An Instrument for Evaluating Uncertainty Visualization Techniques

Andrea Brennen
In-Q-Tel Labs
Waltham, MA 02451, USA
abrennen@iqt.org

Stephanie Tuerk
In-Q-Tel Labs
Waltham, MA 02451, USA
stuerk@iqt.org

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188649

Abstract
Today's data visualization tools offer few capabilities and no representational standards for conveying uncertainty. Our aim is to remedy this by creating a visual vocabulary for uncertainty in data. However, we must first develop an extensible methodology for validating the effectiveness of uncertainty visualization techniques. In this paper we describe a test instrument we have developed to collect empirical data concerning four measures — accuracy, response time, reported confidence, and cognitive load — that can be used to evaluate techniques for visualizing data with uncertainty.

Author Keywords
Data Visualization, Evaluation, Uncertainty, Cognitive Load, Amazon Mechanical Turk.

ACM Classification Keywords
H.5.m Information interfaces and presentation (e.g., HCI): Miscellaneous; H.5.2 Information interfaces and presentation (e.g., HCI): User Interfaces: Evaluation/methodology

Introduction
Data visualizations and information graphics are an increasingly popular medium for communicating quantitative information to decision-makers and to the

Note on Cognitive Load

Cognitive Load Theory, first developed by psychologist John Sweller in the 1980s, is based on an assumption that a person's cognitive capacity in working memory is limited. Unnecessary complexity can consume working memory and detract from learning. The theory differentiates between *Intrinsic Load* (IL), which comes from the inherent difficulty of a task (relative to one's prior knowledge) and *Extrinsic Load* (EL), which is imposed by instructional features that are not beneficial for learning. Researchers have advocated that an explicit aim of instructional design should be to minimize Extrinsic Load [8] and we do not think it is much of a leap to assume this aim would benefit visualization design as well.

public. Charts, plots, maps and interactive infographics help summarize large quantities of data, reveal or highlight relationships in that data, and communicate quantitative information in a way that is engaging and understandable. However, each step in the process of collecting, processing, analyzing and visualizing data introduces the potential for error, bias and uncertainty. If data visualizations do not explicitly convey these factors, they may imply more certainty than is (or can be) known about underlying data. This can encourage overly-confident interpretations of the data depicted, and impact decision-making in undesirable ways. We believe that relevant uncertainties in data and data analysis should be represented, to the extent that they are known. However, today's data visualization tools offer few capabilities and no representational standards for conveying uncertainty. To remedy this, we are building upon existing conventions to create a robust visual vocabulary for conveying uncertainty in data and supporting the development of technologies that will help people use this vocabulary successfully. We hope to make uncertainty a "first class citizen" in data visualization. To achieve this aim, we need an extensible method of validating visualization techniques.

In this paper we propose criteria for evaluating uncertainty visualizations and describe an instrument we developed to measure quantitative indicators of these criteria. We plan to use this instrument to help us develop and validate a visual vocabulary for uncertainty. We imagine this work will benefit both producers and consumers of data visualizations by helping visualization producers communicate their results in a more complete way, and by making visualization consumers (decision-makers, in particular) more aware of common sources of uncertainty that arise in data collection and analysis. While our work is still very much in progress, we seek feedback from the CHI community to help guide our future efforts.

Background

There are vast gaps in our collective knowledge about how people interpret data visualizations. These are exacerbated for visualizations that involve difficult concepts like uncertainty, which create countless opportunities for inconsistencies between the information one encodes into a visualization and the information someone else retrieves from that visualization. Alan MacEachern made the important recommendation that uncertainty should be visually encoded through "free" visual variables that are not being used to represent other information [10]. Boukhelifa et al recommended using "imprecise" visual variables such as "blur" or "sketchiness" to convey uncertainty, suggesting that these encodings intuitively imply a lack of precision [3]. However, despite recommendations such as these, there is little empirical data on the relative effectiveness of uncertainty visualization techniques. Recently, a small number of empirical studies have extended the precedent of graphical perception tests to uncertainty visualizations. For example, Correll and Gleicher identified limitations of error bars [5] and Hullman et al suggested the value of a novel animated visualization technique — Hypothetical Outcome Plots (HOPs) — by demonstrating situations in which this technique outperformed conventional plot types [7]. However, this existing work is far from comprehensive.

In the context of controlled experiments, *accuracy* and *response time* are the most common criteria used to

evaluate data visualizations. A common assumption is that a "better" visualization makes it easier for respondents to retrieve information, leading to quicker response times. However, complications arise when uncertainty visualizations are evaluated based solely on accuracy and response time. First, the complexity added by visualizing uncertainty may make a slightly longer viewing time desirable if it helps a viewer more thoroughly understand the data. Second, in the context of a controlled experiment, it is difficult to distinguish between inaccurate responses due to an ineffective visualization and those that are the result of a user's judgment or a reflection of his or her prior experience (e.g. with the topic, with statistics in general, with uncertainty indicators, etc.) Third, in some situations the purpose of visualizing uncertainty might be to diminish how decisively a viewer can draw conclusions from that data. In these cases, time and accuracy measurements tell us little about the success of a specific technique.

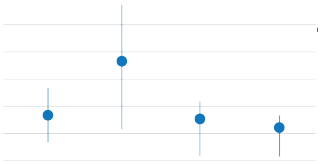
Given these confounding factors, we suggest two additional criteria for evaluating uncertainty visualizations: *reported confidence* and *cognitive load,* a measure of the difficultly one has learning new information and completing tasks that require an understanding of that information. We assume that an uncertainty visualization is more effective when it is associated with higher accuracy, lower cognitive load, and confidence assessments that accurately reflect the certainty of the data. While prior studies [2], [12] have incorporated cognitive load into criteria for evaluating visualizations, we know of no efforts to compare the cognitive load imposed by uncertainty visualization techniques.

Method

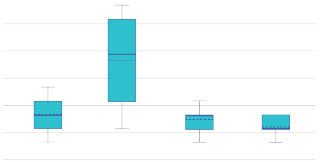
We hypothesize that the choice of technique used to display uncertainty in a visualization impacts how well a viewer is able to comprehend the content of that visualization. In this section we describe the study design and test instrument we developed to test this hypothesis. We use a between-subjects study design where visualization technique (or plot type*)* is treated as the independent variable. We use our test instrument to measure four dependent variables: *accuracy*, *reported confidence*, *response time*, and *cognitive load.* This test instrument requires several assets: a scenario, a dataset with uncertainty, multiple visualizations of the uncertainty in the dataset, comprehension questions (designed to measure respondents' understanding of the uncertainty), and cognitive load questions (that ask participants to self-report the mental effort required to answer the comprehension questions).

We recruit participants from Amazon Mechanical Turk, a precedent established in previous visualization experiments [6]. Then, we randomly assign each participant to one of several test groups and administer our test instrument. All respondents receive the same scenario description and are asked the same comprehension and cognitive load questions, but each group receives a different visualization. With this set up, we can attribute statistically significant variations in responses among the groups to be an effect of the visualization. Over time, we can replicate this extensible format for different visualizations, datasets, and scenarios, allowing us to build out a vocabulary of validated techniques.

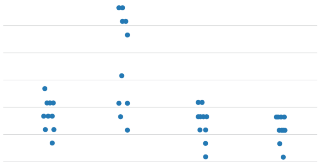
Figure 1: Four techniques for visualizing cardinal uncertainty. (A fifth is shown in Figure 2.)



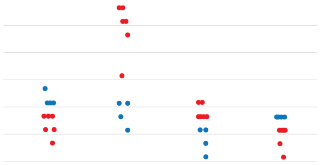
Dot-extents (dot with error bar)



Box Plot



Dot (scatter) Plot, with jitter



Varied Dot Plot, with jitter

Measures

- ***Accuracy.*** We measure accuracy with two types of comprehension questions. One type asks respondents to retrieve a specific piece of information from the visualization. The second type asks them to make a judgment about the data that accounts for the uncertainty shown. In both cases, respondents are given a statement and asked to determine if that statement is true or false.
- ***Reported Confidence.*** Respondents answer these questions in a way that requires them to provide a confidence assessment. They are asked to position a slider on scale ranging from -100 to +100, where "+100" indicates they are "completely certain" the given statement is true, and "-100" indicates they are "completely certain" the statement is false. They may select any integer from 1 to 99 (or -1 to -99) to indicate a confidence assessment that is less than "completely certain" or select "0," to indicate the statement is "equally likely to be true or false."
- ***Response Time.*** Following a precedent set by previous visualization studies, we record the time required for participants to answer questions.
- ***Cognitive Load***. We measure cognitive load by asking respondents to self-report their perceived mental effort after each comprehension question [2]. Specifically, we ask respondents to rate the ease of each question on a 5-point Likert scale ranging from "very easy" to "very difficult" [9]. At the completion of the test, respondents are also asked to rate the overall "usefulness" of the visualization.

Test Assets

The type of uncertainty contained in the test dataset should inform the selection of uncertainty visualization techniques. This is important because our instrument provides a method for evaluating a set of techniques relative to one another, but also, in relation to a particular type of uncertainty and one or more specific tasks (the comprehension questions). We recommend choosing one high-level uncertainty data type — geospatial, temporal, cardinal/numerical, categorical, or quality/reliability — and using a test dataset in which that uncertainty is described in one way — for example, as a probability distribution, a range of values, a nominal value, or on an ordinal scale. In Figures 1 and 2 we provide examples of test assets that we developed to evaluate techniques for visualizing *cardinal uncertainty* — uncertainty associated with counts, amounts, and numerical quantities. We developed a synthetic dataset describing the number of terrorism-related fatalities that occurred in "Country X" between 2010 and 2013, as reported by 10 different news sources. This synthetic dataset was inspired by the Armed Conflict Location and Event Data Project (ACLED), a curated repository of global conflict event data that includes fatality counts reported by different sources [1]. While it is not uncommon for different sources to report different fatality estimates for the same event, conflicting reports create uncertainty about the actual number of fatalities that occurred. Drawing on conventions, we selected five techniques for visualizing this uncertainty — a gradient, a dot with error bars, a box plot, a scatter plot, and a bi-color scatter plot. Figure 2 shows a visualization of the test dataset based on the "gradient" technique. The other four techniques are shown in Figure 1.

Preliminary Results

In December 2017, we conducted a study based on these test assets. We recruited 250 participants via Amazon Mechanical Turk and divided them into 5 test

groups. Each group was shown a different visualization of the same data and all participants were asked the same two comprehension questions, one of which is shown in Figure 2. A full analysis and discussion of results is forthcoming, but we provide a brief summary of preliminary results here, to indicate the validity of our test instrument. By deploying our test instrument, we were able to demonstrate statistically significant differences in the accuracy of participants' responses, the confidence they reported in those responses, and the (perceived) mental effort required to answer the comprehension questions. The data we collected supports our hypothesis, suggesting that showing participants different visualization of our synthetic dataset had a demonstrable effect on how well and how easily they were able to interpret the uncertainty in that dataset.

Figure 3 shows the percentage of participants in each test group who responded "True" to the comprehension questions. The Varied Dot Plot contained additional information that likely contributed to the much lower "True" response rate for that plot type, but a chi-squared test indicated that at the $p < 0.05$ level, there was a statistically significant effect of plot type on the rate of "True" responses among the other four plot types. An adjusted cognitive load or "ease" score for all five plot types is shown in black. A one-way analysis of variance (ANOVA) test indicated statistically significant differences in the mean cognitive load per plot type. Post hoc comparisons using the Tukey Honest Significant Difference (HSD) test [11] suggested statistically significant differences in "ease" for several pairwise comparisons. Figure 4 shows the mean reported confidence per plot type, calculated from the absolute values of all "True" and "False" responses. A

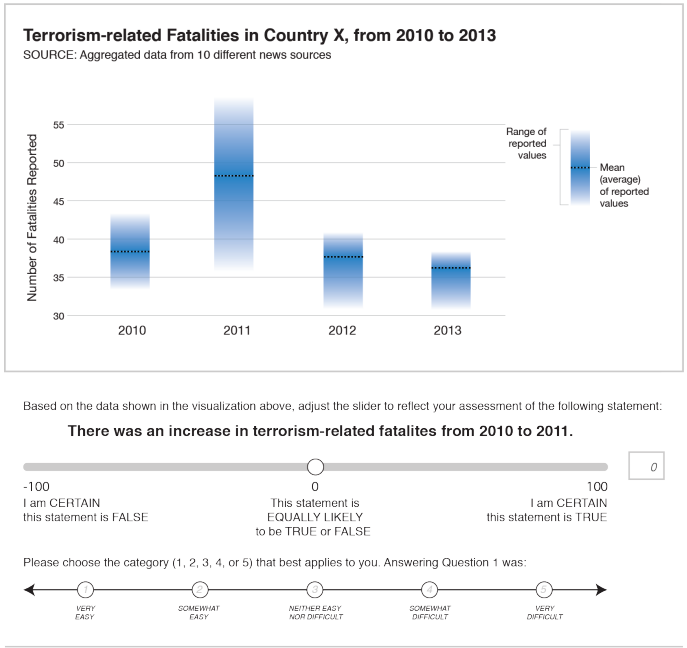


Figure 2: An example plot created for the instrument that uses a gradient to visualize the uncertainty in the test dataset.

one-way ANOVA test confirmed a statistical difference between the mean confidence values. A more thorough analysis of results is forthcoming, but here we offer this brief summary of preliminary results to indicate that our instrument appears to be a useful mechanism for collecting data on uncertainty evaluation techniques.

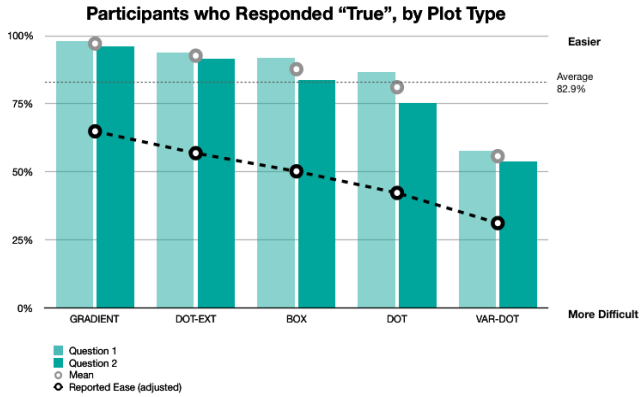


Figure 3: The percentage of participants in each test group who responded "True" to the comprehension questions.

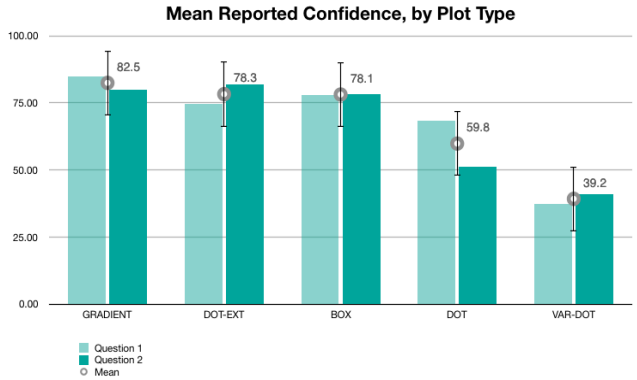


Figure 4: The mean reported confidence for each plot type, using the absolute value of all "True" and "False" responses.

References

1. ACLED. Armed Conflict Location and Event Data Project. http://www.acleddata.com/
2. Barnes, Spencer. 2016. Appearance and explanation: advancements in the evaluation of journalistic information graphics. *Journal of Visual Literacy*. Vol. 35, No. 3.
3. Boukhelifa, N., Bezerianos, A., Isenberg, T. and Fekete, J.D. 2012. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 18(12), 2769-2778.
4. Cleveland, William S. And Robert McGill. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*. Vol. 79, No. 387.
5. Correll, Michael and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics* 20(12), 2142-2151.
6. Heer, Jeffrey and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10).
7. Hullman, Jessica, Paul Resnick, Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLoS ONE* 10(11).
8. Kalyuga, Slava. 2011. Informing: A Cognitive Load Perspective. *Informing Science.* Vol. 14.
9. Leppink, J., F. Paas, C.P. Van der Vleuten, T. Van Gog, J.J. Van Merrienboer. 2013. Development of an instrument for measuring different types of cognitive load. *Behavioral Research Methods* 45(4).
10. MacEachren, A.M. 1992. Visualizing Uncertain Information. *Cartographic Perspectives* (13).
11. *NIST/SEMATECH e-Handbook of Statistical Methods* http://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm, 2018.
12. Zagermann, Johannes, Ulrike Pfeil, Harald Reiterer. 2016. Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (BELIV '16).