, the participants' confidence tracked closely with their accuracy, indicating that they were good judges of their own ability to assess the visualizations.

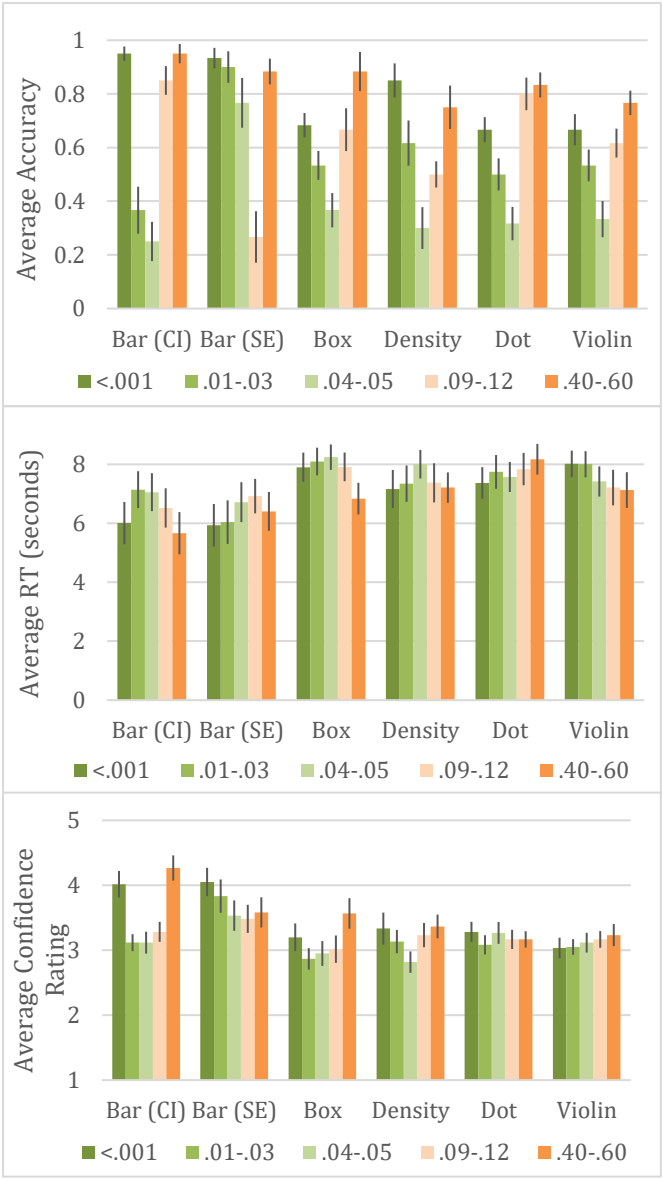

Figure 2: Average accuracy (proportion correct), response time, and confidence ratings for each plot type and p-value range. Visualizations that showed statistically significant differences are shown in shades of green and visualizations that did not are shown in shades of orange. Error bars represent within-subjects standard error of the mean.

When looking at the interaction between plot type and p-value level, an interesting pattern emerged. For the box, density, dot, and violin plots, we observed the same general pattern of lower accuracy for p-values close to 0.05 and higher accuracy for more extreme values, as shown in Figure 2. However, the pattern for the two types of bar plots looked different. The accuracy results for

the two types of bar plots are reorganized in Figure 3 to allow for easier comparisons of the results for the different p-value ranges.

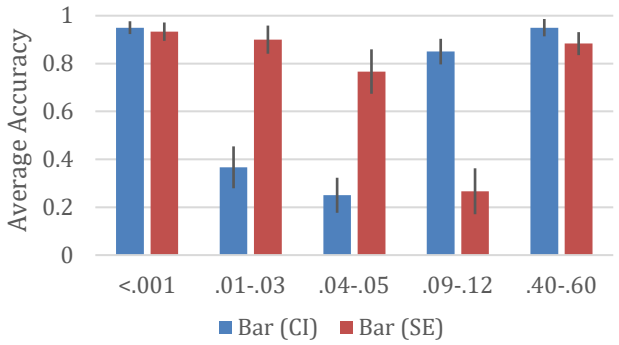

Figure 3: Average proportion correct at each p-value range for the two types of bar plots. Error bars represent within-subjects standard error of the mean.

We ran a simple mixed effects model for each type of bar chart, predicting accuracy from the fixed effect of p-value range and random effect of participant. For both types of bar graphs, the participants were highly accurate for $p < 0.001$ and $p > 0.40$. However, the middle ranges revealed differing patterns. For bar charts with SE error bars, participants also performed well for the .01-.03 and .04-.05 p-value ranges. Only the .09-.12 p-value range displayed significantly lower accuracy than the other p-value ranges. In contrast, the bar charts with 95% CI error bars had significantly lower accuracy in the .01-.03, .04-.05, and .09-.12 p-value ranges relative to the two most extreme p-value ranges. In other words, the participants consistently misinterpreted CI error bars, believing that significant differences were not significant. The opposite misinterpretation was common for the SE error bars, with participants indicating that the difference between two datasets was significant when the p-value was between 0.09 and 0.12.

3.3 Eye Movement Results

Fixations were calculated using Smart Eye's default algorithm, where any sample for which the velocity over the preceding 200 ms is less than 15°/s is deemed a fixation. The first fixation in each trial was excluded from the analysis, as was any fixation with a duration less than 100 ms. See Figure 4 for the average number of fixations for each plot type and p-value range.

A mixed effects model predicting total number of fixations in an image from the fixed effects of plot type, p-value range, sample size, standard deviation, accuracy, and confidence along with random effects for participant and stimulus (with Satterthwaite approximation of degrees of freedom) revealed that bar plots tended to receive fewer fixations than the other plots overall ($t(125) = 8.07$, $p < .001$), with bar plots with SE error bars receiving fewer fixations than those with 95% CI error bars ($t(117) = 5.55$, $p < .001$). No significant differences (at an $\alpha < .05$ level) in number of fixations to an image were found when comparing adjacent p-value ranges (e.g., $p < .001$ compared to $.01 < p < .03$). Numerically, participants had higher numbers of fixations when they responded incorrectly than correctly, but that trend did not reach statistical significance in our model ($t(1346) = 1.71$, $p = .088$).

An analysis of average fixation duration for each type of plot showed that bar plots tended to receive longer fixations than the other plots overall ($t(142) = 7.29$, $p < .001$), with bar plots with SE error bars having longer fixations than those with 95% CI error bars ($t(118) = 4.91$, $p < .001$).

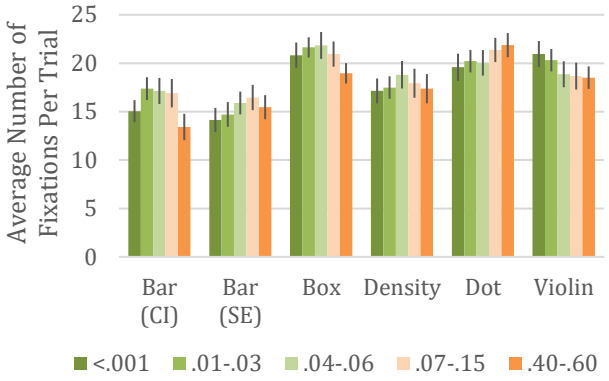

Figure 4. Average number of fixations per trial for each plot type (top) and each plot type broken down by p-value range (bottom). Error bars represent within-subjects standard error of the mean.

Given the differences in the behavioral results for the two types of bar graphs, we compared the pattern of fixations to different regions of interest (ROIs) for these graphs. ROIs were created for the A Data, B Data, A Label, B Label, Error Bars, and Other (i.e., background regions of the visualization that were not contained in any other ROI). Figure 5 shows the proportion of fixations to each of these ROIs. A mixed effects model predicting the proportion of fixations to each ROI with fixed effects for type of bar chart (SE or CI error bars) and ROI type, along with random effects for subject revealed that the plots with SE error bars had a significantly higher proportion of fixations to the A Data ROI ($t(2985) = 2.089$, $p = .037$) and the plots with 95% CI error bars had a significantly higher proportion of fixations to the Error Bars ROI ($t(2985) = 4.223$, $p < .001$). No significant differences in the proportion of fixations were found between the two types of bar plots for the B Data or Other ROIs (all p-values > .05).

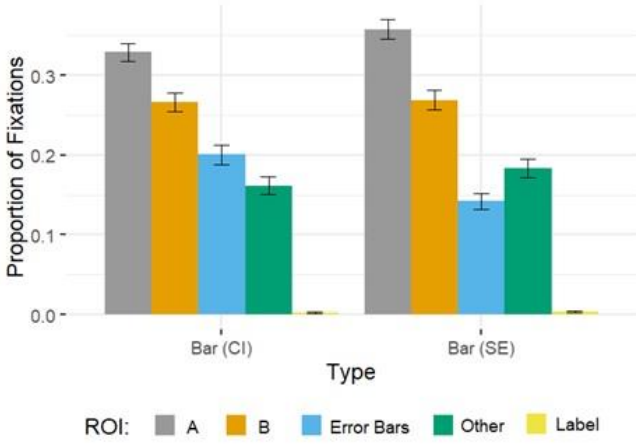

Figure 5. Average proportions of fixations per trial for each region of interest for the two types of bar plots. Error bars represent within-subjects standard error of the mean.

4 Discussion

When participants were asked to determine whether the difference between two groups of data was statistically significant, the way the data were visualized influenced the accuracy of their judgments, their response times, their confidence in their decisions, and their eye movement behavior. Overall, participants had the best performance for the bar graphs. They were faster,

more accurate, and more confident for the two types of bar graphs than for the box, density, dot, or violin plots. They also had fewer, longer fixations for the bar plot stimuli than for the other types of visualizations. This pattern is consistent with prior research on visual search and scene perception, which indicates that people have shorter fixations when searching for information and longer fixations when viewing the most informative regions of scenes [32]. When comparing the two types of bar plots to one another, we found that the participants had fewer, longer fixations to the bar graphs with SE error bars. However, they were no faster and only marginally more accurate and confident with bar charts with SE error bars relative to the 95% CI error bars.

Although prior research has indicated that violin plots, which explicitly encode the distribution of the data, can reduce some of the cognitive biases associated with bar plots [9], that research used novice participants who had minimal experience with interpreting statistical graphs. In the present study, we found that participants who are experienced with creating and interpreting statistical graphs performed best when using bar plots. It is important to note that our task was different from tasks that have been used in prior studies, and it was intended to test the participants' ability to interpret the central tendency and variability in the datasets. There are two reasons that the bar plots may have led to superior performance. First, the bar plots explicitly encode the information that was important for this task through the bar height and the error bars. The dot, violin and density plots allowed participants to infer this information but did not mark it explicitly. While participants can summarize information that is not explicitly marked (e.g., [11,21]), explicit marking may lead to easier or more effective comprehension of that underlying element. The explicit markings on the bar plots may have provided more support to the participants for this task. However, note that the box plots also provide explicit markings of central tendency and variability, yet they consistently led to worse performance than the bar plots in terms of accuracy, response times, and confidence. This leads us to the second reason that the bar plots outperformed all the other plots: familiarity. The survey results indicated that the participants were more familiar with bar plots than with any other type of plot, while they were least familiar with the box plots. Our results indicate that this lack of familiarity affected the participants' performance, negating the benefit that we would expect to see based on the fact that the box plots explicitly marked information that was crucial to the task.

Even though our participants performed best for the bar plots, we still observed biases in their responses. For the bar plots with SE error bars, participants frequently made the mistake of saying that differences in the $0.09 < p < 0.12$ range were significant. Their responses were biased in the opposite direction for the bar plots with 95% CI error bars. For these plots, participants frequently said that differences in the $0.01 < p < 0.05$ range were not significant. These results are consistent with prior studies that have found that researchers consistently misinterpret error bars [3]. In the study by Belia and colleagues, participants were asked to adjust a pair of error bars to reflect a difference equivalent to $p = 0.05$. They found that participants were generally too strict with CI error bars (with a mean response corresponding to $p = 0.009$) and too lax with SE error bars (with a mean response corresponding to $p = 0.109$). The results of our study mirror this same pattern of interpretation.

Patterns of eye movements across the ROIs for the two types of bar graphs were largely similar, although participants devoted a significantly lower proportion of their fixations to the error bars in the plots with 95% CI error bars. This pattern, combined with the behavioral results, could indicate that the participants were less comfortable with those types of error bars, even though all of the participants reported that they were familiar with them. The majority of the participants worked in the field of psychology, and while the field has been pushing for the use of CIs for many years, they are still not widely used in psychology publications [10], with SE error bars being used much more frequently.

5 Conclusions

This experiment represents a small step toward creating a better understanding of how factors such as task, experience, familiarity, and biases impact how viewers interpret common types of data visualizations. It highlights the importance of considering factors beyond low-level visual properties when designing visualizations to communicate information. This area of research is particularly important in the context of developing visualizations that can effectively convey information about uncertainty. Using error bars to show the variability in a dataset is one of the most well-known and straightforward ways to convey uncertainty. Yet even people who have experience with creating and interpreting plots with error bars are subject to biases in their interpretation of these visual cues.

Our experiment points to the need for more empirical research that makes direct comparisons between different types of visual representations, particularly for visualizations that incorporate information about variability or uncertainty. As we observed in this study, there are often consistent patterns in how viewers interpret specific cues. By identifying these systematic effects, we can build on the growing literature on visualization cognition and identify methods for reducing these perceptual and cognitive biases. This will improve the ability of visualization designers to communicate effectively and to design visualizations with the cognitive needs and limitations of their audience in mind.

6 Acknowledgments

We would like to thank Andy Wilson, James Crowell, Camille Goudeseune, Hank Kaczmarski, and Alejandro Lleras for their assistance. This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] Bandlow, A., Matzen, L. E., Cole, K., Dornburg, C., Geiseler, C., Greenfield, J., McNamara, L. & Stevens-Adams, S. (2011, July). Evaluating Information Visualizations with Working Memory Metrics. In *International Conference on Human-Computer Interaction* (pp. 265-269). Springer, Berlin, Heidelberg.

[2] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

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

[4] Bonneau, G. P., Hege, H. C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., & Schultz, T. (2014). *Overview and state-of-the-*

art of uncertainty visualization. In Scientific Visualization (pp. 3-27). Springer, London.

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

[6] Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society, 88(5)*, 651-668.

[7] Chi, M. T., Glaser, R., & Farr, M. J. (2014). *The nature of expertise*. Psychology Press.

[8] Choonpradub, C., & McNeil, D. (2005). Can the box plot be improved. *Songklanakarin Journal of Science and Technology*, *27*(3), 649-657.

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

[10] Cumming, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60 (2),* 170-180.

[11] Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 12*(19), 2316-2325.

[12] Goldberg, J., & Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information visualization*, *10*(3), 182-195.

[13] Griethe, H., & Schumann, H. (2006, March). The visualization of uncertain data: Methods and problems. In *SimVis* (pp. 143-156).

[14] Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*(3), 411-435.

[15] Heuer, R. J. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence.

[16] Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181-184.

[17] Ho, H. Y., Yeh, I. C., Lai, Y. C., Lin, W. C., & Cherng, F. Y. (2015, June). Evaluating 2D flow visualization using eye tracking. In *Computer Graphics Forum* (Vol. 34, No. 3, pp. 501-510).

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

[19] Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

[20] Li, J., Martens, J. B., & Van Wijk, J. J. (2010). Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, *9*(1), 13-30.

[21] Kramer, R. S. S., Telfer, C. G. R., & Towler, A. (2017). Visual comparison of two data sets: Do people use the means and variability? *Journal of Numerical Cognition*, *3*(1), 97-111.

[22] McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, *32*(1), 12-16.

[23] Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136-164.

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

[25] Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., ... & Chang, R. (2016). Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics, 22(1)*, 529-538.

[26] Padilla, L., Creem-Regehr, S., Hegarty, M., & Stefanucci, J. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications*

[27] Padilla, L., Creem-Regehr, S. H., & Thompson, W. (2019). The powerful influence of marks: Visual and knowledge-driven processing in hurricane track displays. *Journal of experimental psychology: applied*.

[28] Padilla, L., Ruginski, I. T., & Creem-Regehr, S. H. (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive research: principles and implications, 2*(1), 40.

[29] Padilla, L., Kay, M., & Hullman, J. (In Press). Uncertainty Visualization. In R. Levine (Ed.), *Handbook of Computational Statistics & Data Science*: Springer Science.

[30] Potter, K., Rosen, P., & Johnson, C. R. (2011, August). From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *IFIP Working Conference on Uncertainty Quantification* (pp. 226-249). Springer, Berlin, Heidelberg.

[31] R Development Core Team (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://r-project.org (ISBN 3-900051-07-0).

[32] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3).

[33] Ruginski, I. T., Boone, A. P., Padilla, L., Liu, L., Heydari, N., Kramer, H. S., . . . Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation, 16*(2), 154-172.

[34] Saccone, E. J., Landry, O., & Chouinard, P. A. (2019). A meta-analysis of the size-weight and material-weight illusions. *Psychonomic bulletin & review*, *26*(4), 1195-1212.

[35] Wickham (2009). *ggplot2: Elegant graphics for data analysis*, Dordrecht New York: Springer. Retrieved from http://ggplot2.org (ISBN 978-0-387-98140-6).