In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation

Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, Matthew Kay

Abstract—Understanding and accounting for uncertainty is critical to effectively reasoning about visualized data. However, evaluating the impact of an uncertainty visualization is complex due to the difficulties that people have interpreting uncertainty and the challenge of defining correct behavior with uncertainty information. Currently, evaluators of uncertainty visualization must rely on general purpose visualization evaluation frameworks which can be ill-equipped to provide guidance with the unique difficulties of assessing judgments under uncertainty. To help evaluators navigate these complexities, we present a taxonomy for characterizing decisions made in designing an evaluation of an uncertainty visualization. Our taxonomy differentiates six levels of decisions that comprise an uncertainty visualization evaluation: the behavioral targets of the study, expected effects from an uncertainty visualization, evaluation goals, measures, elicitation techniques, and analysis paradigms. Applying our taxonomy to 86 user studies of uncertainty visualizations, we find that existing evaluation practice, particularly in visualization research, focuses on Performance and Satisfaction-based measures that assume more predictable and statistically-driven judgment behavior than is suggested by research on human judgment and decision making. We reflect on common themes in evaluation practice concerning the interpretation and semantics of uncertainty, the use of confidence reporting, and a bias toward evaluating performance as accuracy rather than decision quality. We conclude with a concrete set of recommendations for evaluators designed to reduce the mismatch between the conceptualization of uncertainty in visualization versus other fields.

Index Terms—Uncertainty visualization, user study, subjective confidence, probability distribution.

1 Introduction

Data-driven presentations have become commonplace in public-facing domains like the news as well as in the scientific literature. A newspaper article might depict differences in the probabilities of a set of political candidates winning an election. A government agency might present future temperature predictions of a model analyzing climate change trends. By conveying the possibility that a point estimate may vary, uncertainty visualizations enable people to make more informed decisions. As public trust in science declines [19] and overconfidence in noisy effects reportedly affects a number of empirical disciplines [42], uncertainty visualizations are more important than ever.

It is the task of research in uncertainty visualization to provide evidence of the impacts of proposed uncertainty visualization techniques, so as to inform practice. However, how to design an effective evaluation of an uncertainty visualization is rarely addressed in research focused on creating uncertainty representations. A researcher or practitioner seeking to evaluate a visualization with users must instead rely on general purpose frameworks designed to ensure that evaluation designs are appropriate given the nature of the visualization contribution (algorithmic, interaction technique, encoding, etc.) [67, 57].

Recently, scholars have pointed to the challenges evaluating uncertainty visualizations compared to evaluating other visualizations [38, 55, 79]. For example, researchers in judgment and decision making describe eliciting and analyzing subjective accounts of uncertainty as a process fraught with its own uncertainty [72], though the difficulty of uncertainty elicitation is seldom considered in studies of uncertainty visualization. Statisticians and other scholars have long debated how to define normative accounts of uncertainty [2, 21, 27, 83]; without clear agreement on what uncertainty is, it is difficult to imagine an agreed upon approach to evaluating uncertainty comprehension. Canonical work on judgment under uncertainty argues that people display cognitive biases when making decisions involving uncertainty [51, 50]. Historically, such biases have been attributed to heuristics, the use of simple judgments as a proxy for difficult judgments [49, 90, 91, 92]. A common misconception is that reliance on heuristics is bound to cause biased judgments. On the contrary, heuristics are adaptive and often lead to accurate judgments. In the context of evaluating uncertainty visualizations, this means that even when uncertainty is presented in a non-optimal format–such as error bars, which lead to perceptual bias [69, 11] and underweighting of uncertainty [40]–responses based on heuristics will sometimes be correct. Consequently, it can be difficult to create the conditions in which heuristics are flawed and cognitive biases occur. In the face of these and other unique challenges that uncertainty information presents to evaluation, the integrity of our knowledge on how to best visualize uncertainty is at stake.

We take a closer look at the evaluation of uncertainty visualizations through the largest systematic review of existing uncertainty visualization practice. We present a survey of the techniques used in 86 uncertainty visualization evaluations published between 1987 to 2018 in a variety of disciplines. Our first contribution is a taxonomy for distinguishing between uncertainty visualization evaluation approaches (Fig. 2). Our taxonomy distinguishes key considerations at six levels of decisions comprising an uncertainty evaluation approach. *Behavioral Targets* describe what aspects of the impact of uncertainty visualization(s) on user behavior are examined, such as its impact on Performance or the Quality of User Experience. *Expected Effects* describe how a study defines when an uncertainty visualization is successful or not. *Evaluation Goals* pertain to the study design: was the goal to Compare multiple visualizations, Determine the impact of presenting uncertainty versus not, or another goal? *Measures* include directly elicited measures such as satisfaction ratings, probability estimates, subjective confidence, etc. as well as derived measures representing transformations such as error, or the extent to which a decision maximizes utility. *Elicitation* describes the various ways that user's responses can be gathered, from interaction with a physical apparatus to a multiple choice question. *Analysis* describes how an evaluator determines if the *Expected Effects* were achieved, such as Frequentist Null Hypothesis Significance Testing (NHST) or Bayesian Estimation, for example.

Our second contribution is an analysis of *evaluation paths*, where we compare the co-occurance of codes for 86 evaluation studies across the different levels of our taxonomy. For example a common evaluation path describes a Behavioral Target (*L1*) of *Performance* with

- *Jessica Hullman is with Northwestern University. E-mail: jhullman@eecs.northwestern.edu.*
- *Xiaoli Qiao is with University of Washington. E-mail: xiaoliq@uw.edu.*
- *Michael Correll is with Tableau Software. E-mail: mcorrell@tableau.com.*
- *Alex Kale is with University of Washington. E-mail: kalea@uw.edu.*
- *Matthew Kay is with University of Michigan. E-mail: mjskay@umich.edu.*

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

an Expected Effect (*L2*) of *Increasing Accuracy*, conducted by (*L3*) *Comparing the Impacts of Multiple Visualizations* through a Measure (*L4*) of participant's *Probability* estimates from which *Error* is derived, which were Elicited (*L5*) via a *Slider* and Analyzed (*L6*) using a *Frequentist NHST* approach (Fig. 1). We find that existing evaluation practice, particularly in Visualization research, focuses on comparing multiple uncertainty visualizations to one another using a small number of Performance and Satisfaction-based measures elicited using constrained inputs such as multiple choice or Likert-style responses. These studies implicitly assume that correct responses are arrived at using procedures that resemble statistical decision making, that people's meta-cognitive assessments of how effective an uncertainty visualization is are reliable, and that elicitation techniques are mostly interchangable without affecting responses.

Our final contribution is a set of recommendations designed to encourage more transparent evaluations aligned with the state of the art in knowledge on uncertainty comprehension. Our results confirm and extend recent recommendations made by others [38, 55, 54]. We describe how evaluators tend to interpret subjective uncertainty in several incompatible ways, as either directly comparable to statistical uncertainty or as an inherently subjective meta-cognitive assessment of confidence. We describe what lessons can be learned from formal analyses of decision making and subjective uncertainty using techniques from experimental economics, psychophysics, and judgment and decision making. We highlight the potential for decision-based approaches to enable more realistic evaluations relative to accuracy or efficiency, the value of evaluations that aim to understand *why* visualizations produce different behavior, and the value of including no uncertainty and textual controls as a means of developing a deeper base of knowledge around the impacts of uncertainty visualization, among others.

2 Related Work

2.1 Improving Visualization Evaluation

Our work extends a larger body of work aimed at improving evaluation methods in visualization. Researchers have contributed overviews of qualitative and quantitative approaches [57, 94] and models for ensuring that one selects an evaluation that is appropriate for a given task, context, or contribution type [43, 67, 84]. For example, Munzner's well known Nested Model [67] stresses the dependencies between stages of the design process (e.g., characterizing the task and data, designing visual encodings), where an error at one stage can affect subsequent stages. Our taxonomy similarly distinguishes between different "stages" of decisions in the design of evaluations for uncertainty visualizations.

Several prior works describe trends in evaluation in the visualization community through systematic review of studies (e.g., [44, 57]). Lam et al. use a review of 361 visualization papers to contribute a set of scenarios describing the most common goals and outputs of different forms of evaluation. Their work takes a broad view that evaluation can occur at any stage in a design process. Isenberg et al. [44] borrow from Lam et al.'s [57] codes in a review of 581 papers that evaluate the impacts of visualizations on users.

We restrict our review to studies that correspond to Lam et al.'s scenarios for understanding visualization: formal studies conducted to determine a visualization's effects. Similar to both prior works, we include a distinction between evaluations that target Performance versus User Experience. Additionally, in our taxonomy we include a third type of behavioral target of Semantics and Interpretation of uncertainty depictions which we observed in multiple studies. Because the goal of our taxonomy is to surface intended goals and evaluation designs for uncertainty information, we employ a more detailed coding scheme than these prior surveys. For example, in contrast to the reviews of Lam et al. and Isenberg et al., which used a total of 17 and 8 codes respectively, we employ a coding scheme comprised of 76 total codes distributed across six different levels of evaluation design decisions (3 to 25 codes per level).

2.2 Evaluating Uncertainty Visualization

Various standards boards and public-facing organizations have described the importance of conveying uncertainty as a component of scientific data [24, 71, 88]. Visualization researchers and cartographers have proposed a large number of techniques for visualizing uncertainty, which can be classified using various taxonomies [30, 45, 62, 76, 78, 86, 89]. We have observed that few studies used to validate new techniques *motivate* the evaluation design against alternatives, in addition to providing sufficient information for understanding a presented study (i.e., detailing the questions posed to participants, the participant sample, and how the measures were calculated). This aligns with Kinkeldey et al.'s [55] observation that few of 44 cartographic uncertainty visualization evaluations justified their tasks.

Some researchers have commented on the unique aspects of evaluation of uncertainty visualization. Harrower [33] argues that evaluators should be more concerned with "does it help" and less focused on "which vis is better," given the inherent complexity of uncertainty information. Boukhelifa and Duke [8] mention assessing how uncertainty information is used as one challenge in visualizing uncertainty, while Bonneau et al. [6] distinguish three types of evaluations: theoretical evaluation using design principles (e.g., [94]), low-level visual evaluation (e.g., [52]), and task oriented user study (e.g., [82]).

Several recent papers comment on the challenges of designing a realistic uncertainty visualization evaluation while maintaining sufficient experimental control to test predictions [55, 79]. Kinkeldey et al. summarize studies reported in 34 publications that describe an evaluation of how geospatial uncertainty visualizations communicate [55]. In a follow up review on an overlapping set of 43 studies, the authors focus instead on assessing the impact of uncertainty visualization on decision-making and risk assessment [54]. These works characterize types of studies (e.g., laboratory, etc.), types of assessment (objective, involving correct answers, versus subjective, exploring user intuitions), types of uncertainty, visualization techniques examined, application domains, and participant and task characteristics. They note that in many cases evaluations focus on simplified, low-level visual judgment tasks, and appear to be designed in an *ad hoc* manner. They also find a tendency to measure decision-making rather than perceptions of risk, which we also observe in our larger sample. They recommend greater systematicity and focus on realistic user tasks.

Focusing on evaluation across visualization subfields, Hullman [38] summarizes patterns in a small sample of uncertainty visualization studies, characterizing study goals and types of measures. They review challenges to the epistemological nature of uncertainty, and provide specific suggestions related to eliciting subjective probability such as using frequency framings, familiar probability "anchors" (e.g., coin flips), and comparing multiple forms of elicitation.

Our taxonomy extends these prior classifications with a more comprehensive set of distinctions about decisions at each stage of visualization evaluation, informed by categories that emerged in our analysis as well as the state-of-the-art in judgment and decision making.

3 Methods: Creating the Taxonomy

3.1 Scope: Evaluative User Studies

To be included in our sample, we required that a study included at least one *visual representation of uncertainty*, and at least one *research question concerning the impact of an uncertainty visualization* on a user's performance, impressions, or behavior. The goal of these criteria was to eliminate, for example, studies focused on different forms of textual representations of uncertainty or different framings of uncertainty (e.g., frequency) that did not involve visualizations per se. Also eliminated were studies that presented uncertainty to viewers but without posing any research questions concerned with how visual representation of uncertainty impacted behavior or responses.

As user studies, each study necessarily included at least one means of *eliciting responses or actions from users of a visualization*. Our final criterion for inclusion was that each study included a *quantitative analysis of users' responses or behaviors* as impacted by a visualization. This criterion is meant to exclude studies that did not judge how

Table 1. Categories of publication venues and application domains in our sample of 86 uncertainty visualization evaluations.

Publication Venue	#	Application Domain	#
Automotive Ergonomics	2	Aviation/Defense	1
Cartography/GIS	19	Astrophysics	1
Cognitive Psychology	9	Cartography/GIS	15
Computer-Aided Medicine	2	Domain-general	18
SciVis	9	Graphs/Trees	2
Health informatics	4	Health/Medicine	16
Human Factors	6	Meteorology	14
Decision Making	4	Manufacturing	2
InfoVis	23	Management	1
Meteorology	1	Transit	6
Security	1	Volumetric Data	5
Ubicomp	6		

prevalent or important different results are (i.e., studies that refrain completely from reporting on the frequency of different results, even textually through terms such as "most users"). Because we are interested in gaining a comprehensive view of available techniques for uncertainty visualization evaluation, we put no restrictions on the disciplinary venue where a study was published.

3.2 Sample

We seeded our list of publications containing evaluation studies with Potter's online library of uncertainty visualization studies [77]. This resource contains 241 publications presenting uncertainty visualization techniques or studies published between 1990 and 2013 in core venues associated with the research fields of InfoVis, SciVis, Cartography, Medicine, and Psychology, among others. One author first examined each of the 241 papers to remove those that did not contain evaluative user studies as defined above (leaving 48 studies). To identify studies from 2013 to the present, we used a set of queries containing methodological terms like "user study" or "controlled experiment" as well as indicators of uncertainty visualization like "uncertainty visualization", "uncertainty representation", "risk visualization", "graphical representation of risk", and "risk representation". We queried Google Scholar for each combination of method term and uncertainty visualization term. We manually went through each list of results, stopping when we reached an entire page (10 papers per page) of papers that were no longer relevant to uncertainty visualization. This resulted in an additional 52 studies (100 total). During the coding process, we eliminated 14 of these 100 studies as a result of not being able to obtain the paper or noticing a priori that a study did not fulfill one of our inclusion criteria (e.g., we removed an fMRI study where uncertainty visualizations were shown only to gauge how the brain reacts to uncertainty).

Table 1 summarizes the number of studies by publication discipline and application domain.

3.2.1 Coding Procedure

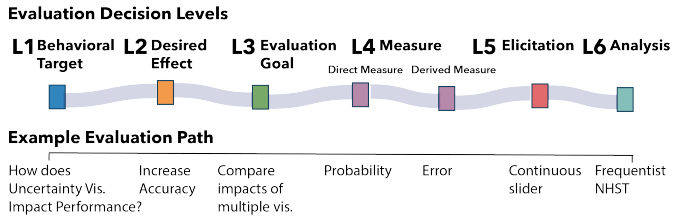

Fig. 1. A depiction of our evaluation taxonomy (top), which distinguishes evaluation decisions made at six different levels; and an example evaluation path (bottom) from the set of 372 paths that we coded across a sample of 86 publications.

We created an initial "top-down" set of distinctions based on our own knowledge of procedures in visualization and related fields such as cognitive psychology. This basic taxonomy differentiated high-level goals (e.g., "improving decision making"), intermediate goals (e.g., "improving accuracy in data extraction"), and evaluation designs (e.g., "comparing multiple uncertainty visualization treatments"). Implementation decisions were considered separately and were categorized based on what measure was elicited (e.g., a probability estimate), how it was elicited (e.g., a slider), and analyses approaches (e.g., Frequentist NHST). The taxonomy was prescriptive in that we included nodes for techniques that are recommended in literature on expert probability elicitation [72], such as graphical elicitation [28] and formal analyses of decision making [14].

We next iteratively applied a "bottom-up" approach in which we assessed the approaches used in our sample. We isolated each measure in a study, tracing each distinct analytical comparison that was applied to this measure. We labeled the research question, expected effect, evaluation design, elicitation method, and analysis approach for each comparison (creating an "evaluation path"; e.g., Figure 1).

As we analyzed more studies, we periodically added new codes to our taxonomy and refined the distinctions between levels of the taxonomy. For example, we opted to reframe high-level study goals more specifically as Behavioral Targets, and intermediate goals as Expected Effects describing the direction of a hypothesized effect in order to more cleanly distinguish these two levels. We also added sub-levels to differentiate elicited measures (e.g., subjective confidence, probability, etc.) from derived measures (e.g., alignment between confidence and accuracy, error, etc.) After all substantive changes, we recoded any affected paths. Our taxonomy includes 76 total codes, with an average of 12.7 codes per each of the six levels.

The first and second author evaluated all studies, resolving inconsistencies between coders and discussing ambiguous codes. Our final coded sample includes 372 paths (mean per publication: 4.3 paths).

4 Uncertainty Visualization Evaluation Taxonomy

4.1 Overview

Our taxonomy proposes six levels of decisions that characterize an uncertainty visualization evaluation. We distinguish between aspects of an evaluation that describe its *Research Values and Aims* (L1, L2), and those that describe its *Research Design* (L3 - L6).

Applying our taxonomy to the 86 studies in our sample produced 372 instances of evaluation paths. We present each code, along with its definition, frequency, and examples below, and depict all coded paths in Figure 2. Our supplement, available at `https://github.com/jhullman/uncertaintyVisEval` (10.5281/zenodo.1324465), provides interactive visualizations that support finding paths with associated study information.

4.2 L1, L2: Research Values and Aims

Research values and aims are comprised of *Behavioral Targets (L1)* describing the focal dimension of the uncertainty visualization(s)' impact and *Expected Effects (L2)* representing the directions of expected outcomes that are framed as either more or less desirable (e.g., more risk aversion, more confidence as ideal behavior).

4.2.1 L1: Behavioral Targets

We assigned each evaluation path one of three broad categories of visualization effects that a study can attempt to isolate:

- **Performance** (241; 64.8%): how effectively a user can extract information, make inferences, or make decisions with a visualization (e.g., [40, 82, 85]).
- **Interpretation & Semantics** (64; 17.2%): the ease with which a user associates uncertainty with an encoding (e.g., [7, 63]). (eg.,
- **Quality of User Experience** (67; 1.8%): the user's valuation of the visualization, such as their preference or satisfaction (e.g., [29]).

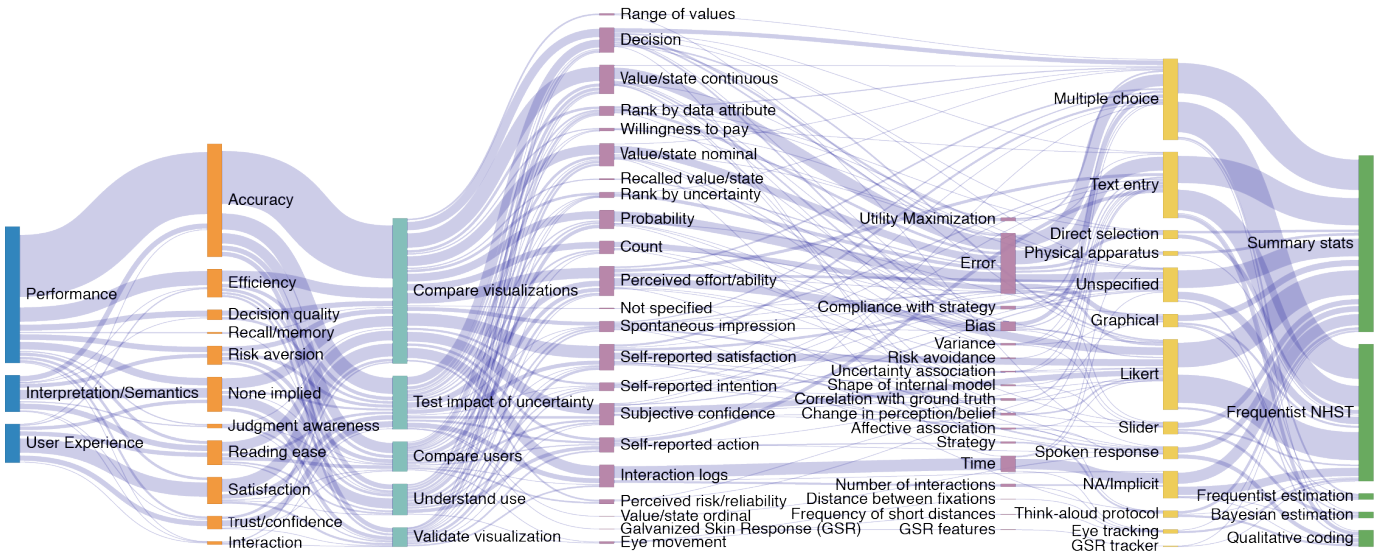

Fig. 2. 372 evaluation paths that we observed across a sample of 86 publications with uncertainty visualization evaluations. The number of inlinks and outlinks differ for some nodes due to the same evaluation path representing multiple codes at a single level (e.g., Analysis).

The first two categories map closely to those applied in prior surveys of visualization evaluation (i.e., Kinkledey et al's [54] "effect" and "communication", which can be traced to subsets of Isenberg et al.'s [44]'s codes for assessing visualization evaluation. We decided to create a separate category for "Quality of User Experience" as Desired Effects associated with these Behavioral Targets tended to differ from Desired Effects associated with Behavioral Targets of Performance and Interpretation & Semantics in that the correctness and/or format of the user's internal representation was not of interest.

Our results make clear that most evaluation paths (241; 64.8%) in our sample framed the Behavioral Target (L1) as an assessment of Performance. Performance targets imply that visualizations are tools for supporting analytical tasks.

4.2.2 L2: Expected Effects

In addition to targeting a specific dimension of behavior, evaluation studies typically presuppose the *direction* of an effect of interest. We defined Expected Effects including:

- **Accuracy** (134; 36%): Difference from a ground truth response (e.g., [37, 40]).
- **Confidence/Accuracy Alignment** (5; 1.3%): Correlation between confidence and accuracy (e.g., [11]).
- **Impact on Decision-Making** (3; 0.8%): Effects on decisions where no ground truth is implied (e.g., [15]).
- **Decision Quality** (14; 3.8%): Difference from a rational decision standard (e.g., [23, 68]).
- **Memorability** (3; 0.8%): How well a fact is remembered (e.g., [39]).
- **Risk Avoidance** (23; 6.2%): The degree to which a user's response attempts to avoid risk (e.g., [35]).
- **Intuitiveness/Reading ease** (31; 8.3%): How "naturally" a visualization supports correct interpretations (e.g., [7]).
- **Awareness of Judgment Process** (8; 2.1%): The user's ability to recognize aspects of their judgment (e.g., [15] 4.4).
- **Learnability** (0): The user's ability to improve their performance over time
- **Confidence** (24; 6.5%): The degree of belief in the validity or truth of a judgment, data set, visualization, etc. (e.g., [16]).
- **Satisfaction** (38; 10.2%): The user's aesthetic valuation of a visualization (e.g., [22, 61]).
- **Interaction** (4; 1.1%): How much interaction a visualization receives in time, number of clicks, etc. (e.g., [32, 35].
- **Efficiency** (37; 9.9%): How well a visualization presents the relevant information in a way that supports quick judgment (e.g., [59]).

We coded evaluation paths that did not imply any expected direction for an effect on user responses as intending to:

- **Understand internal model** (48; 12.9%): How a judgment is made (e.g. [13, 87]).

Our goal was to capture the assumptions that drive evaluators toward focusing on a certain type of response as a signal of whether an uncertainty visualization works. As such, decisions at L4 Measures tend to be dependent on L2.

324 (87%) of the evaluation paths we coded implied an Expected Effect (L2), either explicitly through a hypothesis or prediction, or implicitly in how the goal of uncertainty visualization was framed in contextualizing the task. Over one third of the evaluation paths we coded described tasks where ground truth could be used to establish what an accurate versus inaccurate response looked like. Among the 48 (13%) paths that did not imply an Expected Effect, those associated with Performance targets tended to represent attempts to understand users' mental models or strategies as a means of understanding why judgments were better or worse. When associated with an Interpretation & Semantics target, these paths similarly tended to represent attempts to understand mental models or strategies, but where this understanding was framed as a primary goal, rather than being in service of supporting Performance or a better User Experience (e.g., [11, 87]). We did not observe any cases where a Behavioral Target of the Quality of User Experience was assessed with a subgoal of trying to understand the user's internal model.

We coded paths that did not imply a desired direction for an effect as aiming to "Understand the user's internal model".

4.3 L3 - L6: Research Design

The research design of a study is comprised of *Evaluation Goals (L3)* represented by experimental design considerations, *Measures (L4)* represented by elicited responses, *Elicitation (L5)* describing how the measures are elicited, and *Analysis Paradigms (L6)* describing how measures are modeled.

4.3.1 L3: Evaluation Goal

We observed several forms of comparisons that a study could use:

- **Compare impacts of multiple uncertainty visualizations** (255; 68.5%): Results of one or more tasks are used to rank two or more uncertainty visualizations (e.g., [11, 40, 53]).
- **Determine the impact of presenting uncertainty** (98; 26.3%): At least one visualization that does not contain uncertainty information is evaluated (e.g., [68]).
- **Validate effectiveness of an uncertainty vis** (5; 1.3%): A user study is used to show that a visualization improves some response(s) (e.g. [80]).
- **Understand why/how a visualization works** (54; 14.5%): Responses to a visualization(s) are analyzed to identify a mechanism or confirm that a hypothesized mechanism describes use (e.g., [87]).
- **Understand interactions with user characteristics** (53; 14.2%): The effect of properties of the user (expertise, location, etc.) on responses is analyzed (e.g., [1]).

Notably, 170 (45.7%) of the 372 evaluation paths that we coded that compared multiple visualizations did not include any other evaluation goals. We are nonetheless encouraged by the fact that roughly a quarter of paths (98) involved a comparison to a no-uncertainty baseline condition, a technique that others have suggested should be more common in the uncertainty visualization evaluation literature [33].

4.3.2 L4: Measures

Measures prescribe what aspects of users' behavior or beliefs are elicited to assess whether expected effects have been achieved.

- **Decision measures** (31; 8.3%)
  - **Decision** (31; 8.3%): A hypothetical choice action (e.g., describing when they would leave for the bus [23]).
  - **Allocation** (0%): An amount of currency wagered to endorse a choice (e.g., [74]).
- **Attribute-based measures** (141; 37.9%)
  - **Value of [nominal | ordinal | continuous] variable** (67; 18%): Specifying a value along a provided scale (e.g., asking the user to estimate the forecasted temperature from a display [46]).
  - **Count** (15; 4%): Specifying a count (e.g., of features of a certain type).
  - **Probability** (29; 7.8%): Specifying a probability directly as a probability or frequency (e.g., [40]) or as a Likert-style question asking for the relative probability of an event (e.g., [1]).
  - **Variance** (4; 1.1%): A value or set of values describing how much an outcome can vary (i.e., a range, standard deviation, standard error, etc. as in [31]).
  - **Rank by [data attribute | uncertainty attribute]** (22; 5.9%): The top *k* features sorted by a numerical attribute.
- **Self-report measures** (161; 43.3%)
  - **Spontaneous Impression** (16; 4.3%): An unconstrained description of one's reaction to a visualization (e.g., [8]).
  - **Subjective Confidence** (28; 7.5%): A valuation of one's confidence in information or a judgment.
  - **Perceived Risk/Reliability** (15; 4%): How risky, or conversely how reliable, one believes an outcome is.
  - **Perceived Ability/Effort** (29; 7.8%): How hard a visualization is to read or use.
  - **Self-reported Satisfaction** (41; 10.8%): An aesthetic valuation of a visualization.
  - **Self-reported Intention** (10; 2.7%): A reported intention to use, recommend, etc. some information or artifact (e.g., [17]).
  - **Willingness to Pay** (6; 1.6%): A reported intention involving allocating money to some information or artifact(s).
  - **Self-reported Action** (17; 4.6%): Whether one used, remembered, recommended, etc. some information or artifact.
- **Implicit measures** (36; 9.7%):
  - **Galvanized Skin Response** (GSR) (1; 0.3%): Change in the electrical resistance of the skin.
  - **Interaction logs** (35; 8.9%): A log of a user's behavior as they use a visualization.
  - **Eye movement** (2; 0.5%): An eye tracker log of voluntary or involuntary eye movement.

Often evaluators use directly elicited measures as input to transformations. For example, both Error and Bias are derived measures that can be calculated for a number of lower level measures given ground truth. We therefore distinguish between the measure that is directly elicited (L4a, above) and *Derived Measures* (L4b, below) which describe the possible transformations of the lower level measures.

- **Descriptive Error Measures** (130; 35%):
  - **Error** (112; 30.1%): Can be calculated as binary, continuous, or using Signal Detection theory (TP, TN, FP, FN).
  - **Bias** (18; 4.8%): Can be calculated as an ordinal(i.e., $<$, $=$, $>$) or continuous value.
- **Aggregate Descriptive Measures** (16; 4.3%):
  - **Precision** (2; 0.5%): How consistent a user's responses are.
  - **Variance** (4; 1.1%): How consistent a set of users' responses are.
  - **Relation of response with ground truth** (5; 1.3%): How much a subjective measure like confidence or perceived importance in a feature aligns with ground truth.
  - **Change in response** (5; 1.3%): How much a response (confidence, probability estimate, belief, etc.) changes after using a visualization (e.g., evaluating how much people's judgments about climate change realities change after viewing a visualization [35]).
- **Decision Measures** (14; 3.8%):
  - **Utility maximization** (7; 1.9%): How closely a response decision aligns with the optimal decision under utility theory.
  - **Compliance with optimal strategy** (7; 1.9%): How closely the user's responses follow the optimal decision strategy.
- **Explanatory Measures** (7; 1.9%):
  - **Shape of response function** (3; 0.8%): The shape of a user's inferred internal judgment function (e.g., [87]).
  - **Strategy** (4; 1.1%): The process describes how a user made a judgment.
- **Semantic Measures** (8; 2.2%):
  - **Risk Aversion** (2; 0.5%): The degree to which a direct measure, like a Willingness to Pay, implies a desire to avoid risk.
  - **Affective associations** (3; 0.8%): The degree to which a direct response (e.g., a rating along a scale spanning between two adjective pairs) is associated with one form of affect.

  - **Uncertainty association** (3; 0.8%): The degree to which a direct response is associated with uncertainty.

- **Implicit Measures** (38; 10.2%):
  - **Number of Interactions** (5; 1.3%): A count describing the frequency of an action (a mouse click, a page view, etc.)
  - **GSR Features** (1; 0.3%): For example, number or variance in GSR signals (e.g., [93]).
  - **Eye fixation measures** (2; 0.5%): Distance between sequential fixations, frequency between short distances, etc (e.g., [56]).
  - **Time** (29; 7.8%): The duration of a task completed by the user (session, question response, etc.).

Our results indicate that the most common category of Measure (L4a) in our sample was Self-Report Measures, with multiple specific measures (Subjective Confidence, Self-reported Satisfaction, and Perceived Ability/Effort each occurring in more than 5% of all evaluation paths). Attribute-based measures were also common, including Values of Nominal, Ordinal, and Continuous variables but also to some extent Probability and Rankings of entities by uncertainty or another quantitative variable. Decision Measures and Implicit measures were both notably less common.

More than half of all evaluation paths included Derived Measures (L4b). Not surprisingly, error measures were the most commonly used at roughly 30%, with the implicit measure time as the second most common at roughly 8%. Deriving bias (error with direction) occurred in roughly 5% of evaluation paths.

4.3.3 L5: Elicitation

Elicitation describes how measurements are generated. For example, a study that compares how different uncertainty visualizations (*L3 Compare impacts of multiple uncertainty visualizations*) impact users' confidence in their decisions (*L4 Subjective confidence*) could ask users to rate their confidence on a scale from 50 (random) to 100 (certain) implemented as a continuous slider with numerically labeled endpoints (*L5 Slider*). Research outside of visualization suggests that measures related to subjective uncertainty are sensitive to the elicitation process [28, 72]. By distinguishing among formats for collecting users' responses, our goal is to increase the level of awareness of elicitation as a critical design choice in uncertainty visualization evaluation and visualization evaluation more broadly.

We observed elicitation through:

- **Physical apparatus** (6; 1.6%): Interactions with physical objects (e.g., casts of the human body [85]).
- **Direct selection** (10; 2.7%): Input recorded directly through the stimuli (e.g., user clicks on location on map [12]).
- **Graphical** (15; 4%): The use of a visualization to gather responses, either by asking a user to construct a representation or by asking a user to adjust a representation (e.g., the position of a visualized outcome on a plot relative to error bars [3]).
- **Standard survey inputs** (245; 65.9%)
  - **Slider** (11; 3%): A visual analogue scale depicting a continuous range.
  - **Multiple choice** (83; 22.3%): A list of discrete options via radio button, checkbox, etc.
  - **Likert** (91; 24.5%): A stepped rating scale consisting of 5, 7, etc. points along a continuous range.
  - **Text entry** (60; 16.1%): An unconstrained (e.g., free text entry) text box, one constrained to only accept numeric answers, etc.
- **Oral** (20; 5.4%)
  - **Think-aloud protocol** (6; 1.6%): The user's utterances as they interact with a visualization.
  - **Spoken response** (14; 3.8%): A spoken question response.
- **Indirect** (35; 9.4%)
  - **GSR tracker** (1; 0.3%): The user wears a GSR device.
  - **Eye tracking** (2; 0.5%): The user sits in front of an eye tracker.
  - **Implicit interaction logs** (32; 8.6%): No direct elicitation.

The majority of evaluation paths involve use standard survey inputs, with Multiple Choice and Likert-style inputs being notably more common than Text-entry or Sliders. Combining this information with the prevalence of error measures and self-report measures, these results suggest that many uncertainty visualization evaluations rely on either self-reported assessments of how well a visualization worked, or accuracy measures calculated on highly constrained sets of options (e.g., 5 or less being typical of most multiple choice questions we observed). The next most common form of elicitation was Implicit, which most commonly took the form of logging time spent with a visualization. Also worth noting, 44 (11.8%) paths did not mention how they elicited a specified measure.

4.3.4 L6: Analysis

The analysis paradigm describes how the elicited measurements are summarized and used to assess to what degree the desired effects have been achieved. We distinguished between the following approaches:

- **Intermediate analyses**
  - **Qualitative coding** (33; 8.9%): Categorizing responses based on their similarities.
  - **Summary statistics** (310; 83.3%): Sample statistics like means, variance, or other frequency information.
- **Summative analyses**
  - **Frequentist NHST** (242; 65%): Null hypothesis significance testing within a Frequentist paradigm.
  - **Frequentist Estimation** (12; 3.2%): Frequentist parametric or non-parametric Frequentist approaches to infer and report a sampling distribution (e.g., 95% CIs presented in place of significance tests)
  - **Bayesian NHST** (0): Null hypothesis significance testing within a Bayesian paradigm.
  - **Bayesian Estimation** (12; 3.2%): Parametric or non-parametric Bayesian approaches to infer and report a sampling distribution

Two of these methods (Qualitative Coding and Summary Statistics) tend to be used as intermediate steps to concluding whether or not a visualization was successful. For example, Summary Statistics are often presented to add context to results obtained through NHST or Estimation. Qualitative coding is used most often on free form inputs such as free text responses describing subjects' spontaneous impressions to a visualization.

The majority of paths used Summary Statistics. 77 (20.7%) paths used Summary Statistics *without* relying on inferential statistics (NHST or estimation) to draw final conclusions. Not surprisingly, Frequentist NHST analyses dominate the use of inferential statistics in the uncertainty visualization studies we coded. We observed only 24 paths that used estimation, half of which used Bayesian models to estimate bias and variance separately; and no examples of Bayesian NHST.

5 Results and Recommendations

In contrast to code frequencies by level of the taxonomy, looking at co-occurring decisions in evaluation paths provides a more holistic view of how choices and assumptions at one level of the taxonomy may influence decisions at other levels. We characterize common paths, then present recommendations to reflect on where the patterns that emerge suggest missed opportunities in light of research on uncertainty elicitation and comprehension in other fields.

5.1 Characterizing Evaluation Paths

5.1.1 Interpretation & Semantics Paths

While the majority of studies with Performance or User Experience research questions implied a priori Expected Effects, studies with Interpretation and Semantics research questions were less likely to imply a specific direction of expected effect. Even where an Expected Effect was posited to address an Interpretation and Semantics research question (e.g., suggesting that an encoding should be perceived as more intuitive for presenting uncertainty [8, 64]), Interpretation and Semantics evaluation paths tended to assume a more exploratory approach to soliciting responses to a visualization. Such evaluations often elicited unconstrained spontaneous impressions of a visualization (16; 4.3%) or self-reported actions (17; 4.6%) using written inputs or oral approaches such as interviews and think-aloud protocols (21; 5.6%).

We observed an important subset of Interpretation and Semantics that used more constrained tasks to elicit internal representations of a probability distribution. Tak et al. [87] characterized participants' internal representations of a probability distribution by repeatedly asking them to specify the amount of uncertainty at a given location on a 2D visualization. Similarly, Bisantz et al. [4] asked participants to describe how much linguistic expressions (e.g., "probable") captured a probability value by asking participants to specify the "membership" of a given probability for an uncertainty representation along a continuous scale from "Not at all" to "Absolutely." Hullman et al. [39] asked users to "draw" their prediction for a distribution using a continuous or discrete representation of a probability density function.

Other Interpretation & Semantics paths focused on eliciting evidence of a user's mental process to confirm whether people were employing specific heuristics or biases in interpreting a visualization. For instance, Correll and Gleicher [11] evaluated uncertainty visualizations of confidence intervals to see if participants were subject to "within-the-bar" bias [70], where outcomes within the visual area of bar charts were perceived as more likely than those outside of the bar. Similarly, Correll et al. [12] examine how different color maps for uncertainty can result in participants under- or over-weighting uncertainty information when making decisions. Neither study's sole focus was a comparison with a ground truth correct decision. Instead, responses were examined for systematic patterns of decision-making that would indicate a bias or heuristic that might be undesirable in real-world decision-making tasks. Sometimes self-reports were utilized for this purpose; for example, Padilla et al. [73] used debriefings to evaluate the use of heuristics, breaking down patterns of responses by heuristic strategies used to make judgments about uncertainty.

Less commonly, Interpretation & Semantics paths examined how amount of uncertainty influenced bias or error [46, 47, 68], recognizing the findings of prospect theory in decision science [49].

5.1.2 Accuracy vs. Decision Quality

Multiple researchers have called for a greater focus on realistic decision making in uncertainty visualization evaluation [38, 55, 54, 79]. The focus on accuracy and performance measures that we and others [38, 55, 54, 79] have observed risks glossing over the difference between being able to identify information in a visualization and being able to use it effectively in a decision. Our analysis demonstrates the differences between these two approaches (Fig. 3).

Assessing Accuracy (L2) was roughly 10 times more common than assessing Decision Quality (L2) in our sample. Even when Decisions (L4) were elicited, by posing a hypothetical scenario in which the user should 'act', Increasing Accuracy was roughly twice as common compared to Increasing Decision Quality.

Paths that did focus on Decisions often provided a more thorough motivation for incentives, from designing decision payoffs based on formative studies of how people valued certain outcomes [23] to the development of multiple specific hypotheses related to the impact of incentives on users' decisions [9].

5.1.3 Assessing Confidence

Approximately 20% (72) of paths we observed emphasized constructs associated with one's subjective sense of confidence, including Subjective Confidence (28), Perceived Risk/Reliability (15), and Perceived Effort/Ability (29). Subjective confidence and Perceived Effort/Ability were typically framed as eliciting a participant's *subjective sense of the accuracy or effortfulness of their own judgment or of the external (visualized) information*. Perceived Risk/Reliability asked the user to report how likely they thought they were to be affected by some event, representing an *explicitly subjective version of probability*. Conversely, when studies asked participants to estimate perceived uncertainty via probabilities, rankings, or other measures that error was derived from, these constructs were framed as inquiring about *objective information*. Subjective confidence was typically treated as a secondary form of effect, with priority given to the "objective" measures to determine, for instance, which of multiple visualizations was best.

The lack of motivation for how these constructs do or should differ suggests that subjective uncertainty is not a well defined construct in visualization evaluation. Indeed, we see clear differences in assumptions about how to treat even the measures that are clearly framed as subjective. Some researchers address confidence and perceived risk as subjective feelings with no ground truth, aligning with theories that subjective confidence should not necessarily be expected to behave like statistical confidence [27]. For example, Blenkinsop et al. [5] interpret directional shifts in confidence between visualization conditions, saying things like "participants had more confidence" or participants expressed "low confidence in the use of these displays". In making such assumptions, evaluators maintain transparency about what they are measuring but sacrifice normative conclusions.

In contrast, some researchers take a normative approach by comparing confidence to a ground truth, assuming that these constructs carry a similar meaning across individuals, and that that meaning is more or less directly comparable to an "objective" statistical account. Such approaches assume that confidence provided for a judgment should be a signal of a participant's probability of being correct for that judgment, or of some objective measure of probability. For example, Ibrekk and Morgan [41] analyzed intrasubject correlations, looking for an association between how sure a participant was about their response being correct and the correctness (i.e., error) of the response. Correll and Gleicher [11], on the other hand, evaluated how confidence decayed as a potential outcome got further from the mean, as well as how confidence correlated with effect size and p-value.

5.2 Recommendations for Evaluators

Our analysis surfaced biases, e.g., toward Performance and Accuracy, confirmatory goals and analysis, as well as opportunities for expanding the default "toolbox" such as beyond standard constrained HTML input elicitation through Multiple choice and Likert-style responses. Drawing on our observations and research on uncertainty comprehension and elicitation from outside of visualization, we propose recommendations toward more transparent, and internally and externally valid evaluations of uncertainty visualizations.

5.2.1 R1: Consider Explaining As An End Goal

As described above, studies that posed Behavioral Targets of Interpretation & Semantics were the most likely to work toward a goal of understanding how or why users acted in certain ways when using an uncertainty visualization. A simple way to do so is to elicit participant's descriptions of how they made a judgment, or what information they found helpful, to provide greater context for differences that might be observed. More sophisticated strategies for explaining uncertainty visualization including analyzing results for evidence of heuristic use [11, 12, 40, 70] or prospect theory [47, 48, 68]. We propose that uncertainty visualization evaluators in particular have a responsibility to attempt to elicit and explain why observed differences between conditions might exist due to the complexity of uncertainty comprehension and associated heuristics.

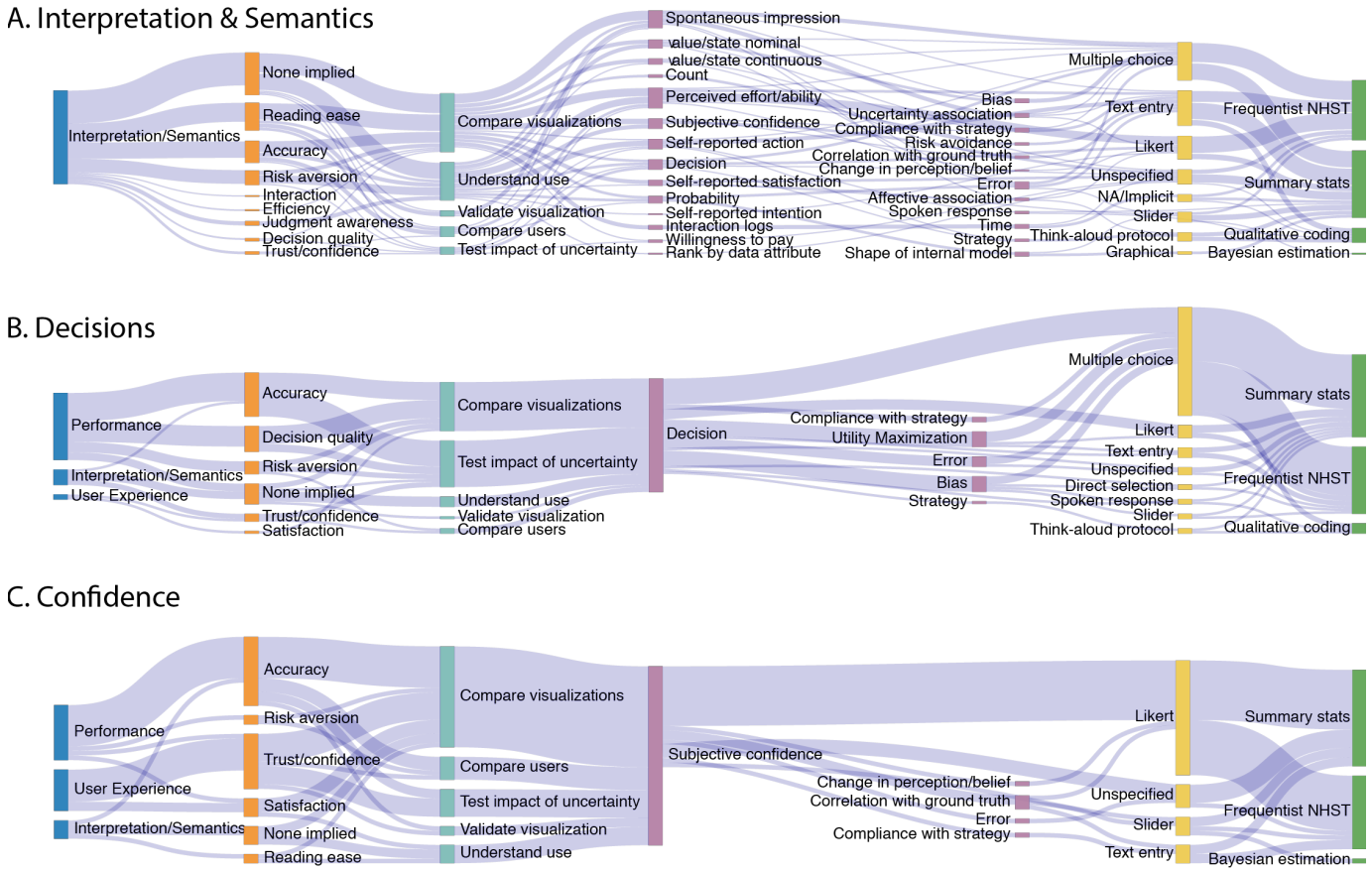

Fig. 3. Paths with Interpretation & Semantics targets (A), and Decisions (B) and Subjective Confidence as measures (C).

5.2.2 R2: Use Decision Frameworks for Realism & Control

An accuracy approach can establish, at base, how well a user can extract a probability or other information necessary to make an informed decision—but not that the user would know how to use that probability to make an informed decision. If estimation error were the only contributing factor in differences in decision quality between representations, designers should simply choose the most perceptually effective uncertainty encoding (i.e., that which best achieves L2's goal of Increasing Accuracy) in all cases. However, research shows that different framings can affect interpretation of uncertainty: e.g., frequency framings can improve Bayesian reasoning about probability [25], and some representations may be more susceptible to *deterministic construal errors* (e.g. mistaking a confidence interval around a predicted daily temperature as a deterministic prediction of the daily high/low temperatures [46]). Thus it is not a foregone conclusion that higher Accuracy (L2) also leads to better Decision Quality (L2).

Eliciting decisions, in the form of hypothetical actions like where to locate an airport or park [58] or which location the user would suggest a person buy [22] allows the evaluator to examine whether users understand uncertainty in light of a specific decision context. Other questions that can be answered via decision tasks include whether a subject feels confident enough to use the information given the potential rewards or consequences of the decision. In a real-world decision context, participants have consequences for choosing incorrectly, and may even have conflicting decision criteria or trade-offs to make (e.g., balancing the chance of an incorrect snow forecast against the cost of failing to ice the road [46]).

Economic frameworks, such as Utility Maximization (L4 Derived) offer one approach to assessing decision quality. Joslyn and LeClerc [48, 46], for example, give participants a budget and have them decide whether to salt the road given weather forecasts with different displays, with costs for failing to salt if it does snow. The remaining budget after many trials becomes a measure of decision quality. Similarly, Fernandes et al. [23] follow up a prior study [53] that assessed the Accuracy (L2) of probability extraction from probabilistic predictions of bus arrival times with an incentivized experiment where participants decide when to catch simulated buses. Participants are paid according to a utility function, and the ratio of their expected payment under the decisions they made to the payment under an optimal strategy is used to assess Decision Quality (L4 Derived).

These approaches offer a controlled means of decision assessment that attempts to include realistic tradeoffs. Given multiple trials, they can be used to compare how well people can learn to use uncertainty displays as they make more decisions and receive feedback.[1] We observed that in contrast to accuracy studies, researchers who applied such approaches also tended to include controlled comparisons between no uncertainty and a visual representation, which can help answer in what situations uncertainty helps or not.

5.2.3 R3: Account for the Ambiguity of Confidence

All of these approaches, when used as described above, do not account for several known properties of confidence reporting. The "hard-easy effect" is a counter-intuitive pattern of behavior in which people report overconfidence when presented with a difficult task and underconfidence when presented with an easy task [18, 66]. This is partially explained by people's tendency to report low confidence for incorrect judgments, especially when made in the face of an easy task [81]. Hence, an analysis like the one used by Ibrekk and Morgan, which uncovered no clear differences between visualizations based on confidence reporting, may have been confounded by the difference between confidence reported for a judgment that is likely to be wrong versus one that is likely to be right. Only two studies in our sample of 86

[1] For observed learning rates to be reliable as a signal of how easy to learn different visualizations are in the world requires that the conditions of decision making in the experimental task be a reasonable proxy for decision making in the real world. This is debatable.

that explicitly accounted for the difference in confidence reports for a wrong versus a right answer, in this case by weighting confidence values for a correct judgment significantly more than those for an incorrect judgment [20, 52].

A second issue that was not explicitly addressed in any of the confidence evaluation paths we observed is that confidence reports are known to be subject to considerable noise. For example, experience, effort, and information availability can lead to increased confidence without comparable increases in accuracy [10, 75], confidence can be inconsistent even for the same task, and participants may report confidence using idiosyncratic methods like rounding values. This can mean that subjective confidence is not a direct linear function of statistical confidence, and that trial order and other contextual factors should be accounted for. We conclude that evaluators should clearly state their motivation for why they consider a certain confidence outcome (e.g., higher confidence) to be superior and test for effects of binary accuracy and experience (e.g., trial number) on results.

5.2.4 R4: Validate Elicited Responses

Research outside of visualization describes how subjective accounts of uncertainty can be sensitive to the elicitation method as well as other properties of a task context or user [72]. We observed several ways that evaluators in our sample sought to increase the validity of the responses they gathered. We highlight a few of these techniques to encourage further use of validations for uncertainty visualization.

Recognize the effects of priors: A few studies allowed for the fact that users may have different prior beliefs or perceptions of an event for which uncertainty is shown by using pre- and post-visualization questions. Herring et al. [35] assessed the change in participants' beliefs about climate change after seeing a visualization. Ibrekk and Morgan [41] examined whether explaining a visualization helped users better understand it relative to their base knowledge.

Calibrate the user: In real life, uncertainty is often experienced as the percentage of the time when an event occurs versus does not (e.g., catching versus missing one's daily bus when leaving the house at 8am). By definition, under uncertainty it is possible to make the "right" decision (e.g., getting to the bus stop at the optimal time to minimize average wait), but still experience the "wrong" outcome. Studies that measured decision quality against utility theory were most likely to provide feedback on the uncertain outcomes they asked participants about [23, 48, 68], though some studies that used simpler error measures also incorporated feedback for realism and calibration [9].

Use graphical elicitation to reduce noise: Graphical elicitation of subjective uncertainty has been shown to reduce error in subjective probability reporting [28]. Even when the elicited measure is not subjective probability, graphical elicitation increases the likelihood that participants will process the visualization sufficiently to provide their best guess. For example, in Belia et al.'s study of error bar interpretation (a task that is known to be error prone) participants were asked to adjust the position visualized mean with an error bar so that it was just statistically significantly different from another mean [2].

Allow for "no answer" as a valid response: A few evaluation paths specifically allowed for the fact that a participant faced with uncertain options might find the information too ambiguous to feel confident enough to point to a difference [41, 73]. For example, Padilla et al. [73] added an additional study in their work to allow for a judgment that multiple locations along a possible hurricane path were approximately equal inexpected damage. This approach accounts for the case where a user is not confident enough to point to a difference.

Account for numeracy and other subject-specific characteristics: In addition to studies that explicitly evaluated how user characteristics impacted use of a visualization (L3 - Understand interactions with user characteristics), multiple studies in our sample controlled for ability covariates that could influence performance. For example, evaluators measured numeracy (e.g., [32, 47, 53]), subjective numeracy [73], and spatial ability (e.g., [53, 60]) using established scales.Using standardized instruments: Some evaluators also employed standard elicitation instruments. This tended to occur in studies of Bayesian reasoning, where problems are well defined [47, 60], but we also observed evaluators using standard scales for aspects of user experience [79], such as the User Experience Questionnaire [49].

Compare to textual uncertainty: Another way in which evaluators sought to increase the validity of their results was by comparing uncertainty visualizations to text representations (e.g., [3, 20, 39]) for added insight about why visualizations may or may not have helped. Including text conditions also acknowledged the potential for visualizations—and in particular uncertainty visualizations—to present more information than needed relative to text [80].

6 Discussion: Subjective Uncertainty as Construct

Our results suggest a mismatch between how uncertainty comprehension is viewed in visualization evaluation versus how it is viewed in fields that deal more directly with cognition (e.g., judgment and decision making). Our results suggest that visualization evaluators show either an accuracy-efficiency bias that assumes that it is sufficient to measure the effectiveness of a visualization by comparing either how accurately or efficiently people can make judgments from multiple uncertainty visualizations using standard survey style prompts or rely on explicit value judgments, where users are asked directly whether uncertainty visualization helped them to make a judgment. Further, we find a tendency toward confirmation over explanation, where the majority of evaluation studies do not try to explain the effects that they find. We also find that simpler alternatives (No uncertainty or text representations) are considered in a minority of evaluations.

As others have noted, the accuracy-efficiency bias introduces a risk that the effects identified in a study will not persist in real decision tasks, where decision makers are incentivized by real world consequences. We note the importance of decision feedback for calibrating decisions, and we describe how decision paradigms informed by economic theory can help maintain both the control and realism that others have argued is difficult to achieve in this context [55, 79].

Further, evidence from other discplines suggests that people are not very good at making accurate judgments about their own ability to make judgments under uncertainty [72]. Again, attempts to understand strategies could prove useful for surfacing whether a person's sense of confidence in themselves or a visualization is not aligned with the evaluator's expectations for use. We also suggest that how well a person can assess their performance should be explicitly modeled where possible, such as in using confidence reports [81].

Possible threats to validity arise when evaluators seek to confirm hypotheses without also seeking to understand and explain why differences exist between uncertainty visualizations. One possible reason for differences in user performance is that people tend use heuristics to simplify judgments from uncertainty (e.g., [26, 51, 50, 92]). Finally, omitting simpler presentations makes it less likely that evaluators will realize when uncertainty information is unnecessary for the task or being disregarded. The evaluation paths we observed that attempt to uncover strategies used to respond to a task therefore provide an important lesson for future evaluation work.

The studies in our sample rarely, if ever, acknowledged differences among alternative conceptions of uncertainty such as risk, ambiguity, and error. Research outside of visualization provides empirical evidence for different conceptions of uncertainty held by decision makers [34, 36, 60, 65]) and may be informative for visualization researchers considering meanings of subjective uncertainty.

7 Conclusion

We present a taxonomy of methods for evaluating uncertainty visualizations and describe the results of a qualitative analysis applying our framework to 86 publications which represent the state of uncertainty visualization evaluation. Our results indicate that current evaluation practices focus primarily on a small set of Performance and User Experience concerns in order to compare uncertainty visualization designs. While these studies about Performance and User Experience tend to seek "confirmatory" evidence for the superiority of some visualization technique, a much smaller set of studies addressing issues of Interpretation tend to be more "explanatory". We characterize overall trends in *evaluation paths* (i.e., the co-occurrence of codes in our taxonomy)

which indicate distinctions between methods for measuring Accuracy and Decision, as well as different methods for eliciting and assessing Subjective Confidence. Drawing on related research in judgment and decision making, we recommend specific steps that researchers should take when designing uncertainty visualization evaluations in order to strive for valid and transparent findings.

8 Acknowledgements

The first author thanks Dan G. Goldstein for feedback on the work.

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] M. Allais. Rational man's behavior in the presence of risk: Critique of the postulates and axioms of the american school. *Econometrica*, 21(4):503–46, 1953.

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

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

[5] 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.

[6] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*, pages 3–27. Springer, 2014.

[7] 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.

[8] N. Boukhelifa and D. J. Duke. Uncertainty visualization: why might it fail? In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4051–4056. ACM, 2009.

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

[10] J. Chung and G. Monroe. The effects of experience and task difficulty on accuracy and confidence assessments of auditors. *Accounting & Finance*, 40(2):135–151, 2000.

[11] 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.

[12] M. Correll, D. Moritz, and J. Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018. to appear.

[13] M. Daradkeh. Exploring the use of an information visualization tool for decision support under uncertainty and risk. In *Proceedings of the The International Conference on Engineering & MIS 2015*, page 41. ACM, 2015.

[14] D. D. Davis and C. A. Holt. *Experimental economics*. Princeton university press, 1993.

[15] S. Deitrick. Evaluating implicit visualization of uncertainty for public policy decision. *Proceedings AutoCarto 2012*, 2012.

[16] M. Dong, L. Chen, L. Wang, X. Jiang, and G. Chen. Uncertainty visualization for mobile and wearable devices based activity recognition systems. *International Journal of Human–Computer Interaction*, 33(2):151–163, 2017.

[17] X. Dong and C. C. Hayes. Uncertainty visualizations: Helping decision makers become more aware of uncertainty and its implications. *Journal of Cognitive Engineering and Decision Making*, 6(1):30–56, 2012.

[18] J. Drugowitsch, R. Moreno-Bote, and A. Pouget. Relation between belief and performance in perceptual decision making. *PLoS ONE*, 9(5), 2014.

[19] Edelman. 2017 edelman trust barometer, 2017.

[20] L. D. Edwards and E. S. Nelson. Visualizing data certainty: A case study using graduated circle maps. *Cartographic Perspectives*, (38):19–36, 2001.

[21] D. Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

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

[23] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. 2018.

[24] T. I. O. for Standardization (ISO). Guide to the expression of uncertainty in measurement, 1997.

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

[26] G. Gigerenzer, P. M. Todd, t. ABC Research Group, et al. *Simple heuristics that make us smart*. Oxford University Press, 1999.

[27] C. Glymour. *Why I am not a Bayesian*. Citeseer, 1981.

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

[29] M. Greis, E. Avci, A. Schmidt, and T. Machulla. Increasing users' confidence in uncertain data by aggregating data from multiple sources. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 828–840. ACM, 2017.

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

[31] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE transactions on visualization and computer graphics*, 22(1):539–548, 2016.

[32] C. Hansen, S. Zidowitz, F. Ritter, C. Lange, K. Oldhafer, and H. K. Hahn. Risk maps for liver surgery. *International journal of computer assisted radiology and surgery*, 8(3):419–428, 2013.

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

[34] R. HASTIE. Concepts in judgment and decision research-anderson, bf, deane, dh, hammond, kr, mcclelland, gh, shanteau, jc, 1982.

[35] J. Herring, M. S. VanDyke, R. G. Cummins, and F. Melton. Communicating local climate risks online through an interactive data visualization. *Environmental Communication*, 11(1):90–105, 2017.

[36] R. M. Hogarth and H. Kunreuther. Decision making under ignorance: Arguing with yourself. *Journal of Risk and Uncertainty*, 10(1):15–36, 1995.

[37] I. Huff, C. Weigle, and D. C. Banks. Ensemble-space visualization improves perception of 3d state of molecular dynamics simulation. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 163–170. ACM, 2008.

[38] J. Hullman. Why evaluating uncertainty visualization is error prone. *Proc. BELIV*, 2016.

[39] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE transactions on visualization and computer graphics*, 24(1):446–456, 2018.

[40] 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.

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

[42] J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[43] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, page 6. ACM, 2008.

[44] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[45] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, 2003.

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

[47] S. Joslyn, K. Pak, D. Jones, J. Pyles, and E. Hunt. The effect of probabilistic information on threshold forecasts. *Weather and Forecasting*, 22(4):804–812, 2007.

[48] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, 18(1):126, 2012.

[49] D. Kahneman. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278, 1979.

[50] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[51] D. Kahneman, P. Slovic, and A. Tversky. editors, judgment under uncertainty: Heuristics and biases, 1982.

[52] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Animated hypothetical outcome plots help untrained observers judge trends in ambiguous data. *Visualization and Computer Graphics, IEEE Transactions on*, 25(1), 2019.

[53] 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.

[54] C. Kinkeldey, A. M. MacEachren, M. Riveiro, and J. Schiewe. Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science*, 44(1):1–21, 2017.

[55] C. Kinkeldey, A. M. MacEachren, and J. Schiewe. How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51(4):372–386, 2014.

[56] C. Kreuzmair, M. Siegrist, and C. Keller. High numerates count icons and low numerates process large areas in pictographs: Results of an eye-tracking study. *Risk Analysis*, 36(8):1599–1614, 2016.

[57] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2012.

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

[59] H. Li, C.-W. Fu, Y. Li, and A. Hanson. Visualizing large-scale uncertainty in astrophysical data. *IEEE transactions on visualization and computer graphics*, 13(6):1640–1647, 2007.

[60] R. Lipshitz and O. Strauss. Coping with uncertainty: A naturalistic decision-making analysis. *Organizational behavior and human decision processes*, 69(2):149–163, 1997.

[61] C. Lundström, P. Ljung, A. Persson, and A. Ynnerman. Uncertainty visualization in medical volume rendering using probabilistic animation. *IEEE transactions on visualization and computer graphics*, 13(6):1648–1655, 2007.

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

[63] 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.

[64] E. B. Mandinach and E. Gummer. Navigating the landscape of data literacy: It is complex. *Washington, DC and Portland, OR: WestEd and Education Northwest*, 2012.

[65] J. G. March and J. P. Olsen. *Ambiguity and choice in organizations*. Universitetsforlaget, 1979.

[66] E. C. Merkle. *Psychonomic Bulletin Review*, 16(1):204–213, 2009.

[67] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6), 2009.

[68] L. Nadav-Greenberg and S. L. Joslyn. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, 2009.

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

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

[71] NRC. Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts, 2006.

[72] 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.

[73] L. M. Padilla, G. Hansen, I. T. Ruginski, H. S. Kramer, W. B. Thompson, and S. H. Creem-Regehr. The influence of different graphical displays on nonexpert decision making under uncertainty. *Journal of Experimental Psychology: Applied*, 21(1):37, 2015.

[74] P. W. Paese. Uncertainty assessment accuracy and resource allocation outcomes: an empirical test of a presumed relation. *The Journal of psychology*, 127(4):443–450, 1993.

[75] P. W. Paese and J. A. Sniezek. Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Organizational Behavior and Human Decision Processes*, 48(1):100–130, 1991.

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

[77] K. Potter. Uncertainty visualization and data references.

[78] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.

[79] P. S. Quinan, L. M. Padilla, S. H. Creem-Regehr, and M. Meyer. Towards ecological validity in evaluating uncertainty. In *Proceedings of Workshop on Visualization for Decision Making Under Uncertainty (VIS'15). http://vdl. sci. utah. edu/publications/2015_vdmu_ecological-validity*, 2015.

[80] C. Roessing, A. Reker, M. Gabb, K. Dietmayer, and H. P. Lensch. Intuitive visualization of vehicle distance, velocity and risk potential in rear-view camera applications. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 579–585. IEEE, 2013.

[81] J. I. Sanders, B. Hangya, and A. Kepecs. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, 90:499–506, 2016.

[82] 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.

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

[84] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7. ACM, 2006.

[85] A. L. Simpson, B. Ma, E. M. Vasarhelyi, D. P. Borschneck, R. E. Ellis, and A. James Stewart. Computation and visualization of uncertainty in surgical navigation. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 10(3):332–343, 2014.

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

[87] 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.

[88] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, 1994.

[89] 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.

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

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

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

[93] J. Zhou, S. Z. Arshad, X. Wang, Z. Li, D. D. Feng, and F. Chen. End-user development for interactive data analytics: Uncertainty, correlation and user confidence. *IEEE Transactions on Affective Computing*, 2017.

[94] T. Zuk and S. Carpendale. Theoretical analysis of uncertainty visualizations. In *Visualization and data analysis 2006*, volume 6060, page 606007. International Society for Optics and Photonics, 2006.