Why Evaluating Uncertainty Visualization is Error Prone

Jessica Hullman
Information School
University of Washington
jhullman@uw.edu

ABSTRACT

Evaluating a visualization that depicts uncertainty is fraught with challenges due to the complex psychology of uncertainty. However, relatively little attention is paid to selecting and motivating a chosen interpretation or elicitation method for subjective probabilities in the uncertainty visualization literature. I survey existing evaluation work in uncertainty visualization, and examine how research in judgment and decision-making that focuses on subjective uncertainty elicitation sheds light on common approaches in visualization. I propose suggestions for practice aimed at reducing errors and noise related to how ground truth is defined for subjective probability estimates, the choice of an elicitation method, and the strategies used by subjects making judgments with an uncertainty visualization.

Keywords

uncertainty visualization; subjective probability distribution; elicitation

1. INTRODUCTION

Visualizing uncertainty in data—in the form of variance, precision, accuracy, reliability or related concepts?—has a relatively long history. Francis Galton visualized a hypothetical distribution of heights back in 1869 [20]; visualizations were used in China in the earlier 1800's to show predictions related to children?s health [55]). Research and practice have since proposed many techniques.

But to quote a future agenda for visualization tools to help combat information overload: "There is no accepted methodology to represent potentially erroneous information... There is no agreement on factors regarding the nature of uncertainty, quality of source, and relevance to a particular decision or assessment." In fact, many visualizations that we encounter don't show uncertainty at all [19].

One possible reason for the lack of depiction of uncertainty in many situations where visualization is used–from business reporting to the media to scientific contexts–is that it is generally straightforward to show value in a visualization that does not depict uncertainty, but much more complex to show value in a visualization of uncertainty.

BELIV '16, October 24 2016, Baltimore, MD, USA

ISBN 978-1-4503-4818-8/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2993901.2993919

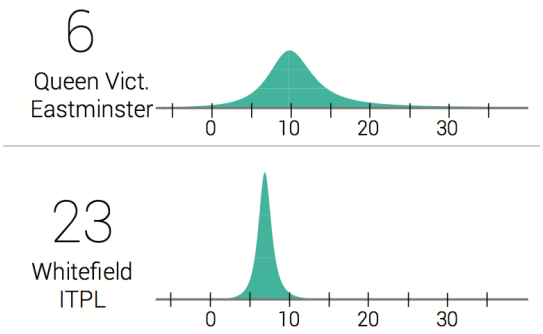

Figure 1: PDFs used to show uncertainty in predicted bus arrival times [36].

Consider, for example, the two probability density functions shown in Figure 1, each of which represent the probability of a given bus arriving at a stop at different times (where 0 represents the current time). For a user who wants to know which of the two busses, both of which are viable routes to her home, will arrive first, the expected value for each bus time distribution alone is sufficient information to make this binary decision. Is the presentation of uncertainty for each bus helpful? One could compare her ability to make informed decisions with and without the uncertainty, but for questions like this one, which does not require distributional information, it is unlikely that displaying uncertainty will prove helpful. Tasks that are appropriate to assess the depiction of uncertainty are more complex: e.g., What are the chances that the number 6 bus will arrive first? How probable is it that the number 6 bus will arrive within 5 minutes from now? That one of the busses will arrive within 5 minutes? By asking a number of such questions, a researcher can elicit evidence of the viewer's subjective probability distribution, and compare that distribution to the observed data.

However, even a seemingly simple question like "How probable is it that the number 6 bus will arrive first?" can be complex to answer. If subjects in a study do poorly, how can the researcher know whether the question was understood, implying a failure of the visual presentation, versus misinterpreted, implying a problem with training or the clarity of the task? If the participant answers a question correctly, what properties should other questions have to ensure that the

correct response was based on the full information provided, rather than on the use of common heuristics for making decisions under uncertainty? Determining the right question, the most direct way to ask it, and the appropriate ground truth to compare it to is particularly challenging for evaluators of uncertainty visualizations.

The reasons behind these challenges are varied. At a foundational level, epistemological questions about the correct interpretation of probability, and peoples' judgments of uncertainty, have long been debated by mathematicians and philosophers. Defining ground truth becomes a philosophical exercise rather than a standard experimental task. Additionally, evidence suggests that the validity of elicited probability distributions is highly sensitive to the elicitation method [21]. But determining these preferred methods is itself a complex task due to the ground truth dilemma. Numerous studies in psychology indicate that people are prone to various heuristics, biases, and other errors when responding to decisions involving uncertainty. For example, people often prefer not to see uncertainty information due to its complexity and abstractness, and may consequently ignore it.

Despite these challenges, very few papers in the visualization literature on uncertainty directly address the complexity of evaluation as a focus. Brodlie et al. [10] ask, Why is uncertainty so hard?, yet none of the nine reasons they propose relate to the psychology of interpretation. Harrower [28] suggests shifting the focus in uncertainty visualization evaluation to aspects of process rather than the outcome alone, including understanding the learning curve associated with understanding representations and how uncertainty changes thinking with a visualization; few studies, however, attempt this level of understanding. The goal of this paper is to provide a discussion of the psychology of eliciting and measuring uncertainty for a visualization audience.

I first outline existing evaluation paradigms used for uncertainty visualization. I discuss how research around the philosophy of uncertainty, the elicitation of subjective probabilities, and the role of heuristics in reasoning shed light on under-discussed aspects of evaluation. I conclude with a discussion of potential remedies and areas for future work.

2. BACKGROUND

2.1 Types of uncertainty

Taxonomies frame uncertainty as symptomatic of processes used to create a visualization, from data collection to dissemination ([41, 43, 48, 26, 57, 53]). Several important distinctions are made within the visualization literature and broader literature alike regarding the source of uncertainty. First, uncertainty can arise from randomness inherent in a process, such as the random error that results from attempts to measure a parameter (e.g., annual rainfall in Seattle) through samples (e.g., taking precipitation level readings every day for a year). Uncertainty induced by randomness is called aleatory uncertainty. Epistemic uncertainty, on the other hand, refers to cases where a quantity is uncertain due to a lack of knowledge, but is in principle knowable. For example, given a particular measurement instrument, say, a thermometer intended for cooking, aleatory uncertainty describes the random fluctuation one would expect in temperatures even when the measured environment is consistent (e.g., a turkey that has been cooked to exactly 360 degrees). An example of epistemic uncertainty occurs if one asks instead about how the themometer would work in a different set of conditions to which it has not yet been applied, such as in measuring the temperature of water in the deep sea. Within the visualization literature, aleatory uncertainty is more commonly the focus of visual representations.

Uncertainty is generally modeled quantitatively, in which case it concerns an uncertainty quantity called a random variable (continuous or discrete). When multiple random variables are presented in a visualization, the potential for dependence between the variables requires representing the uncertainty as a joint probability density function (e.g., the product of the pdfs of the individual variables).

2.2 Existing evaluation paradigms

The typical goal in evaluating uncertainty visualization is to elicit evidence of a viewer's subjective probability distribution and then to compare this to a ground truth. In most controlled experiments, responses are elicited from subjects for two or more representations of uncertainty and performance compared. An exception is studies that compare performance with a visualization that depicts uncertainty information to one that does not [42, 18]. In some cases, alternative targets (preference ratings, spontaneous interpretations) are used in addition or instead of accuracy as dependent measures. I provide a brief overview of the various evaluation paradigms that occur in a set of surveyed studies evaluating uncertainty visualizations.

In keeping with the canonical approach to evaluating visualizations, **accuracy** is the most common outcome measure in evaluations of uncertainty visualization. In studies that target accuracy as a dependent variable, a common criterion for declaring an uncertainty visualization technique superior is to observe a big enough (e.g., statistically significant) difference in the degree to which the subjective uncertainty estimates of subjects who used one visualization treatment resemble the ground truth compared to that for the other visualization treatment. However, the specific forms used to elicit information about the subject's subjective distribution vary considerably in the form and precision of the information they elicit.

Absolute measures of accuracy elicit subjective probabilities or other measures that can be derived directly from subjective probabilities. Several recent studies [32, 36, 56] use direct elicitation of numeric probability estimates, following early work on novice interpretations of uncertainty visualizations by Ibrekk and Morgan [33]. Accuracy is determined by the absolute error between subjective probabilities and actual probabilities of values given the data distribution. Work by Hullman et al. [32] evaluates subjective estimates of joint probabilities in addition to estimates of probabilities for single random variables. Several studies employ absolute values derived from subjective distributions as outcome measures, including asking subjects to estimate the proportion of land use data that are reliable given a map with uncertainty depicted [18], and to judge the reliability of map features, which is then compared to the coefficient of variation for those features [42]. Particularly notable is the work of Tak et al. [56], who elicit subjective probability estimates for 9 different data points in order to analyze representations of subjects' subjective probability distributions posthoc. Though their work measures how accurately

subjects' estimates resemble the data distribution, the reconstruction process places a unique emphasis on the overall subjective distribution for each subject, including properties like its shape. Few studies directly elicit summary properties of a subjective distribution besides probabilities, such as quantiles, probability intervals, or measures of scale or dispersion.

Relative measures of accuracy include asking subjects to find regions of least certainty [44], or to rank targets by uncertainty [8, 6, 7]. Most responses are analyzed in comparison to the relative position of the features in the data. Other relative methods are scored not by comparing a subjective ranking to the data-based ranking, but by defining elicited levels of uncertainty using absolute values on the ground truth. For example, Sanyal et al. [50] use a relative framing, asking subjects to identify the "highest uncertainty" and "lowest uncertainty" features in 1D, 2D, and 3D visualizations. To score responses, they interpret highest and lowest uncertainty to be the features in the top and bottom 10th percentile of the data.

Also in keeping with canonical visualization evaluation metrics, **response time** is an outcome measure in several studies [18, 37, 40]. **Confidence ratings** or certainty levels in one's response are commonly elicited, either as a primary dependent variable or measure of differences between treatment in addition to accuracy or response time. Several studies use subjects' confidence in their responses to identify whether significant differences exist based on display [40, 8, 36]. An alternative technique is to examine the relative strength of association between the subject's confidence in their responses and the accuracy of their subjective probability estimates [36] or statistical properties of the data distributions such as the cumulative distribution function of a t-distribution for a single random variable or the p-value for a significance test for a pair of random variables [13].

Several studies adopt **spontaneous semiotic interpretations** as a dependent variable to evaluate what viewers assume different encodings to mean. Boukhelifa et al. [9] solicit spontaneous interpretations of sketchiness to understand how frequently viewers' attribute the sketchy encoding to uncertainty. MacEachren et al. [44] asks subjects to rate the suitability of various encodings for uncertainty.

Several studies ask the subject to make a domain-specific decision using the data [18, 37, 40] that is modeled after real world tasks, such as choosing an appropriate location given criteria [40].

In addition to accuracy, response time, and confidence, other subjective preferences for visualization techniques are commonly solicited as an overall measure of the effectiveness of a technique. Subjects have been asked to rate helpfulness [18, 1, 42], complexity of the task [1], degree of visual overload [46], ease of use [36, 38], visual appeal [36], and preference [38].

Most studies do not differentiate how well experts versus novices perform with different visualization techniques. Exceptions include the work of Aerts et al. [1], Evans [18], and Blenkinsop et al [8], who distinguish performance across users based on experience level. Also relatively rare are studies that focus specifically on non-expert viewers who may lack statistical education. Exceptions include Ibrekk and Morgan [33], Boukhelifia et al. [9], Correll and Gleicher [13], Tak et al. [56], and Hullman et al. [32].

3. THE PSYCHOLOGY OF UNCERTAINTY

I summarize three factors with direct influence on evaluation of uncertainty visualizations: the nature or definition of probability and subjective probability distributions, the sensitivity of responses to the elicitation method, and the potential for judgments made using heuristics to resemble normative responses.

3.1 Interpreting probabilities

3.1.1 Frequentist framing

To measure the accuracy of judgments that a viewer makes using an uncertainty visualization calls for an agreed upon ground truth. In many visualization evaluation settings, one can simply compare a user's judgments to the data that was shown. However, when the goal is to evaluate how accurate viewers' subjective probability distributions are based on one or more uncertainty visualizations, the choice of ground truth becomes significantly more complex. The *epistemological stance* that a research adopts strongly influences the validity of various interpretations of knowledge about uncertainty that is elicited from subjects.

When the target is aleatory uncertainty, frequency is a commonly assumed ground truth [61]. Viewers can be asked to provide subjective probabilities for events (e.g., $x \in X > 30$), which are compared to the frequency of the event in the data set (or perhaps the average frequency of the event across many generated replications). Often, however, uncertainty visualizations are intended to depict error associated with a sample statistic like the sample mean. For example, error bars commonly depict a 95% confidence interval, an interval such that if many such intervals were constructed, 95% of the intervals would contain the true mean. The appropriate ground truth is therefore the sampling distribution of the mean that is calculated using inferential statistics.

One might assume that viewers should in this case provide subjective probabilities for events related to the "true" mean (population parameter); after all, the sample mean for which the error bar is constructed is an estimate of the true mean. However, within a frequentist framework, the true mean is fixed, and probability–defined as frequency for an infinite repetition of experiments–can only be assigned to repeated events.

Many pervasive misinterpretations of 95% confidence intervals, however, indicate that this distinction is very frequently overlooked [30]. One of the most common misinterpretations of a 95% confidence interval is that the probability that the mean lies in the interval is 95%. This interpretation is incompatible with the frequentist definition of probability. Hence, to elicit subjective probabilities about inferential statistics, researchers must take care in framing questions that ask for probabilities, and focus instead on events that can be repeated (e.g., calculation of sample means).

The common misinterpretation of a confidence interval is frequently attributed to viewers erroneously interpreting the range from a Bayesian perspective [12, 16]. Bayesian statisticians argue that probability is best thought of as a quantity representing degrees of belief, or personal probability. As a result, in a Bayesian framework, subjective probabilities can be elicited for hypotheses or parameters themselves. In the latter case, ground truth can be established using Bayesian analysis.

3.1.2 Is subjective probability valid?

A deeper question still is whether it is reasonable to assume that probability is the most natural form for peoples' subjective understandings of uncertainty. Arguments that probability is uniquely valid as a measure of subjective uncertainty can be traced to elicitation methods that use a utility function framework. A rational agent's willingness to accept betting odds given a utility function that expresses the value of making a decision, conditional on the true values of the quantities in question, is used to derive probabilities [51, 15]. Later definitions similarly assume a scoring rule to define a person's subjective probability [14]. Hence it is assumed that the person whose subjective probabilities are being elicited is incentivized to carefully weigh the likelihood of different events, and to make the optimal decision according to the utility function. By offering a person enough different decision options, with varying rewards, it is thought that one can observe the "true" subjective probability distribution.

However, the assumption that probability is the right mental representation of subjective uncertainty is a topic of considerable scholarly debate. Arguments have been made against several premises on which the probability representation is based. For example, the normative theory relies on an assumption that personal probabilities obey the axioms of probability, including updating using Bayes' rule as new evidence is observed. Degrees of belief are further believed to be distributed according to the laws of probability, with degrees of belief of 0 representing that the person finds those events preposterous or impossible, while a reported probability of 50% means agnostic [47].

Some scholars have questioned how reasonable it is to assume that degrees of belief must obey the axioms of probability, that changes in degrees of belief must be conditioned on evidence, or that a person can balance probabilities and rewards appropriately [23]. In practice, some evidence indicates that people are capable of assigning probabilities to events in non-coherent ways: ways that do not obey the axioms of probability, nor the assumption that degrees of belief are rationally conditioned by evidence [47].

The possibility of incoherent probability estimates raises the question of how such behavior should impact analysis of results in an uncertainty visualization study. Imagine an evaluation paradigm in which a subject provides probability responses for mutually exclusive events, but their subjective probabilities do not sum to 1. Should an accuracy measure be adjusted to reflect this? Or should a researcher use the probabilities only as a form of relative information, or not at all?

3.1.3 Suggestions for practice

Challenges to evaluating uncertainty visualization that involve the validity of subjective probability are very difficult to "solve" through specific practices; they concern the very nature of subjective understandings of uncertainty. However, several high level guidelines can help ensure that a researcher's stance on probability is transparent and coherent with existing statistical paradigms.

- Avoid using frequency as a ground truth when asking for subjective probabilities about a parameter value. Instead, use alternative framings about the sample statistics that are compatible with frequency, such as questions about the probability of the sample mean falling in various ranges upon replicating the data collection and inference process. A researcher may also want to describe how hypothetical replications of the data collection and inference process might come about (e.g., a team is hired to run many replications of the original experiment) to ground subjects' interpretations.

- Consider individually incentivizing any questions in a study that depend directly on the subjective probability, such as by incorporating utility functions [39].This can help ensure that elicited probabilities are at least closer to the assumptions on which the probabilistic representation of subjective uncertainty is based [15].

- Include checks of "probability-coherence" in the evaluation procedure and the analysis of results. Do participants probabilities sum to 1 where expected? When basic expectations do not hold, consider whether the responses may express a deep lack of understanding about the task.

3.2 Elicitation error

The literature on eliciting subjective probability distributions in fields like decision science and behavioral economics suggests that the elicitation method (or response mode) can strongly affect the integrity and noisiness of responses. The response modes used in uncertainty visualization evaluations in the visualization literature (see Sec. 2.2) take one of several forms: 1) numeric response representing a probability or confidence level, 2) decision between alternatives, often accompanied by a confidence rating, 3) Likert scale rating in which subjects choose a level of probability, confidence, or associated value, or 4) listing or ranking of features with a specified probability level. The majority of authors provide little to no justification of their chosen response modes. The consequences of using a less reliable method is that the experimental comparisons between displays are subject to greater noise, which may even lead to conclusions of differences between displays that do not exist.

For example, Goldstein and Rothschild [25] find that eliciting an entire distribution from a respondent using a *graphical interface* (see also [24, 52]) and then computing simple statistics (such as means, fractiles, and confidence intervals) on this distribution leads to greater accuracy than the standard method of asking about the same statistics directly. These improvements affect both individual and aggregate estimates. Specifically, they show that a graphical method in which subjects construct a probability distribution leads to responses that better capture the ground truth distribution for a set of viewed stimuli than verbally asking subjects about subjective distribution properties like the median, extremes, and fractiles. In their graphical interface, bins are fixed and subjects assign probabilities to bins by dragging stacks of markers into the bins. This method also eliminates problematic characteristics of some subjective probabilities, such as the failure to sum to 1. For example, when 100 markers are used, each represents a 1% probability, and so the subjects' distributions must sum to 1.

Graphical elicitation methods that require putting a discrete number of elements into bins realize another recommended elicitation technique: *framing probabilities as natural frequencies* (counts). For example, rather than asking subjects "What is the probability that the difference in

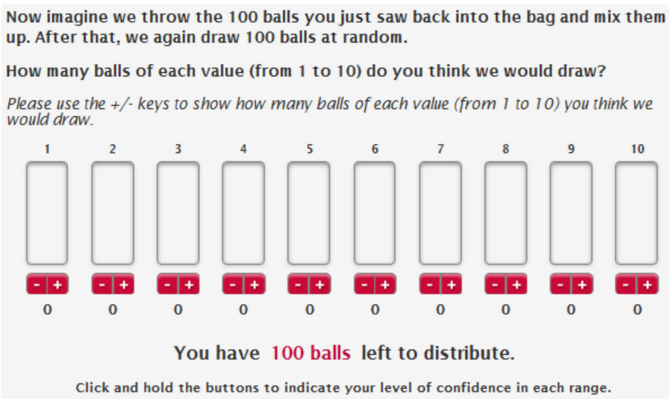

Figure 2: Graphical distribution builder for eliciting subjective probability distributions [25].

sample means between the groups will be greater than 30 points?", or alternatively, asking for a percentage, one should ask, "How many times, out of 100 repetitions of this study, would you expect to see a difference in sample means between the two groups that is greater than 30 points?") Frequency framings have been shown to reduce noise in subjective estimates about probability, such as in eliciting Bayesian reasoning [22, 31]). Only a small handful of papers in the uncertainty visualization literature surveyed above adopt a frequency framing (e.g., [32, 36]).

Regardless of whether one uses natural frequency, probability, or alternative (e.g., odds) formats, researchers must decide whether to elicit *absolute versus relative judgments*. Relative methods are typically less precise due to the large number of comparisons that must be conducted in order to rank a large number of events by probability. However, absolute methods are difficult to reconcile with the potential for context effects, which have been shown to affect many psychophysical judgments, such as when the same stimuli (e.g., a sound of a particular pitch, or circle of a particular size) produces different responses based on what stimuli appeared before it [3].

There is some evidence to suggest that probability judgments appear to be less relative than other psychophysical judgments [47]. In contrast to judgments of phenomena like size or loudness, probability scales have a clear bounded range of possible responses, with endpoints that are well understood (0% meaning no chance, 100% meaning certain). People have intuitions about other values on the scale (e.g,. 50% is equivalent to a coin flip) to further standardize their responses.

Likert style responses are used in a number of uncertainty visualization evaluation studies. Psychology research indicates that scales organized into categories often result in skewed responses based on the *range-frequency principle* [49]. Two tendencies that have been observed when subjects assign levels or categories to stimuli using response scales. First, people tend to use the different regions of the scale to cover broadly equal intervals or fractions between the smallest and largest stimulus presented. Second, people tend to use different regions or categories with broadly equal frequencies. Hence, a researcher should be aware that judgments made with Likert-style response scales may differ from judgments made with scales that are not framed categorically, and consider whether an alternative response mode is possible.

A tendency to balance responses across categories may also occur when elicitation methods ask subjects to identify features with given uncertainty or confidence levels (e.g., [8, 6, 7, 44, 50]). A second consideration when using these approaches concerns the level of clarity with which the categories are described. Terms such as "highly certain", when presented without an implied numeric range (e.g., having a probability of occurring that is 90% or more), are particularly likely to be sensitive to context effects and can lead to variability between people based on different interpretations (e.g., [62]).

One way to think about the sensitivity of responses to the elicitation interface is by acknowledging that subjective probability distributions are spontaneous. It is most likely that a subject is constructing a belief distribution upon being asked, rather than simply articulating a fully formed mental representation of their beliefs [47]. As a result, their ability to remember information relevant to the subjective probability estimation becomes important (e.g., how long it has been since they experienced evidence related to the event in question). Heuristics related to availability (see 3.3) can result.

Another fact that should be acknowledged in eliciting probabilities is that it is often inaccurate to draw general conclusions about elicitation from successes with a given method. A large body of research in judgment and decision making, as well as other fields, has indicated that very small changes to a method (e.g., "framing effects") can produce considerable changes in solicited responses. As an example, experiments do not provide a clear signal on whether methods that have viewers construct a representation of an entire distribution at once are better than alternatives. While Goldstein and Rothschild [25] found that having subjects graphically construct a probability distribution before answering questions reduced noise relative to asking questions directly, some other research that asks subjects to construct a probability representation has less success. *Interval estimation*, in which subjects are asked to construct an interval with a given probability of containing a statistic, is a well studied method that has repeatedly been shown to lead to overprecision, or overconfidence in one's interval (e.g., [45, 54]). Like various other sensitivities to elicitation methods, these effects do not disappear with more expertise [11]. Instead, researchers have found that interval-based methods that are *not* constructive, including asking subjects to evaluate many intervals instead [27] exhibit less biased responses. Hence despite broad similarities, elicitation methods should be considered independently.

3.2.1 Suggestions for practice

In addition to consulting research on the specific methods found to be successful above, other simple techniques may help visualization researchers elicit less noisy probability estimates:

- Use frequency formats rather than asking for probabilities to reduce noise in subjects' estimates [22, 31] . Be aware that the ease with which subjects can interpret the "out of" number (denominator) can influence results (see, e.g., [63]). In some situations, numbers out of 100 may be easiest as subjects can also apply any prior knowledge they might possess that has been

framed in terms of probabilities. However, if the event in question is one for which it is difficult to imagine many repeated simulations, a smaller number (e.g., out of 10) may be easier to think about.

- To help subjects overcome resistance to providing subjective probabilities and reduce noise in the interpretation of probabilities, consider "anchoring" a scale with familiar probabilities, such as coin flips.

- Use pilot testing to hone an elicitation interface prior to an evaluation study. Suggested measures for determining the effectiveness of a scale include checking how often 90% confidence intervals derived from elicited distributions bracket a correct answer across a number of problems [27].

- Alternatively, a simpler approach to piloting that is likely to still provide valuable feedback is have subjects provide distributions using various formats, and then judge which way they believe to be most useful in allowing them to articulate their opinions.

3.3 Heuristics mimic understanding

Resistance to thinking about how uncertainty affects judgments has led some experts to propose that people should not be shown precise numerical uncertainty estimates in many practical situations (e.g., [2] as cited in [34]). However, with just a few exceptions, evaluation studies in uncertainty visualization assume that uncertainty will be shown, and test multiple different displays. One way in which uncertainty visualization evaluations could acknowledge the inherent cognitive load of uncertainty (and address Harrower's vision of understanding how uncertainty changes thinking [28]) is to include "no-uncertainty" controls for comparison.

When asked, most people can produce a probability for a wide range of questions involving uncertainty. But often they do so using heuristics, a form of intuition that provides a mental shortcut for hard decisions. Most heuristics work by substituting a simple but less accurate representation to turn a difficult decision–about a situation with multiple parameters and uncertainties–into an easier one. This phenomena is sometimes called "attribute substitution" [35].

For example, faced with a bar chart with error bars depicting confidence intervals (e.g., any of the charts in Fig. 3) and a question like "How likely is it that the first bar will be greater than the second if many more samples are gathered?", subjects may avoid the complicated task of judging the probability using the error bars and instead rely on a simpler cue–say, the difference between the means–to estimate how reliable a perceived pattern is [32]. Big difference → reliable difference, small difference → unreliable difference. This works when the data resembles those situations on which the heuristic is based, but fails badly in cases that depart from the typical case, such as small but reliable differences.

This example demonstrates belief in the law of small numbers [58], a form of representativeness bias [60]. In making decisions, people underweight sample size, focusing instead on the degree to which a stimuli (or sample, or instance) resembles a common pattern in large samples (or population, or category). Also applicable to uncertainty visualization evaluation is an inverse tendency: *non-belief* in the law of

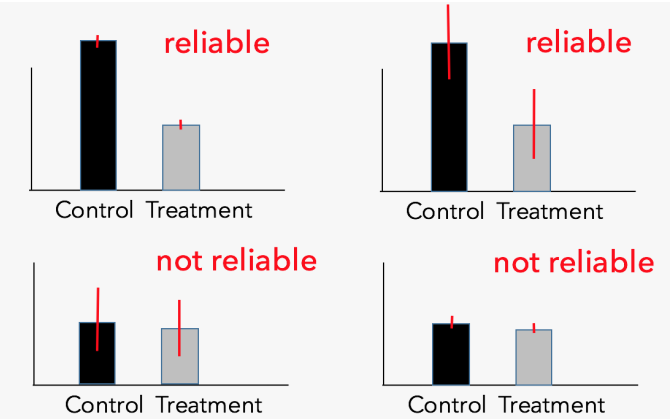

Figure 3: Subjects may follow heuristics based on the difference in means, rather than incorporating the error information, when judging differences in bar charts with error bars [32].

large numbers [5]. The inverse tendency suggests that even when given ample evidence that an event is highly certain, people are reluctant to judge it as certain. Hence, the extremes of a probability (or certainty, or confidence) scale may be used less than suggested by a normative response, producing asymmetries in judgment errors that differ are difficult to explain in analysis.

Another heuristic with the potential to bias the results of evaluations of uncertainty visualizations is availability. This heuristic refers to the tendency for people to overweight (e.g., see as more probable) events that come to mind more easily [59]. Subjective probabilities have been shown to differ in particular in how much likelihood is attributed to low frequency events [47]. In particular, when a person *experiences* an uncertain event, subjective probabilities assigned to low frequency events tend to be *higher* than when the person develops the subjective probability assessment from a *description* of the probability of the event [29]. It is an open question whether these effects also apply to uncertainty visualization technique that differ in how they represent uncertainty for the user, for example animated outcome-based techniques (e.g., [4, 18, 17, 32]) versus more descriptive techniques (e.g., error bars or other static representations of an entire distribution at once).

A form of context effect that can affect both relative and absolute methods is *value biasing*. Value biasing occurs when the value or utility of an event influences the probability assigned to it. If an event for which a subjective probability is elicited is perceived to have very positive or negative effects (e.g., the probability of an earthquake in one's region), these value judgments may influence the subjects' strength of beliefs [47]).

Heuristics are a particular challenge for uncertainty visualization evaluation because there are many cases in which their usage cannot easily be differentiated from more informed judgments. Consider, for example, a study of the effectiveness of error bars in which counter examples of the use of the difference in means to judge reliability are not tested. Subjects using this heuristic may not be distinguishable from those that are not. Additionally, heuristics can lead to high confidence in the accuracy of one's answers for

some stimuli. Given that heuristics are mental shortcuts that reduce decision complexity, it makes sense that subjects might feel more certain of their accuracy: answering a question using a heuristic is a more fluent experience.

3.3.1 Suggestions for practice

While it may not be possible to eliminate heuristics from subjects responses in evaluating uncertainty visualizations, several practices can help a research detect and evaluate the impact of heuristics:

- Be proactive about brainstorming possible heuristics that may affect judgment when designing evaluations. Include stimuli that are expected to produce answers that deviate more or less from ground truth. Check expectations about the heuristic against the elicited responses.
- Look for signs of heuristics in responses by looking for unusually large variance [64]. This can signal that people are using nonpredictive cues beyond the presented data and uncertainty.
- Ask subjects to describe the strategies they are using as they answer questions about uncertainty with a visualization.
- Consider including the degree of heuristic use as a dependent variable in an evaluation study. For example, if a heuristic is found to affect responses, compare the number of responses that are consistent with the heuristic to the number that are not across the visualization treatments to see if some visualizations reduce usage.
- Be aware that personal assessments of the value of events may impact judgments. Consider asking directly for value judgments if it is anticipated that subjects may have strong opinions on the value of different hypothetical events that are asked about.

4. CONCLUSION

Evaluating a visualization that depicts uncertainty is fraught with complexities, yet little attention is given to the interpretation or elicitation of subjective understandings of uncertainty in the uncertainty visualization literature. Awareness of the psychology of uncertainty, as documented in research focused on elicitation of subjective probabilities in fields like judgment and decision making, can help researchers avoid unnecessary noise or bias in responses. Practices that acknowledge and aim to reduce "uncertainty" in the definition of probability and subjective probability, the choice of elicitation method, and the strategies used by subjects to make judgments with an uncertainty visualization are proposed. Future work should further explore how various aspects of the psychology of uncertainty and probability elicitation impact uncertainty visualization.

5. 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] J. S. Ancker, Y. Senathirajah, R. Kukafka, and J. B. Starren. Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, 13(6):608–618, 2006.

[3] N. H. Anderson. On the role of context effects in psychophysical judgment. *Psychological Review*, 82(6):462, 1975.

[4] L. Bastin, P. F. Fisher, and J. Wood. Visualizing uncertainty in multi-spectral remotely sensed imagery. *Computers & Geosciences*, 28(3):337–350, 2002.

[5] D. J. Benjamin, M. Rabin, and C. Raymond. A model of nonbelief in the law of large numbers. *Journal of the European Economic Association*, 14, 2015.

[6] A. M. Bisantz, S. S. Marsiglio, and J. Munch. Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 47(4):777–796, 2005.

[7] A. M. Bisantz, R. T. Stone, J. Pfautz, A. Fouse, M. Farry, E. Roth, A. L. Nagy, and G. Thomas. Visual representations of meta-information. *Journal of Cognitive Engineering and Decision Making*, 3(1):67–91, 2009.

[8] S. Blenkinsop, P. Fisher, L. Bastin, and J. Wood. Evaluating the perception of uncertainty in alternative visualization strategies. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 37(1):1–14, 2000.

[9] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2769–2778, 2012.

[10] K. Brodlie, R. A. Osorio, and A. Lopes. A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. Springer, 2012.

[11] B. Clemen. Assessing 10-50-90s: A surprise. *Decision Analysis Newsletter*, 20(1):2, 2001.

[12] J. Cohen. The earth is round (p<. 05): Rejoinder. 1995.

[13] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2142–2151, 2014.

[14] B. de Finetti. Probabilities of probabilities: A real problem or a misunderstanding. *New Developments in the Applications of Bayesian methods*, pages 1–10, 1977.

[15] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The statistician*, pages 12–22, 1983.

[16] Z. Dienes. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3):274–290, 2011.

[17] C. R. Ehlschlaeger, A. M. Shortridge, and M. F. Goodchild. Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23(4):387–395, 1997.

[18] B. J. Evans. Dynamic display of spatial

data-reliability: Does it benefit the map user? *Computers & Geosciences*, 23(4):409–422, 1997.

[19] R. Finger and A. M. Bisantz. Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science*, 3(1):1–25, 2002.

[20] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[21] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[22] G. Gigerenzer and U. Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

[23] C. Glymour. *Why I am not a Bayesian. In Theory and Evidence*. University of Chicago Press, 1981.

[24] D. G. Goldstein, E. J. Johnson, and W. F. Sharpe. Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, 35(3):440–456, 2008.

[25] D. G. Goldstein and D. Rothschild. Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1):1, 2014.

[26] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.

[27] U. Haran, D. A. Moore, and C. K. Morewedge. A simple remedy for overprecision in judgment. *Judgment and Decision Making*, 5(7):467, 2010.

[28] M. Harrower. Representing uncertainty: Does it help people make better decisions. *Ithaca, NY: University Consortium for Geographic Information Science. Accessed October*, 16:2012, 2003.

[29] R. Hertwig, G. Barron, E. U. Weber, and I. Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8):534–539, 2004.

[30] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, 2014.

[31] U. Hoffrage and G. Gigerenzer. Using natural frequencies to improve diagnostic inferences. *Academic medicine*, 73(5):538–40, 1998.

[32] 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.

[33] H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis*, 7(4):519–529, 1987.

[34] S. Joslyn and J. LeClerc. Decisions with uncertainty: the glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013.

[35] D. Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697, 2003.

[36] 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*, pages 5092–5103. ACM, 2016.

[37] D. A. Kobus, S. Proctor, and S. Holste. Effects of experience and uncertainty during dynamic decision making. *International Journal of Industrial Ergonomics*, 28(5):275–290, 2001.

[38] R. Kosara. *Semantic Depth of Field-Using Blur for Focus+ Context Visualization*. PhD thesis, Kosara, 2004.

[39] S. M. LaValle. *Planning algorithms*. Cambridge University Press, 2006.

[40] M. Leitner and B. P. Buttenfield. Guidelines for the display of attribute certainty. *Cartography and Geographic Information Science*, 27(1):3–14, 2000.

[41] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, (13):10–19, 1992.

[42] A. M. MacEachren, C. A. Brewer, and L. W. Pickle. Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 30(9):1547–1561, 1998.

[43] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

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

[45] D. A. Moore and P. J. Healy. The trouble with overconfidence. *Psychological review*, 115(2):502, 2008.

[46] 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.

[47] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[48] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[49] A. Parducci. Contextual effects: A range-frequency analysis. *Handbook of perception*, 2:127–141, 1974.

[50] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. J. Moorhead. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1209–1218, 2009.

[51] L. J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

[52] W. F. Sharpe, D. G. Goldstein, and P. W. Blythe. The distribution builder: A tool for inferring investor preferences. 2000.

[53] M. Skeels, B. Lee, G. Smith, and G. G. Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2010.

[54] J. B. Soll and J. Klayman. Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):299, 2004.

[55] S. M. Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[56] S. Tak, A. Toet, and J. van Erp. The perception of visual uncertainty representation by non-experts. *Visualization and Computer Graphics, IEEE Transactions on*, 20(6):935–943, 2014.

[57] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. In *Electronic Imaging 2005*, pages 146–157. International Society for Optics and Photonics, 2005.

[58] A. Tversky and D. Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

[59] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[60] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, pages 141–162. Springer, 1975.

[61] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. *Bull. Amer. Math. Soc*, 51(7):498–504, 1945.

[62] P. D. Windschitl and G. L. Wells. Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, 2(4):343, 1996.

[63] K. Yamagishi. Upward versus downward anchoring in frequency judgments of social facts. *Japanese Psychological Research*, 39(2):124–129, 1997.

[64] J. F. Yates, L. S. McDaniel, and E. S. Brown. Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes*, 49(1):60–79, 1991.