Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data

Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman

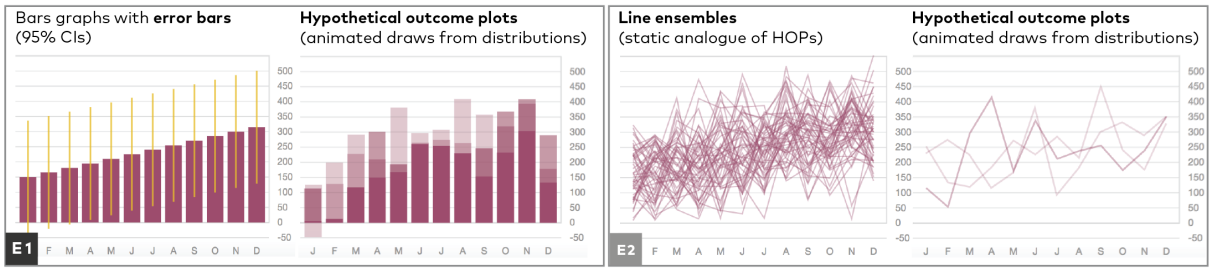

Fig. 1. We present two experiments (E1 and E2) evaluating four different uncertainty visualizations (from left to right): bar graphs with error bars, bar hypothetical outcome plots (HOPs), static line ensembles, and line HOPs.

Abstract—Animated representations of outcomes drawn from distributions (hypothetical outcome plots, or HOPs) are used in the media and other public venues to communicate uncertainty. HOPs greatly improve multivariate probability estimation over conventional static uncertainty visualizations and leverage the ability of the visual system to quickly, accurately, and automatically process the summary statistical properties of ensembles. However, it is unclear how well HOPs support applied tasks resembling real world judgments posed in uncertainty communication. We identify and motivate an appropriate task to investigate realistic judgments of uncertainty in the public domain through a qualitative analysis of uncertainty visualizations in the news. We contribute two crowdsourced experiments comparing the effectiveness of HOPs, error bars, and line ensembles for supporting perceptual decision-making from visualized uncertainty. Participants infer which of two possible underlying trends is more likely to have produced a sample of time series data by referencing uncertainty visualizations which depict the two trends with variability due to sampling error. By modeling each participant's accuracy as a function of the level of evidence presented over many repeated judgments, we find that observers are able to correctly infer the underlying trend in samples conveying a lower level of evidence when using HOPs rather than static aggregate uncertainty visualizations as a decision aid. Modeling approaches like ours contribute theoretically grounded and richly descriptive accounts of user perceptions to visualization evaluation.

Index Terms—uncertainty visualization, hypothetical outcome plots, psychometric functions

1 Introduction

Effective communication of uncertainty, probability, and random sampling is necessary for scientific literacy among the public and for the practice of reproducible science. For example, confusing presentations of uncertainty in weather forecasts may lead people to discount uncertainty in the forecast, inducing a false sense of security about predicted outcomes. This kind of misunderstanding erodes public trust in science [7, 39]. Among scientists, misunderstandings of sampling error and the likelihood of replicating experimental results contribute to rampant use of underpowered studies and the "replication crisis" [37, 62]. A core challenge in communicating uncertainty information is how to help audiences recognize that estimates are subject to variability in the process which produces them [21, 54, 60]. This is especially difficult when audiences are unfamiliar with the statistical abstractions commonly used to express these concepts.

- *Alex Kale is with the University of Washington. E-mail: kalea@uw.edu.*
- *Francis Nguyen is with the University of Washington. E-mail: fmnguyen@uw.edu.*
- *Matthew Kay is with the University of Michigan. E-mail: mjskay@umich.edu.*
- *Jessica Hullman is with Northwestern University. E-mail: jessica.hullman@northwestern.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/

Data visualizations communicate complex information by offloading cognitive work as to automatic perceptual processing [43]. Visual metaphors could help audiences make sense of otherwise inaccessible uncertainty information. However, commonly used uncertainty visualizations often lead to misinterpretations. For example, error bars encoding confidence intervals or standard errors are easily misunderstood [6, 31] perhaps because such frequentist statistics are misinterpreted as indicating the probability of an estimate rather than the variability in the process which produced that estimate. Similarly, the nuances of statistical abstractions make it hard for many people to interpret probability density plots like gradient plots and violin plots.

Recently, Hullman et al. [36] defined a form of animated uncertainty visualization called Hypothetical Outcome Plots (HOPs). HOPs present uncertainty as a set of animated frames, each depicting a sample from a distribution of possible outcomes. Hullman et al. [36] found that HOPs facilitate comparable and much better estimation of univariate and multivariate distributional information, respectively, than error bars and violin plots. Importantly, HOPs express shared variation among multiple variables via the correlation of samples across animated frames, whereas static visualizations are not generally expressive of this shared variability. HOPs are especially flexible across applications since they do not introduce additional graphical marks (e.g., error bars) or encodings (e.g., color, transparency) to encode the variability of a distribution.

However, it remains unknown whether HOPs also have advantages for more applied tasks resembling real world judgments about data. For a realistic investigation of comprehension, we must go beyond the ability to *read probabilities* from a visualization and look at judgment tasks where people must *choose between alternative interpretations* of uncertain data (i.e., models that generated the data) in order to act. We

present two experiments that address gaps in knowledge about the impacts of HOPs by modelling perceptual decision-making in a realistic uncertainty communication task.

We first motivate our selection of an experimental task through a qualitative analysis of how visualizations are used to communicate uncertainty in 22 examples of uncertainty visualization in the news. Our results provide an overview of the visual and cognitive demands associated with judgments of uncertainty visualizations that the public encounter in news contexts.

Our primary contributions are two controlled experiments on the impact of HOPS versus static alternatives on user performance in an uncertainty communication task. Based on our qualitative analysis, we adapt an example of HOPs from the New York Times (NYT) [38] to examine the impact of different uncertainty encodings on the perceptual decision-making of Amazon Mechanical Turk (MTurk) workers. Participants infer which of two possible underlying trends is more likely to have produced a sample of time series data by referencing uncertainty visualizations which depict the two trends with variability due to sampling error (Fig. 3). Each participant makes many judgments about the more likely trend despite noise due to sampling error, using one of multiple uncertainty visualizations (Fig. 1) to aid in the task. For each participant's responses, we fit behavioral models called psychometric functions (PFs) [27, 42] which estimate accuracy on the task as a function of the strength of evidence in the samples judged. PFs enable us to measure the sensitivity of each observer to evidence in the task under different visualization conditions. Our results suggest that users of HOPs make correct judgments at lower levels of evidence than users of error bars and line ensembles.

2 Related Work

2.1 Uncertainty Visualization

Static visualizations, such as error bars, ensembles, and probability density functions (PDFs), are currently the predominant visual encoding for communicating uncertainty. These visualizations express distributions in a single view by encoding summary statistics or showing the whole distribution. Static visualizations align with a design goal of cognitive efficiency [11, 34, 43, 61]. In contrast, animations are thought to be less efficient because they require the integration of information across frames and pose challenges related to memory and attention [64]. However, outside of a few early studies of animated uncertainty visualizations in cartography [16, 17, 26], little empirical work has attempted to evaluate the potential for animation to provide an effective visual metaphor for probabilistic information.

Static uncertainty visualizations often assume that the audience has some baseline of statistical knowledge [6] and a propensity to connect such abstract concepts to graphical representations [15]. Error bars in particular can be difficult to interpret due to inconsistency of conventions governing what constructs (i.e., confidence intervals, standard errors, etc.) they represent [14, 31]. For graphs with enclosed shapes, within-the-bar bias can impact estimates of likelihood such that values within the shape tend to be judged as more likely than values outside [14, 50]. Uncertainty visualizations which use separate marks to encode expected value and uncertainty can elicit a heuristic bias of attention toward expected value and away from uncertainty [36], in keeping with a well-documented tendency for people to underweight or ignore uncertainty in decision-making [63]. In general, static visualizations which only express summary statistics fail to fully convey the underlying distribution since dissimilar datasets can have similar summary statistics [12]. Although variants of PDFs (i.e, violin plots, gradient plots) are, in theory, fully expressive of the underlying distribution, they still rely heavily on the statistical literacy of the observer.

In contrast, HOPs are designed to promote the integration of uncertainty information through *experience rather than description*. While viewers of static visualizations must exert cognitive effort to map graphical encodings to statistical concepts which may be unfamiliar [15, 66], viewers of HOPs can directly extract frequency information [30]. Observers who are allocating attention to a frequency encoding extract frequency information automatically, even when they are not aware of doing so [29]. Research on statistical reasoning suggests that frequency-oriented framing of probability can lead to more accurate statistical reasoning in many judgment contexts [18, 24, 32, 35, 40] and to more accurate elicitation of subjective probability distributions among lay observers [25]. Prior work establishes that HOPs enable more accurate inferences about joint probabilities than static uncertainty visualizations and that HOPs are comparable to static representations for estimating distributional properties of univariate distributions [36]. Users of error bars and violin plots only outperformed users of HOPs when estimating the central tendency of distributions with high variance.

Researchers and practitioners have applied animated and static sampling-oriented approaches to uncertainty visualization. Researchers have used animated sampling-oriented visualizations in medical volume rendering [48] and estimates of geospatial surfaces [16, 17, 26]. Static sampling-oriented quantile dotplots have been used to aid decision-making about bus arrival times [18, 41] and to help users to make graphically elicited predictions about experimental data [35]. A recent study investigated HOPs of hurricane locations [47], suggesting a role for sampling-oriented presentations of uncertainty in weather forecasts. In recent years, there have been implementations of HOPs, interactive simulations, and static sampling-oriented ensembles in the media which aim to communicate uncertainty in the job market [38] and in political elections [1, 9, 55], suggesting there is potential for HOPs to be used to communicate uncertainty information to the public in informal contexts.

2.2 Vision Science in Information Visualization

Vision science research on perception of ensembles, groups of visual objects spread over space and time, is directly applicable to data visualizations. Accumulating evidence suggests that the visual system is capable of quickly, automatically, and accurately extracting rich mental representations of ensembles from visual scenes [3, 4, 33, 44] even when the observer has no explicit memory of some ensemble members. The precise limitations of this mode of perception are a topic of ongoing inquiry. Observers are similarly accurate in perceptual averaging tasks whether stimuli are dynamically changing or arranged on a static spatial plane [4, 44]. However, in the case of dynamically changing stimuli, limitations on the duration and temporal frequencies for which this perceptual averaging is effective are not yet known. Observers can accurately report the average size of dynamic stimuli [2], the average expression of a dynamically changing face stimulus [28], and even more abstract features such as the average lifelikeness of a set of objects [44]. Sampling-oriented visualizations of uncertainty like HOPs seem particularly well matched to the visual system's ability to form gist representations from experience.

Psychophysics, a methodology in vision science, enables researchers to model the relationship between the presentation format of information and the perceptual decision-making of observers. In visualization research, psychophysics is most often used to estimate just-noticeable differences (JNDs) for a visual encoding, the change in some quantified property of that encoding which would be discernible to a subject 75% of the time. JNDs have been applied in the creation of perceptually-uniform color palettes (i.e., CIELAB [20]) and perceptually-driven color palette selection tools [45, 59]. The inferential models which researchers use to estimate JNDs are called psychometric functions (PFs). PFs estimate an observer's accuracy in judging a stimulus as function of the strength of evidence presented. PFs are based on signal detection theory (SDT) [27] (Fig. 2), a model of how we make decisions by weighing evidence in favor of two alternative interpretations. SDT posits that for any judgment between two alternatives internal receivers in the brain register the evidence for each possible alternative. Since the response of these receivers to a presented stimulus is subject to noise, and observers are inconsistent in the criteria they use to decide between alternatives, judgments of a stimulus err systematically. PFs model this systematic error by estimating both the amount of bias and random noise introduced in the process of perceiving the strength of evidence about the decision.

Our experiments incorporate PF analysis and SDT into the context

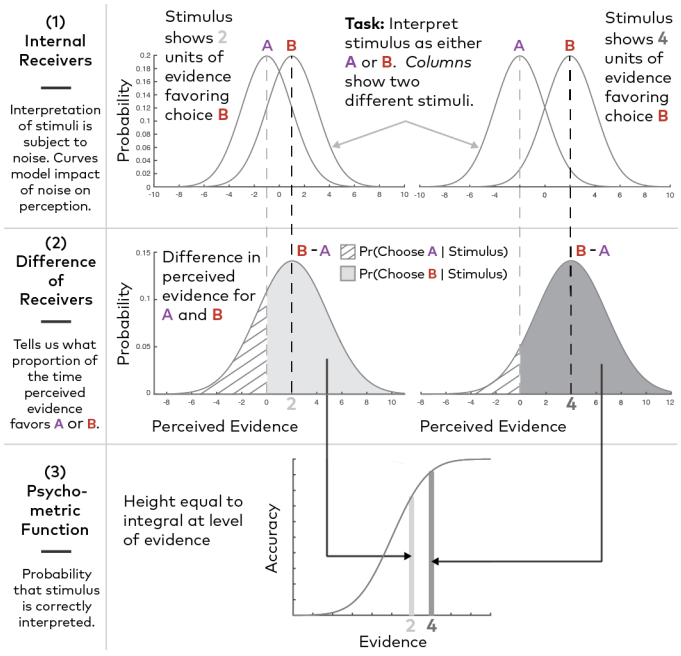

Fig. 2. We illustrate how to use signal detection theory to estimate the accuracy of an observer on a two-alternative forced choice task for two different stimuli (left and right columns). Signal detection theory posits internal receivers for two alternative interpretations of a stimulus, choices A and B. The response of these receivers to a stimulus is probabilistic and represents the perceived strength of evidence in favor of each choice. The two stimuli shown in the left and right column each convey levels of evidence equal to the difference between the black and gray vertical lines on the x-axes. The middle row depicts the difference between the responses of the receivers for choices A and B, shown in the top row. The estimated accuracy of the observer for each stimulus is equal to the integral of the difference distributions in the middle row. Estimated accuracy of stimulus interpretation for each level of evidence is equal to the height of the psychometric function.

of a realistic data interpretation task. Although other studies in the information visualization literature use PFs to investigate JNDs in low-level visual encodings such as color (i.e., [59]), the application of psychophysics in abstract conceptual domains, such as the interpretation of visualized uncertainty, is uncommon in information visualization research. PFs are an ideal way to measure the effectiveness of uncertainty visualizations for supporting alternative interpretations of data because they enable nuanced comparisons of user sensitivity to evidence under different display conditions.

3 Identifying Uncertainty Judgments in the News

Readers of the news often encounter uncertainty information which they interpret in order to make decisions. For example, a person might read a piece of data journalism from the New York Times (NYT) about expected changes under the 2017 GOP tax plan [10][1] and decide to take the standard deduction instead of itemizing their taxes. Since the news is perhaps the most common venue in which people encounter uncertain data in everyday life, we conducted a formative content analysis investigating the use of uncertainty visualizations in the news and how people might interpret visualized uncertainty in this context. We use this analysis to inform our experimental task.

3.1 Selection of Articles

We gathered a convenience sample of 22 articles from mainstream news in the US and UK. We started with recent visualizations from publishers like the NYT, and supplemented this set by conducting Google News searches containing terms like 'uncertainty' or 'probability' and 'visualization' or 'chart' (see Supplemental Materials). Articles were included in our analysis (1) if they are primarily concerned with communicating uncertainty information and (2) if they contain at least one visualization which conveys that there are a *distribution of possible values* which an observation or prediction might take.

While we started with 32 articles which met our first inclusion criterion, we had to exclude ten of these articles because they failed to meet our second criterion. In these cases, the visualizations only encoded single values of each estimate (e.g., a central tendency), and visualized distributions were contextualized as separate events plotted together by geographic location or time (e.g., an estimate for each of multiple countries). That the media use distributions of outcomes for separate events to illustrate uncertainty speaks to a disconnect between public concerns about uncertainty and the notion of uncertainty employed in information visualization research, in which uncertainty is most often defined as an expression of error in an estimate [54, 60]. The sample of articles considered reflects the time frame of the study and should be considered representative of uncertainty visualizations published in the popular press from 2014 to early 2018.

3.2 Qualitative Coding

Our goal was to identify a set of distinctions that could help visualization researchers understand the types of tasks implied by uncertainty visualizations in the news. We used a bottom-up iterative coding process in order to see what distinctions emerged from close analysis of the examples in our corpus. Two authors separately coded each article initially, listing for each a set of concrete questions about the visualized data that the article poses to the audience. We identified these questions by considering how the uncertainty visualizations support the narrative in the text of each article. Through several additional passes on the corpus, the two authors iteratively reviewed and refined the codes. We arrived at a set of yes-or-no distinctions used to label and characterize each question posed by the articles.

3.3 Results and Discussion

3.3.1 Conceptual Distinctions

The first three labels we coded characterize the *low-level visual tasks* implied by the text and the visual encodings. These codes answer the question: Is the relevant information from the uncertainty visualization obtainable through (1) direct reading of encoded values, (2) automatic visual ensemble processing, or (3) shape perception [19]? While direct reading is highly deliberate, ensemble processing is an automatic visual computation about sets of graphical encodings. Shape perception occurs when the contours of graphical elements (e.g., tops of bars, trend lines) are recognized as patterns which have a specific meaning to the observer. Visualizations can often be read in more than one way.

The second set of labels describe *what information is needed* to answer each question. Does the task primarily require the audience to consider (1) central tendency or (2) variance of a distribution of values? These labels are not mutually exclusive. However, to differentiate tasks that emphasize uncertainty from those that emphasize central tendency, we only assign a question with both codes if it is not possible to identify which aspect of the data dominates the task.

The final pair of labels concern the *high-level cognitive operations* which the observer performs with visually extracted information in order to address each question posed by the article. Do the concrete task and visual encoding require observers to make (1) comparisons or (2) inferences? The reader must compare values or distributions when the question posed by the article requires a relative judgment (e.g., differences of value or conditional likelihood). Inference occurs when an observer is asked to reason about an aspect of the data that is not directly encoded (e.g., the central tendency of a set of points).

3.3.2 Frequency of Codes

We identified a total of 87 questions that were implied in the 22 articles we analyzed. The information the reader needed to glean from the visualizations in order to answer these 87 questions could be read directly from the visualization in 60 (69.0%) cases, could be acquired

[1] https://www.nytimes.com/interactive/2017/11/28/upshot/what-the-tax-bill-would-look-like-for-25000-middle-class-families.html

through ensemble processing in 56 (64.4%) cases, and could be read through shape perception in 60 (69.0%) cases. This shows that designers often encode relevant information in multiple different ways (e.g., providing a trend line as well as individual data points).

We observed 19 (21.8%) questions that were primarily about central tendency and 34 (39.1%) questions that were primarily about uncertainty. Only 33 of the questions (37.9%) were concerned with both central tendency and uncertainty without one clearly dominating the task. This suggests that judgments about uncertainty in news contexts are often framed as separate tasks from judgments of central tendency. Empirical evidence in uncertainty visualization [36, 47, 52, 53] and cognitive science [39, 63] suggests that effective uncertainty visualization should promote the integration of both central tendency and uncertainty information because otherwise people tend to substitute biased heuristics for more nuanced judgments of uncertainty.

Of the 87 questions examined, 64 (73.6%) involved comparisons between individual data points and/or distributions, and 72 (82.8%) involved inferences about information not directly encoded in the visualizations. Both comparison and inference seem to be common cognitive operations required to interpret uncertainty visualizations in the news. This suggests that when designing and evaluating uncertainty visualizations for a broad audience, it is not sufficient to simply measure ability of observers to extract directly encoded individual values.

3.3.3 Selecting a Representative Uncertainty Judgment Task

We wanted a task for our experiments that required more sophisticated forms of judgment than simply reading directly encoded values. We also wanted to choose a task that required considering both uncertainty and central tendency. We used the results of our coding to identify a concrete question from a visualization and article originally presented by the NYT: "How not to be Misled by the Jobs Report" [55][2]. In our adaptation of this task (Fig. 3), participants decide which of two underlying trends, growth or no growth in the job market, is more likely to have produced an observed sample of job growth numbers. The choice between trends requires the participant to consider how variance due to sampling error impacts samples of hypothetical jobs numbers. The participant must visually extract the trend in a hypothetical sample of jobs numbers through either ensemble processing or shape perception. Then, considering the central tendency and variance of the two possible trends based on uncertainty visualizations, participants make inferences about the likelihood of a given sample under each trend. Thus, our task entails reasoning spanning all of our distinctions except for the direct reading of individually encoded values.

The task we selected is representative not only of how uncertainty visualization is used in the news but of tasks that are implicated in statistical literacy. By asking participants to make inferences from samples of noisy data, our task engages core statistical competencies. In order to understand applications of statistics in science, people must recognize that samples may not always resemble the model or population from which they are produced [21, 23]. In our task and in public-facing presentations of statistics (e.g., election models), there are components of variance in sampling which are accounted for by underlying trends (signal) and components of variance which are due to sampling error and other random processes (noise). The ability to recognize and parse these sources of variance is implicated in our task and is important to the comprehension of statistics [21, 23].

4 Experiment 1

Experiment 1 mirrors the visualization design and data-interpretation context of the NYT article "How not to be Misled by the Jobs Report" [38]. We evaluate how well HOPs facilitate accurate perceptual decision-making compared to a more conventional encoding of uncertainty— error bars. We choose error bars as the control visualization for this task because (1) they are a common static encoding of uncertainty and (2) they can be used to add uncertainty information to the bar encoding used in the NYT, enabling a controlled comparison that remains faithful to the original presentation.

[2]https://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html

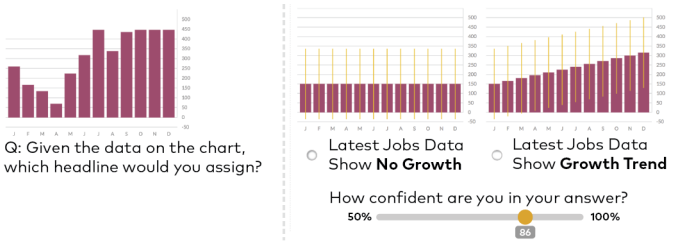

Fig. 3. A depiction of the task interface used in our studies (stimuli for Experiment 1 are shown). The chart that participants judge on the current trial is on the left side of the display. The reference uncertainty visualizations for the "no growth" and "growth" trends are on the right side of the display. Underneath the uncertainty visualizations, the participant uses radio buttons to indicate which trend is more likely and the slider to rate their confidence.

In a crowdsourced experiment on MTurk, we ask participants to discriminate which of two possible underlying trends is more likely to have produced hypothetical samples of jobs added to the economy each month of a simulated year (Fig. 3). Thus, we embed the visualizations from the NYT [38] into a traditional two-alternative forced choice (2AFC) psychophysics experiment.

4.1 Methods

4.1.1 Procedure

Upon accepting the Human Intelligence Task (HIT), participants were redirected to a web page containing instructions on the task. Participants were told they would play the role of a newspaper editor who is presented with a bar chart of jobs added to the economy each month of a simulated year and asked decide on a headline about the growth trend in the job market for that year (Fig. 3). On each trial, the participant made a 2AFC judgment ("growth" or "no growth") about one bar chart and provided a rating of their confidence in this judgment on a scale of 50 (random guess) to 100 (absolutely certain). Each participant completed a block of 60 trials using error bars as a decision aid and another block of 60 trials with HOPs as a decision aid, where the order of the visualization conditions was counterbalanced across observers. At the end of 120 trials, each participant completed a brief demographic survey, including questions about familiarity with statistics and with the specific visualizations shown in the task. The HIT carried a reward of $8 ($29.32 per hour on average).

4.1.2 Measures and Hypotheses

Our dependent measures are parameter estimates derived from psychometric functions (PFs, Fig. 5) and a related approach to modeling confidence data. These are inferential models of perceptual decision-making based on signal detection theory (SDT, Fig. 2).

Psychometric Functions (PFs): For each participant's 2AFC responses under each visualization condition, we fit a PF [42] estimating two parameters: (1) the JND, which is the level of evidence at which the participant is expected to perform with mean accuracy on the task; and (2) the spread of the PF, which describes the noise in the participant's perception of evidence in the task (Fig. 5). We predicted that when uncertainty due to sampling error was visualized using HOPs, as compared to error bars, subjects would have smaller estimated JNDs on average, indicating that observers require less evidence to distinguish the trend underlying a sample from noisy data. We also predicted that users would have smaller estimated PF spreads, indicating that they find stimuli ambiguous across a narrower range of evidence.

Confidence Fitness: Existing approaches to confidence analysis in uncertainty visualization evaluation tend to either assume that more confidence is better (e.g., [8]) or analyze the correlation between confidence and accuracy over a set of judgments (e.g., [14]). However,

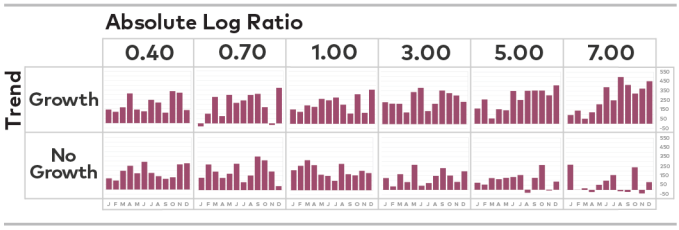

Fig. 4. Example stimuli judged by participants in our task. Scan across the figure to get a concrete sense of how our units of evidence translate into the appearance of a stimulus. Units of evidence are the log ratio of the probability that each stimulus was produced by the growth vs no growth trends, given shared variability due to sampling error. We take the absolute value of the log likelihood ratio so that units of evidence have the same sign regardless of which trend produced the stimulus.

both approaches may be inadequate to deal with the complexity of confidence as a construct; for example, research in judgment and decision-making describes how confidence reporting is often noisy and may not adhere to the laws of probability (i.e., [51]). To overcome these limitations, we adapt an approach from psychophysics [56] which explicitly models the noise in an observer's confidence reporting process.

Confidence fitness can be interpreted as an estimate of the degree to which confidence ratings are coherent with a probabilistic interpretation. The model is based on SDT and assumes that confidence ratings from the ideal observer should reflect the accuracy of their judgments at the given level of evidence [56]. To estimate a ground truth for confidence, we simulated many trials for each observer based on their PF. We estimate the ideal confidence of an observer as the accuracy of these simulated noisy judgments at each level of evidence. Confidence fitness is a latent parameter of the model ranging from 0 to 1 which estimates the degree to which actual confidence reporting is random or ideal. We had no strong *a priori* predictions about confidence fitness.

4.1.3 Stimuli and Trial Generation

Stimuli, Units of Evidence, and Task Difficulty: The PF fitting process requires a single measure of stimulus intensity which approximates the difficulty of the judgment task for each stimulus. Stimulus intensity is used as a ground truth to determine whether or not judgments are correct. In our task, the intensity measure should quantify the strength of the evidence in favor of a growth or no growth interpretation of a given sample of jobs numbers. We calculate the intensity based the probability of each possible trend having produced the sample of jobs numbers. For any given stimulus (set of hypothetical job numbers), strength of evidence is the log likelihood ratio describing the relative likelihood that the stimulus was produced by the no growth or growth trend in the job market (Fig. 4). This produces a log-linear intensity scale where positive values represent evidence in favor of no growth (e.g., bottom row of Fig. 4), negative values represent evidence in favor of growth (e.g., top row of Fig. 4), and zero is the point of maximum uncertainty. Because our measure of intensity should be consistent whether the evidence more strongly favors the growth or the no-growth trend, we take the absolute value of this log likelihood ratio.

$$evidence = |\log_{10}(Pr(sample|noGrowth)/Pr(sample|growth))|$$

Here, *sample* is a given set of job growth numbers, *noGrowth* is a trend in which the jobs added each month are normally distributed about 150k with a standard deviation (SD) of 95k, and *growth* is a trend in which there is a linear increase of 15k jobs per month from 150k in January to 350k in December, with a SD of 95k jobs each month (Fig. 3, right side). For both trends, we matched the mean job growth for each month to the NYT article [38]. However, we differed from the NYT article by using a SD of 95k jobs instead of 55k. We chose this SD to guarantee that there were many visually distinct stimuli to sample for which the underlying trend was ambiguous.

Staircase Sampling Procedure: A major challenge in obtaining valid PF fits is making sure that the observer completes enough trials at stimulus intensities which are ambiguous. Choosing stimuli which the participant can judge correctly, but which are not easy, reduces the uncertainty in the fitting process [57]. However, the researcher must not present too many trials, otherwise issues of participant attrition and fatigue arise. Adaptive sampling procedures are the best solution to this problem. These algorithms sample stimuli at levels of evidence which are ambiguous but not uninformative based on the participant's past performance. We used a three-down, one-up staircase (suggested in [22]) in which the level of evidence was incremented (i.e., made easier) by 3 absolute log likelihood ratio units each time the participant guessed wrong, and the level of evidence was decremented (i.e., made harder) every third time the participant was correct. In order to avoid autocorrelation in performance resulting from participants noticing the sampling procedure, we randomly interleaved 25 trials each from two different staircases (suggested in [13]) as well as 10 gold standard trials at an absolute log likelihood ratio of 9 (very easy). The two staircases differed only in their decrementing step sizes of 2.22 and 1.65 absolute log likelihood ratio units, respectively. Step sizes were chosen based on pilot data and recommendations from simulation studies on how to create staircases with stable convergence [22] and how to sample in order to minimize uncertainty in the parameter estimates from PFs [57]. These staircases promoted meaningful PF fits in a minimal number of trials.

Uncertainty Visualizations: In our task, participants use different uncertainty visualizations as a decision aid showing the no growth and growth trends (Fig. 3, right side). HOPs were generated by repeatedly sampling 12-month sets of jobs added to the economy from each underlying trend, plotting these numbers in bar charts, and animating transitions between frames. Animated transitions between bar values were 500 ms in duration with a 10 ms delay between each bar and a frame rate of 45 hz. Each sample was displayed for 1500 ms in between the animated transitions. Animation parameters followed those used by the NYT. We used a cubic easing function for animated transitions, which was consistent with the NYT visualization.

4.1.4 Participants

We recruited 62 MTurk Masters workers, each located in the United States and with a HIT approval greater than 95%, to participate in the study. A power analysis on pilot data suggested our target sample size should be 50 within-subjects comparisons of HOPs and error bars, where each participant completes a block of 60 trials for each visualization condition and the condition order is counterbalanced across participants. To achieve this, we iteratively sampled small batches of participants ($n < 10$), alternating the starting visualization condition, and applying a set of *a priori* exclusion criteria to the data collected. This was procedure was repeated until we had 50 participants in our sample after exclusions, and condition order was counterbalanced across participants. During the iterative sampling procedure (see Preregistration[3]), one participant was excluded due to exceptionally poor performance, and five participants were excluded from analysis due to poorly converging psychometric function fits. We accidentally collected six participants more than intended, so six participants were chosen at random and excluded from analysis thereafter.

4.1.5 Analysis Approach

Psychometric Function Estimation: We used a maximum likelihood optimization algorithm to fit a cumulative Gaussian distribution to model accuracy as a function of the level of evidence presented in each stimulus (Fig. 5). Recall that evidence is the degree to which the data in a stimulus informs the forced choice between the two possible underlying trends of growth and no growth. The lower asymptote of this cumulative Gaussian was set to a *guess rate* of 0.5 (Fig. 5, lower left), which would be achieved in theory if the participant guessed on each trial. The upper asymptote of the cumulative Gaussian was set to one minus some *lapse rate* (Fig. 5, upper right), the base rate of stimulus-independent errors due to lapses of attention. Following recommendations from a simulation study on PF fitting [65], we estimate

[3] https://osf.io/975us/register/5771ca429ad5a1020de2872e

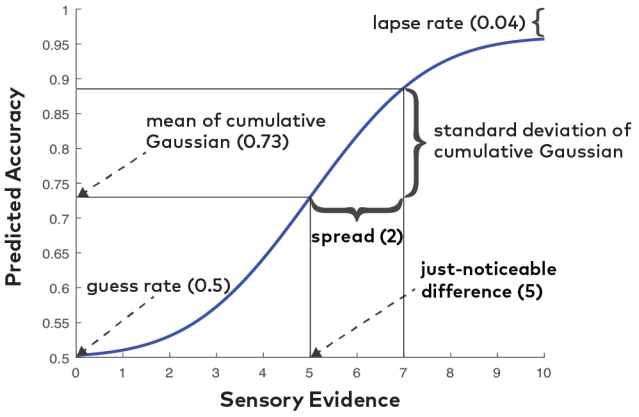

Fig. 5. An illustration of the psychometric function with relevant parameters.

lapse rate as a constrained free parameter between 0 and 0.06 in order to reduce bias in other parameter estimates. The level of evidence corresponding to the mean of the fitted cumulative Gaussian is the *JND* (Fig. 5, lower right). The standard deviation of the cumulative Gaussian is the *spread* parameter of the PF (Fig. 5, middle right). Per SDT (Fig. 2), the spread parameter represents noise in perception of the level of evidence, where PFs with greater spreads indicate that the trend in the jobs report is ambiguous across a larger range of evidence.

Confidence Fitness Estimation: We used maximum likelihood estimation to find an optimal value of confidence fitness for each PF. This algorithm was transcribed from Sanders et al. [56]. We provide a detailed description of the confidence fitness algorithm and our implementation in Supplemental Materials.

Statistical Inference: We used mixed-effects linear models in lme4 for R to estimate each response variable (i.e., JND, PF spread, and confidence fitness) as a function of fixed effects of visualization condition and condition order and a random intercept effect of participant. Further, we believed that an interaction between visualization condition and condition order would explain variance in parameter estimates due to practice or learning. Following the procedure in our preregistered analysis plan[4], for each measure we used ANOVA to compare model residuals with and without the interaction between visualization condition and condition order, and we chose the most parsimonious mixed-effects model which still minimized the variance of residuals. We provide analysis scripts and data in a project repository[5].

4.2 Results and Discussion

4.2.1 Just-Noticeable Differences (JNDs)

We modeled fixed effects for visualization condition and condition order and a random intercept term to account for individual differences, as described above. We did not include the interaction between visualization and starting condition in our model of JNDs because including the interaction did not reduce residual variance.

We found that JNDs were lower in the HOPs condition (Fig. 6, left) than in the error bars condition (Est = -0.68; 95% CI: -1.19 to -0.14; t = -2.49; p = 0.02). The second column of Figure 4 illustrates an effect of roughly this size. This effect suggested that on average when participants used HOPs rather than error bars they required less evidence about the underlying trend in a noisy time series to achieve mean accuracy (around 75%). Put differently, participants could successfully discriminate the underlying trend in the job market for more ambiguous stimuli when using HOPs as a decision aid.

4.2.2 Psychometric Function Spreads

We modeled visualization condition, condition order, and their interaction as fixed effects and a random intercept for individual differences. We found no reliable differences in PF spread estimates between visualization conditions (Fig. 6, middle). However, PF spread estimates were lower among participants who started in the HOPs condition (Est = -1.00; 95% CI: -1.91 to -0.09; t = -2.12; p = 0.04) compared to participants who started with error bars. Recall that narrower PF spreads indicate less noise in the perception of evidence. We speculate that participants may have developed mental representations for the the two possible growth trends during the first block of trials and relied on the uncertainty visualizations less in later trials. There was also an interaction between visualization condition and starting condition (Est = 1.46; 95% CI: 0.34 to 2.58; t = 2.54; p = 0.01), indicating a learning or practice effect. Regardless of the visualization condition under which data was collected, PF spread estimates tended to be larger in the first block of trials than in the second block. We speculate that participants became more sensitive to evidence in the task with practice.

Although we intentionally avoided the use of practice trials or feedback in order to test the behavior of untrained observers, it seems that practice or learning introduced differences across blocks which were not necessarily related to visualization condition. While the effect of starting condition on users' sensitivity to evidence could mean that HOPs facilitate superior learning about sampling error, this effect could also be explained as an artifact of testing visualizations within-subjects and keeping the task consistent across blocks. The ambiguity of this result points to considerations about experimental design in visualization evaluation. If the goal of the experiment is to measure learning, a longitudinal experiment which employs practice trials to control for practice effects is ideal. In studies like ours where the goal is to measure visualization effectiveness, future experiments in this paradigm should employ a between-subjects design, as we do in experiment 2, in order to mitigate ambiguity about which visualization condition is informing a user's mental representation of the task.

4.2.3 Confidence Fitness

We modeled fixed effects of visualization and condition order and a random intercept for individual differences. We did not model the interaction of visualization and starting condition. Although confidence fitness within individuals was seldom consistent across blocks (see Supplemental Materials), we found no reliable effect of visualization or condition order on confidence fitness (Fig. 6, right).

Since the results of the confidence fitness analysis were inconclusive and difficult to interpret, we conducted a more conventional *exploratory analysis* on confidence reporting. We used a mixed-effects linear model on trial-level data to estimate reported confidence based on fixed effects of visualization, starting condition, and their interaction as well as fixed effects of the level of evidence conveyed in each stimulus, whether or not a participant's answer was correct, and their interaction. We also included a random intercept effect of participant.

This model (Fig. 7) showed a small effect of visualization condition (Est = 1.62; 95% CI: 0.80 to 2.43; t = 3.90; p < 0.001) such that users report slightly higher confidence on average when using HOPs. There was also an interaction between visualization and condition order (Est = -1.33; 95% CI: -2.48 to -0.18; t = -2.28; p = 0.02). Simple effects showed that users were more confident on average when they used HOPs in the second block than in any other combination of visualization or block order. More than practice or learning, visualization condition seemed to play a role in confidence reporting, although these effects are of doubtful practical significance. Looking at trial-level predictors, we found that users were slightly more confident on average when they answered correctly (Est = 1.93; 95% CI: 0.33 to 3.52; t = 2.37; p = 0.02). We also found an interaction between correctness and the level of evidence presented in a stimulus (Est = 1.73; 95% CI: 1.36 to 2.09; t = 9.25; p < 0.001). Confidence increased with increasing evidence for trials where the user was correct, but confidence decreased with increasing evidence for trials where the user was incorrect. Sanders et al. [56] found that this interaction was predicted by the estimates of expected confidence produced by their confidence modeling. However, we failed to replicate this finding. The expected confidence ratings produced by our model of confidence fitness failed to predict reported confidence very well (see Supplemental Materials).

[4]https://osf.io/975us/register/5771ca429ad5a1020de2872e

[5]https://github.com/kalealex/jobs-report-hops

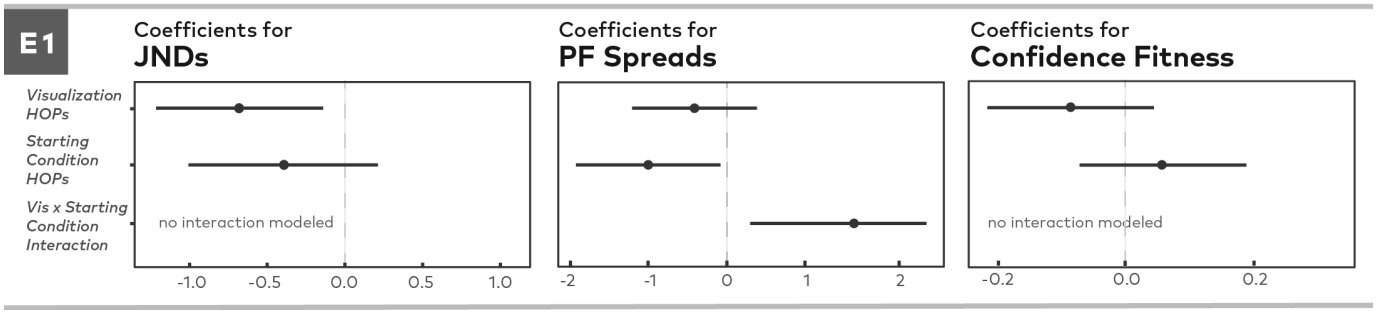

Fig. 6. Estimated regression coefficients for mixed-effects linear models of each psychophysical response variable in Experiment 1 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

Fig. 7. Estimated regression coefficients for our exploratory mixed-effects linear model of confidence reports in Experiment 1 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

5 Experiment 2

In experiment 1 we find that participants are more sensitive to the trend underlying samples from a noisy time series when using HOPs rather than error bars for a decision aid. However, in light of prior work demonstrating that people struggle to interpret error bars [6, 14, 31] and that bar encodings induce biased judgments of probability [50], this result is not entirely surprising. How might these results differ if we compare HOPs to a more effective static visualization than error bars and remove the influence of the within-the-bar bias from our visual encodings? Also, the differences between HOPs and error bars inferred from the results of experiment 1 conflate the effect of *animated vs static* encodings with the effect of *discrete sampling vs summary statistical* representations of probability. In experiment 2, our primary aim is to use the evaluation framework we developed to isolate the effect of animated vs static encodings of uncertainty. Thus, we use a similar study design to compare HOPs composed of lines (instead of bars) to static line ensembles (Fig. 1, middle right). We choose line ensembles as our control condition because they use a frequency representation of uncertainty which is superior to error bars, and they aggregate the discrete samples shown in HOPs into one static view.

5.1 Methods

With the exception of the following changes, the methods used in our second experiment were identical to the methods used in the first.

5.1.1 Uncertainty Visualizations

We tested three uncertainty visualization conditions in our second experiment. Two of these were HOPs with lines instead of bars (Fig. 1, right). We test two HOPs conditions to examine whether the benefits of HOPs persist at very high frame rates where the participant cannot cognize individual samples. The *fast HOPs* condition displayed each sample for only 100 ms in order to test the limits of visual ensemble processing. The *regular HOPs* condition displayed each sample for 400 ms, matching the presentation rate chosen by Hullman et al. [36]. Recall that HOPs in experiment 1 displayed each frame for 1500 ms, following the visualization design of the NYT. In contrast, the faster frame rate of the regular HOPs condition in experiment 2 reflects our *a priori* estimate of optimal animation parameters.

The static visualization condition in our second experiment was a *line ensemble* (Fig. 1, middle right). The line ensemble visualizations displayed the same samples as HOPs using the same line encoding, but lines were aggregated into one static view. We chose to use 50 lines per ensemble with representative sampling from each month in order to maintain the perception of a discrete representation of uncertainty. We did not match the lines per ensemble to the lines shown across all frames of the HOPs because we cannot know which lines users attended to in the HOPs condition. By comparing HOPs to line ensembles we investigate the impact of animating vs aggregating discrete sampling-oriented representations of uncertainty.

5.1.2 Experimental Design

We tested visualization conditions between-subjects, whereas visualization condition was a within-subjects comparison in Experiment 1. We changed to a between-subjects comparison in order to avoid ambiguity about which visualization condition was informing participants' mental representation of the two possible underlying trends. Just like the first experiment, each participant completed two blocks of 60 trials under the same sampling procedure as before. However, in Experiment 2 both blocks of trials were completed under one visualization condition. Thus, we sampled twice the number of trials per PF fit in the second experiment as we did in the first. We chose to sample 120 trials per PF in order to improve the quality of our JND and PF spread estimates and reduce the role of sampling error in our results.

5.1.3 Analysis Approach

We used the same maximum likelihood estimation procedures as in Experiment 1. For statistical inferences on JNDs, PF spreads, and confidence fitness, we used a linear model with visualization condition as a predictor. For Experiment 2, we preregistered[6] a secondary analysis of confidence reporting on each trial, similar to the exploratory analysis in our first experiment. We used a mixed-effects linear model of reported confidence with fixed effects of visualization condition, level of evidence in the stimulus, whether or not the participant answered correctly, and the interaction between evidence and correctness. We also included a random effect for individual differences.

5.2 Results and Discussion

5.2.1 Just-Noticeable Differences (JNDs)

We found that JNDs were lower in the regular HOPs condition (Fig. 8, left) than in the line ensembles condition (Est = -1.22; 95% CI: -2.18 to -0.26; t = -2.52; p = 0.01). This suggested that animating hypothetical outcomes across frames instead of aggregating them into one static view facilitated correct judgments of the underlying trend in samples

[6] https://osf.io/gw4cj/register/5771ca429ad5a1020de2872e

Fig. 8. Estimated regression coefficients for mixed-effects linear models of each psychophysical response variable in Experiment 2 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

Fig. 9. Estimated regression coefficients for our mixed-effects linear model of confidence reports in Experiment 2 (intercept coefficient omitted for space). Dots represent estimated effects, and lines represent 95% CIs.

conveying a lower level of evidence. JNDs were also lower for participants who used fast HOPs rather than line ensembles as a decision aid (Est = -0.74; 95% CI: -1.69 to 0.22; t = -1.52; p = 0.13). However, this effect was small and unreliable. When the frame rate of HOPs was faster, gains in performance were attenuated. Most individual JND estimates were in approximately the same range regardless of visualization condition. The effects we found were largely driven by a subset of individuals, mostly in the line ensembles condition, who showed little sensitivity on the task (JNDs > 9 absolute log likelihood units). It seems that animating hypothetical outcomes at a particular frame rate leads to more consistent levels of sensitivity across observers in judging samples from a noisy time series.

5.2.2 Psychometric Function Spreads

In line with the results from our first experiment, we found that visualization condition did not seem to impact PF spreads. On average, spread estimates were not reliably different for participants in the regular HOPs, fast HOPs, and line ensemble conditions (Fig. 8, middle).

5.2.3 Confidence Fitness

Confidence fitness was not reliably different across visualization conditions (Fig. 9, right). We speculate that animation of hypothetical outcomes, compared to a static depiction of the same outcomes, may have little distorting effect on the efficacy of confidence monitoring.

We ran a mixed-effects linear model on confidence reporting data (Fig. 9) similar to the exploratory analysis in Experiment 1. On average participants reported lower confidence in the fast HOPs condition than in the line ensembles condition (Est = -2.94; 95% CI: -5.70 to -0.17; t = -2.07; p = 0.04). Participants in the regular HOPs condition also reported lower confidence on average (Est = -2.34; 95% CI: -5.09 to 0.44; t = -1.64; p = 0.10), although this effect was not reliable. These results were in contrast with those of Experiment 1, in which participants reported slightly *higher* confidence on average when using HOPs rather than error bars as a decision aid. This apparent reversal of effect may have occurred because line ensembles establish a higher baseline for confidence than error bars, but it could also be noise.

Our findings regarding trial-level predictors were consistent with those of the first experiment. We found that participants were more confident on trials where they were correct (Est = 1.96; 95% CI: 1.04 to 2.88; t = 4.18; $p < 0.001$) and that there was an interaction between correctness and the level of evidence presented (Est = 1.34; 95% CI: 1.15 to 1.53; t = 13.77; $p < 0.001$). We also found a minuscule effect of the level of evidence in a stimulus on confidence (Est = 0.19; 95% CI: 0.01 to 0.37; t = 2.04; p = 0.04). Similar to our first experiment, our raw confidence data followed the predictions of Sanders et al. [56], who created the confidence fitness algorithm. Their model predicted that the relationship between confidence reporting and the level of evidence presented in a stimulus is modulated by correctness of participant responses. However, our implementation did not produce this predictive behavior (see Supplemental Materials).

6 Discussion and Future Work

6.1 Findings and Interpretation

In experiment 1, we find that users are more sensitive to the underlying trend in samples from a noisy time series when using HOPs rather than error bars as a decision aid. This suggests that sampling-oriented presentations of uncertainty lead to better comprehension of uncertainty for the purpose of perceptual decision-making than summary statistical representations of uncertainty. This finding is consistent with the literature on frequency presentations of probability [30, 24, 25, 32], sampling-oriented uncertainty visualizations [18, 36, 35, 40, 47], and the ensemble processing abilities of the human visual system [3, 4, 29, 33, 44]. Our first experiment extends this line of inquiry by (1) applying vision science methodology to uncertainty visualization evaluation and (2) addressing the question of whether HOPs can improve sensitivity to uncertainty information in the public domain.

In our second experiment, we isolate the effect of animation on user perceptions of probability. We compare HOPs, at two different frame rates, showing hypothetical sets of 12 outcomes as lines to static line ensembles, which show the same samples using the same line encoding but aggregated into one static display. We find that participants in the regular HOPs condition (400 ms per sample) are more consistently sensitive to the underlying trend in samples from a noisy time series than participants in the line ensembles condition. However, this effect is attenuated for fast HOPs (100 ms per sample). Perhaps, the ability of the visual system to automatically process ensembles breaks down beyond a certain frequency. Future work in information visualization and vision science should systematically test ensemble processing abilities across a more granular set of frame rates to try to pinpoint the limitations of the visual system in processing this kind of display.

Our analyses of confidence fitness in both experiments suggest that the efficacy of participants' confidence monitoring was not systematically impacted by visualization conditions. Confidence fitness is a unitless latent parameter of a model describing the degree to which reported confidence matches predictions from a Monte Carlo simulation based on signal detection theory. As such, it is difficult to interpret the practical significance of these null results.

6.2 Limitations

6.2.1 Confidence Fitness

Confidence as a construct is difficult to interpret. Confidence fitness contextualizes confidence reporting data in comparison to a theoretical ground truth, quantifying the degree to which confidence reporting fits a particular statistical definition. In principle, this tells us whether reported levels of confidence are warranted based on the presented evidence and the perceptual sensitivity of the user. However, we find that the algorithm does not consistently predict reported confidence, and this suggests that there may be wrong assumptions in the model (see Supplemental Materials). Future work should search for better-fitting models to establish a ground truth for confidence. Alternatively, perhaps a normative account of confidence, which assumes that confidence means the same thing to participants and researchers, is not appropriate considering the intersubjectivity of confidence reporting.

6.2.2 Psychometric Functions

A typical approach when using psychometric functions (PFs, Fig. 5) to model perceptual decision-making is to estimate JNDs and use statistical inference to detect shifts in the sensitivity of participants' perceptions under different conditions, quantified as differences in JNDs. In doing this, we fit a richly descriptive model but throw out all information other than a point estimate in the process of statistical inference.

However, we can leverage other aspects of the PF for inference. We did this by examining often ignored PF spreads, which represent noise in the perception of evidence. Judging by the null effects of visualization on PF spreads in this study, one might think they are an insensitive measure of visualization effectiveness. However, the change in PF spreads over time within an individual measures of how quickly people learn a task on a given interface. Since our experiments were not designed to measure such a learning effect, this remains a promising direction for future work. We also demonstrated how estimates of the PF spread can be leveraged, along with signal detection theory (Fig. 2), to bootstrap estimates of expected confidence reporting [56].

The core problem with the traditional use of JNDs is that PF fitting and statistical inference are relegated to separate computational models. In future work, we intend to rethink perceptual modeling using PFs and incorporate the entire procedure into a hierarchical Bayesian modeling approach [49].

6.3 Design Guidelines

Our findings imply that uncertainty visualizations in public-facing venues such as the news have a measurable impact on perceptions of uncertainty. Designers of uncertainty visualizations for public audiences have a responsibility to use uncertainty representations which promote accurate perceptions of probability.

HOPs and other sampling-oriented displays like line ensembles are particularly helpful when understanding the process that produced data is critical. In such cases, summary statistical encodings like error bars conceal important information about variability behind opaque statistical constructs like standard error [6, 14, 15, 31]. In contrast, showing discrete outcomes from the process which produced the data offers a more interpretable rendering of uncertainty [18, 24, 25, 32, 36, 35, 40], especially when viewers are unlikely to have statistical training.

The trade-off between static sampling-oriented visualizations and HOPs is mostly a question of how these techniques interact with the affordances of the visual system. In the case of static sampling-oriented visualizations, designers should consider whether displaying too many outcomes in one view will disrupt the perception of discrete outcomes, resulting in a density encoding rather than the intended frequency encoding. Density encodings for uncertainty are interpreted less consistently than frequency encodings because they describe probability in the abstract rather than showing users an experiential representation of uncertainty [36, 35, 40]. Displaying too many outcomes in one view also might lead to the problem of crowding, the inability of the visual system to generate accurate percepts of cluttered scenes [4]. In the case of HOPs, designers should consider whether the audience will allocate the necessary attention to integrate outcomes over time. Additionally, the designer should choose a frame rate for HOPs which is not so fast that the viewer cannot cognize individual observations and not so slow that the viewer becomes impatient. Interestingly, we and colleagues in our lab independently chose a similar frame rate to the one tested in prior work on HOPs [36], 400 ms per sample. Future work should examine a range of possible parameters and empirically establish the optimal design for HOPs.

Although communicating uncertainty is recommended in order to emphasize that outcomes may not always resemble their expected value [21, 23], in some cases uncertainty visualizations may not be well-matched to the intended task. If uncertainty is unlikely to aid interpretation (e.g., when reading a point estimate), a direct encoding of central tendency might be preferred. Prior work shows that HOPs are inferior to point estimates with error bars or violin plots if the task is to estimate central tendency of highly variable outcomes [36]. In some scenarios involving decision-making, communicating the uncertainty in the process which produced the data might introduce confusion, anxiety, or indecision [46]. However, failing to show uncertainty may lead to a false sense of security in visualized outcomes [5, 58]. Designers should consider the needs and background of their audience in order to determine how and whether or not to visualize uncertainty.

7 Conclusion

In our study, we demonstrate that animated sampling-oriented uncertainty visualizations can support perceptual decision-making in a realistic context. Since our task requires users to integrate information about the central tendency and variance of distributions and leverage this information to make inferences about the likelihood of samples, performance on this task is a proxy for the kind of reasoning which is necessary to understand the use of statistics in science [21, 23]. The degree to which different uncertainty visualizations support performance on this task should be considered an indication of their efficacy for facilitating nuanced statistical reasoning. Whether or not the gains in performance associated with HOPs on this task can translate into long-term improvements in statistical literacy is an open question. Future research should examine, in a longitudinal study, whether incidental exposure to various visual representations of uncertainty can facilitate learning of statistical reasoning. If it turns out that the benefits of HOPs for complex tasks like ours extend to lasting improvements in statistical reasoning, this would suggest that the incorporation of certain uncertainty visualizations into informal contexts like the news promotes a more statistically literate society.

8 Acknowledgements

The roles of the first and second authors were approximately equal. Alex Kale contributed domain knowledge about psychophysics and experimental design and was responsible for implementing the data analyses. Francis Nguyen contributed domain knowledge about system development and graphical design and was responsible for implementing the experimental interface. Jessica Hullman and Matthew Kay advised the project, contributing expertise in statistical inference and uncertainty visualization. Alex wrote the paper with iterative feedback from Jessica and editing help from other authors. Francis created the figures for the paper with guidance from other authors. Alex and Jessica conducted the qualitative analysis of visualizations in the news. Alex, Jessica, and Francis designed the experimental procedure. All authors selected visualization conditions to evaluate.

We would like to thank Geoffrey Boynton, Steve Franconeri, Michael-Paul Schallmo, and the members of the Interactive Data Lab for their feedback throughout the study.

References

[1] G. Aisch, N. Cohn, A. Cox, J. Katz, A. Pearce, and K. Quealy. Live Presidential Forecast, 2016.

[2] A. R. Albrecht and B. J. Scholl. Perceptually averaging in a continuous visual world extracting statistical summary representations over time. *Psychological Science*, 21(4):560–567, 2010.

[3] G. Alvarez and A. Oliva. The representation of ensemble visual features outside the focus of attention. *Psychological Science*, 19(4):392–398, 2008.

[4] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences*, 15(3):122–131, 2011.

[5] E. B. Andrade. Excessive confidence in visually-based estimates. *Organizational Behavior and Human Decision Processes*, 116(2):252–261, 2011.

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

[7] A. R. Binder, E. D. Hillback, and D. Brossard. Conflict or caveats? effects of media portrayals of scientific uncertainty on audience perceptions of new technologies. *Risk Analysis*, 36(4):831–846, 2016.

[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] M. Bostock, S. Carter, A. Cox, J. Daniel, J. Katz, and K. Quealy. Who Will Win The Senate? *The New York Times*, Apr. 2014.

[10] Q. Bui and B. Casselman. What the tax bill would look like for 25,000 middle-class families. *The New York Times*, Nov. 2017.

[11] S. M. Casner and J. H. Larkin. Cognitive efficiency considerations for good graphic design. *The Cognitive Science Society*, (June):275–282, 1989.

[12] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science*, 216(4550):1138–1141, 1982.

[13] T. N. Cornsweet. The Staircase-Method in Psychophysics. *The American Journal of Psychology*, 75(3):485–491, 1962.

[14] 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, Dec 2014.

[15] R. Cox. Representation construction, externalised cognition and individual differences. *Learning and instruction*, 9(4):343–363, 1999.

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

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

[18] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Conference on Human Factors in Computing Systems - CHI '18*, 2018.

[19] S. Franconeri. personal communication.

[20] S. G. *Digital Color Imaging Handbook*. CRC Press, 2002.

[21] I. Gal. Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, 70(1):1–25, 2002.

[22] M. A. García-Pérez. Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, 38(12):1861–1881, 1998.

[23] J. Garfield. The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3):58–69, 2002.

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

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

[26] M. F. Goodchild. Geographic information systems and science: today and tomorrow. *Procedia Earth and Planetary Science*, 1(1):1037–1043, 2009.

[27] D. Green and J. Swets. *Signal Detection Theory and Psychophysics*. John Wiley and Sons, 1966.

[28] J. Haberman and D. Whitney. Seeing the mean: ensemble coding for sets of faces. *Hum. Percept. Perform.*, 2009.

[29] L. Hasher and R. T. Zacks. Automatic processing of fundamental information: the case of frequency of occurrence. *The American psychologist*, 39(12):1372–1388, 1984.

[30] 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.

[31] 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.

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

[33] B. Hubert-Wallander and G. M. Boynton. Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, 15(2015):1–12, 2015.

[34] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2213–2222, 2011.

[35] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall estimation of effect uncertainty. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2018.

[36] 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.

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

[38] N. Irwin and K. Quealy. How Not to Be Misled by the Jobs Report. *The New York Times*, May 2014.

[39] 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):126–140, 2012.

[40] M. Kay, T. Kola, J. Hullman, and S. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems*, CHI '16, 2016.

[41] M. Kay, T. Kola, J. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proc. CHI 2016*, 2016.

[42] F. A. Kingdom and N. Prins. *Psychophysics: A Practical Introduction*. Elsevier Ltd., first edition edition, 2010.

[43] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[44] A. Y. Leib, A. Kosovicheva, and D. Whitney. Fast ensemble representations for abstract visual impressions. *Nature Communications*, 7:1–10, 2016.

[45] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum (Proc. EuroVis)*, 2013.

[46] R. Lipshitz and O. Strauss. Coping with Uncertainty: A Naturalistic Decision-Making Analysis. *Organizational Behavior and Human Decision Processes*, 69(2):149–163, 1997.

[47] L. Liu, A. Boone, I. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. Uncertainty Visualization by Representative Sampling from Prediction Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2165 – 2178, 2016.

[48] 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.

[49] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2016.

[50] 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.

[51] 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.

[52] 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):37–46, 2015.

[53] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1):40, 2017.

[54] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In A. M. Dienstfrey and R. F. Boisvert, editors, *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[55] K. Quealy and A. Cox. The first g.o.p. debate: Who's in, who's out and

the role of chance. *The New York Times*, July 2015.

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

[57] Y. Shen and V. M. Richards. A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *The Journal of the Acoustical Society of America*, 132(2):957–967, 2012.

[58] E. Soyer and R. M. Hogarth. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3):712–714, 2012.

[59] D. A. Szafir. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2017.

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

[61] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Conn., 1983.

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

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

[64] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *Int. J. Human-Computer Studies*, 57:247–262, 2002.

[65] F. A. Wichmann and N. J. Hill. The psychometric function : I . Fitting , sampling , and goodness of fit. 63(8):1293–1313, 2001.

[66] J. Zhang and D. A. Norman. Representations in distributed cognitive tasks. *Cognitive Science*, 18(1):87–122.