How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results

Jake M. Hofman
Microsoft Research
jmh@microsoft.com

Daniel G. Goldstein
Microsoft Research
dgg@microsoft.com

Jessica Hullman
Northwestern University
jhullman@northwestern.edu

ABSTRACT

When presenting visualizations of experimental results, scientists often choose to display either inferential uncertainty (e.g., uncertainty in the estimate of a population mean) or outcome uncertainty (e.g., variation of outcomes around that mean) about their estimates. How does this choice impact readers' beliefs about the size of treatment effects? We investigate this question in two experiments comparing 95% confidence intervals (means and standard errors) to 95% prediction intervals (means and standard deviations). The first experiment finds that participants are willing to pay more for and overestimate the effect of a treatment when shown confidence intervals relative to prediction intervals. The second experiment evaluates how alternative visualizations compare to standard visualizations for different effect sizes. We find that axis rescaling reduces error, but not as well as prediction intervals or animated hypothetical outcome plots (HOPs), and that depicting inferential uncertainty causes participants to underestimate variability in individual outcomes.

Author Keywords

Uncertainty visualization, effect size, judgment and decision making, confidence interval, prediction interval.

INTRODUCTION

Scientists are often faced with the challenge of conveying uncertainty to their audiences. Broadly speaking, this information can be thought of as belonging to one of two categories: *inferential uncertainty* or *outcome uncertainty*. By inferential uncertainty, we mean the degree to which a particular summary statistic (e.g., a population mean) is known to the scientist. Outcome uncertainty, in contrast, captures how much individual outcomes vary (e.g., around the mean, regardless of how well it has been estimated).

A distinguishing feature between the two is that inferential uncertainty can be reduced by collecting and analyzing more data, whereas outcome uncertainty cannot. For example, the *standard deviation* σ is a population parameter that quantifies outcome uncertainty, whereas inferential uncertainty is measured using *standard error*, calculated as $se = \frac{\sigma}{\sqrt{n}}$ where n is the sample size. Standard errors can be made arbitrarily small by increasing n, but this does not change the standard deviation σ.

For instance, in a study of how heights vary by gender, inferential uncertainty captures how precisely one has estimated the average height of men and women based on the set of measurements that were made. With a large enough sample, one can very precisely estimate the average height within each group. But this does not change the fact that there is a good deal of outcome uncertainty, as individual heights within each group vary substantially around their respective averages.

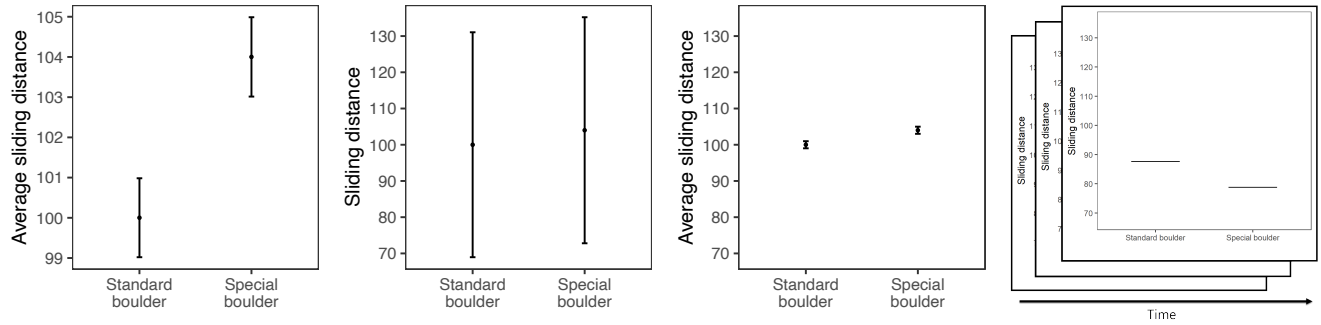
This choice of emphasizing either inferential uncertainty or outcome uncertainty extends to the visualizations that scientists present to their readers. In particular, the distinction between showing inferential uncertainty or outcome uncertainty produces graphical depictions that look quite different from one another when plotted in numerous graphics libraries (see Figures 1a and 1b, which were generated in R). Visualizations of inferential uncertainty facilitate comparing the sampling distribution of the mean in one group to that of another, as shown in Figure 1a. A common chart of this type will display the mean of each group with error bars that extend 1.96 *standard errors* above and below the mean to create a 95% confidence interval (95% CI), though error bars of one standard error are also common. Such visualizations convey uncertainty in estimating the means of each group, and facilitate null hypothesis significance testing (NHST) and "inference by eye" using various heuristics [14].

On the other hand, visualizations of outcome uncertainty emphasize the distribution of individual outcomes, such as the 95% prediction intervals (95% PIs) depicted in Figure 1b. A 95% PI displays group means accompanied by error bars that extend 1.96 *standard deviations* above and below the mean. Error bars of one standard deviation are also common. Visualizations of this type facilitate estimation of standardized effect sizes that account for both variance as well as mean differences. For instance, Cohen's d is a common measure of effect size that normalizes simple mean differences by the (pooled) standard deviation in outcomes: $\frac{\mu_1 - \mu_2}{\sigma}$. One can almost directly estimate Cohen's d from Figure 1b.

Most prior work in uncertainty visualization has sought to determine what visualization technique best facilitates comprehension, assuming *either* outcome or inferential uncertainty is being visualized. In contrast, in this paper we investigate how the type of uncertainty that an author chooses to visualize

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
https://dx.doi.org/10.1145/3313831.3376454



(a) 95% confidence intervals (b) 95% prediction intervals (c) Rescaled 95% confidence intervals (d) Animated samples (HOPs)

Figure 1: The visualization formats we tested in our experiments. All four present information about the same underlying data, but emphasize different aspects of it. Our first experiment compared two conventional visualizations: (a), which focuses on *inferential* uncertainty (uncertainty in predicting the mean), to (b), which presents information about *outcome* uncertainty (variation in individual outcomes). Our second experiment included all four visualizations, adding (c) as a hybrid that contains information about both inferential and outcome uncertainty and (d), which presents animated samples of individual outcomes. We find that (a) and (c) lead people to overestimate the effectiveness of treatments compared to (b) and (d).

affects people's beliefs about the effectiveness of a treatment. To this end, we contribute two large pre-registered randomized experiments in which we show people different visualizations of the same underlying data—some of which focus on inferential uncertainty and others on outcome uncertainty—and query beliefs about the size of treatment effects.

We measure participants' judgments using three proxies that vary predictably with the treatment effects: *willingness to pay for a treatment*, *probability of superiority* in which we ask participants how often they believe they would win a competition with or without the treatment, and *belief distributions*, that is, subjective probability distributions describing outcomes in the treatment and control groups, elicited with a graphical tool called the Distribution Builder [21, 22]. We focus on these measures because they are important for individual-level decision making, where one is concerned with their own individual outcome as opposed to the average outcome over a larger group, as might be the case for policy makers.

Across all measures, we find that visualizations oriented towards inferential uncertainty (95% CIs) led people to overstate willingness to pay by the greatest amount, overstate probability of superiority by the greatest amount, and understate the variability in outcomes by the greatest amount. Presenting 95% CIs on rescaled axes improved answers somewhat, but responses closest to the normatively correct answers across all measures were attained by people presented with visualizations that encoded information about variation in individual outcomes (95% PIs or hypothetical outcome plots (HOPs)).

VISUALLY COMMUNICATING EXPERIMENT RESULTS

An error bar is perhaps the most recognized way of visualizing a range of possible values that a statistic might take. One commonly cited problem with error bars is that they are used to convey multiple statistical constructs, such as standard error, standard deviation, and confidence intervals. Conventions concerning the use of error bars vary between and even within fields. For example, a review of articles published in Nature Methods in 2012 noted that 49% of error bar uses were to convey standard error, 42% were used to convey standard deviation, 5% did not specify what they represented, and only 1 occurrence conveyed a 95% confidence interval [33]. Other authors publishing in a human computer interaction (HCI) venue, however, suggest that 95% CIs are the most common type of interval estimate [6]. Further, 95% CIs are presented as an exemplary display of effect size in proposed guidelines for transparent statistical reporting in HCI [37]. Statistical reformers and psychological standards boards have also advocated for 95% CIs as a more intuitive way of expressing uncertainty [2, 12, 13, 38].

Error bars have also been thought to invite confusion due to the heuristics people apply to them. For example, readers may correctly assume that non-overlap of two error bars on independent means implies a statistically significant difference but incorrectly assume that overlap implies no significant difference [4, 14, 35]. Even scientists can wrongly apply heuristics like judging overlap. A study that asked psychology, neuroscience, and medical researchers to position error bars representing standard errors or a 95% CI found that the majority did not understand the relationship between overlap and statistical significance nor the importance of the study design [4], while a similar study which posed true/false statements about CIs to researchers and graduate students (none of which were true) found that on average, respondents endorsed more than half of them [23]. Other misinterpretations include the belief that error bars display a region of uniform probability [27] and a "within-the-bar bias" that occurs when error bars are superimposed on bar charts [11, 34]. Finally, separating the visual marks encoding underlying data from those encoding uncertainty may encourage users to underweight probabilistic information in favor of measures of central tendency [26].

Despite the wealth of knowledge around biases in reading error bars, surprisingly little work has directly compared the use of error bars to show different statistical constructs associated with effect size reporting. However, lay people,

students, and even researchers are often confused by the critical difference between a sampling distribution, which describes variance in a statistic like a mean, and a population distribution, which describes variance in individual measurements [4, 9, 23]. For example, a recent crowdsourced study found that when presented with sample statistics from an experiment, laypeople greatly overestimated variance in a sampling distribution, though interactive visualizations that enabled them to compare their expectations to the inferred distribution could reduce this bias [25]. In contrast to the many prior studies that have examined the effects of showing only one of these two types of distributions, our work directly compares them to each other. Specifically, we compare how emphasizing the sampling distribution (95% CIs) as opposed to the population distribution (95% PIs) impacts judgments, decisions, and perceptions about the size of an effect.

Alternatives to error bars have been proposed and evaluated. Intrinsic visual encodings of variability use a single retinal variable to convey a distribution without adding additional marks. For example, density plots use height to convey the probability of values, such that the mode appears as the highest point, while violin plots encode probability as width [3, 30] and gradient plots encode probability using opacity [28]. Other options are based on evidence from research in judgment and decision-making [20, 24] that suggests that presenting probabilities using a frequency framing (e.g., 1 out of 10), as opposed to a probability framing (e.g., 10%), can improve judgments. Experimental tasks used to establish an advantage to frequency-based visualizations have targeted Bayesian reasoning [32], risk assessment in a health context [18], reporting of subjective probability distributions [25], probability estimates made from visualizations [26, 31], recall for a visualized distribution [25], judgments of which model generated a data sample [29] and incentivized decisions [17].

Drawing on the human capacity for frequency encoding, HOPs—animated visualizations in which each frame depicts a random draw from a distribution—have been shown to lead to more accurate estimates of probability of superiority than violin plots and error bars [26], and more accurate judgments of which model produced data samples than error bars or static ensembles [29]. HOPs directly encode probability of superiority, making them useful regardless of dependencies between variables, unlike most conventional plots for presenting distributions [26]. While far from being mainstream, researchers have proposed using HOPs in scientific papers, both in Portable Document Format [26] and in online, interactive versions of scientific papers [16]. In Experiment 2, we compare judgments made from HOPs to those made from error bars conveying population and sampling distribution constructs to gain a better understanding of whether directly encoding probability of superiority can help curb biases in estimates of treatment effectiveness.

EXPERIMENT 1

We designed our first experiment to measure differences in how people perceive the same effect when it is communicated in one of two conventional formats: a) means and standard errors, which focus on uncertainty in measuring the mean and b) means and standard deviations, which present information about variation in individual outcomes around the mean. Specifically, we compare error bars that show 95% CIs (using 1.96 *standard errors*), as in Figure 1a, to error bars that show 95% PIs (using 1.96 *standard deviations*), as in Figure 1b.

In theory these presentations contain the same information about the underlying data so long as one knows the sample size, but the former focuses on inferential uncertainty whereas the latter emphasizes outcome uncertainty. We are interested in how people's perceptions of treatment effectiveness compare when they are shown only one of these two graphical formats. In particular, estimating the sample standard deviation from visualizations like Figure 1a requires the reader to inflate the error bars by the (square root of the) sample size. As this seems difficult for even sophisticated readers, we imagined that people would perceive different effect sizes for visualizations shown in format a) compared to format b).

We also used this experiment to investigate whether differences between visualization formats can be mitigated by simply adding extra text in the captions that appear alongside a given figure. For instance, it is not uncommon for a scientific paper to summarize both sample statistics as well as inferential statistics in text, but to visualize only one of these. Perhaps simply adding information about 95% PIs in the caption for a plot showing 95% CIs changes the inferences readers make about the distribution of outcomes under the treatment while still communicating information about inferential uncertainty.

Before running the experiment, we formulated and pre-registered[1] the following hypotheses:

H1. **Willingness to pay.** Participants who are shown visualizations with 95% CIs will exhibit different willingness to pay for the same treatment compared to those who are shown 95% PIs.

H2. **Probability of superiority.** Participants who are shown visualizations with 95% CIs will report different estimates for the probability that undergoing the treatment provides a benefit over a control condition.

We examined these hypotheses for two caption alternatives: when the figure caption matched the information in the visualization, and when the figure caption contained extra information beyond what is directly shown in the visualization.

Experimental Design

To test these hypotheses, we created a scenario that we deemed would be easily understandable by laypeople and representative of many situations where one uses analytical results to decide whether to make an investment or take a precaution. We presented the information in this scenario in different formats, and measured differences in how people perceived the effect of the treatment. Specifically, we told participants that they were athletes competing in a boulder sliding game, playing against an equally skilled competitor named Blorg. The goal of the game is to slide a boulder on ice farther than the opponent's boulder. There is an all-or-nothing 250 Ice Dollar prize for the

[1] Pre-registrations, data, and analysis code for both experiments are available at https://osf.io/rcfy5/.

contestant who slides their boulder the farthest. Participants were given the opportunity to rent a superior boulder (i.e., undergo a treatment) that is expected (but not guaranteed) to increase their sliding distance for their next and final competition. They were then shown a visualization that provided statistics about both the standard and special boulders with an accompanying caption. Finally, they were asked how much they were willing to pay for the special boulder and to estimate the probability of winning if they used it.

We manipulated the information shown to participants in a 2 x 2 design that varied both the visualization and text that they saw. Participants were randomly assigned to see a figure with either 95% CIs or 95% PIs, and were independently randomly assigned to see an accompanying caption that was either specific to the visualization they were shown or that contained extra information beyond what was directly presented in the visualization. Crossing these two levels of visualization and accompanying explanation created the four between-subjects conditions in our experiment. Screenshots of all conditions are provided in the supplemental material.

Data & Stimuli

We constructed the stimuli in our experiment to correspond to a Cohen's d of approximately 0.25 and a probability of superiority of 57%, typical of the effect sizes observed in fields such as psychology, neuroscience, and medicine [5, 7, 8]. We achieved this using the following parameters for the standard and special boulder: slides from the standard boulder were normally distributed with a mean of 100 meters and a standard deviation of 15.3 meters, whereas slides from the special boulder had the same standard deviation but a mean of 104 meters. We chose a standard deviation of 15.3 meters so that the 95% PI, derived from 1.96 standard deviations above and below the mean, would span an easily readable, round number range of 70 to 130 meters. We then took 1,000 samples from each of these distributions and used them to compute 95% CIs on the mean and 95% PIs on individual outcomes for each of the boulders. We plotted the results and matched the number of tick marks on the vertical axis, as shown in Figures 1a and 1b. We eschewed bar charts to prevent "within-the-bar bias" [11]. We created captions to explain the mean, 95% PI, and 95% CI, and phrased each of these in terms of what would happen during 1,000 potential future slides of each boulder.

Participants

We recruited 2,400 participants from Amazon's Mechanical Turk to take part in the experiment. We chose this sample size based on the results of a previous pilot, so that we had approximately 80% power in detecting differences between conditions at a 5% significance level. Removing the 49 participants who participated in both the previous pilot and this experiment left us with 2,351 participants. All participants were located in the U.S. with Mechanical Turk approval ratings of 97% or above. Participants were randomly assigned to one of the four conditions (95% CI w/ matching text, n = 569; 95% CI w/ extra text, n = 584; 95% PI w/ matching text, n = 647; 95% PI w/ extra text, n = 551). Each participant received a flat payment of $0.75.

Procedure

Participants were first presented with the rules of the game described above, and told that they had the option to rent the special boulder for a one-time use. The instructions indicated that it was within the rules of the game for them to rent the special boulder, and that they could be assured that their opponent would not have access to it and would instead use a standard boulder. On the next screen they were shown text with statistics about the standard boulder as per the condition they were randomly assigned to. They were also told that there was no reason to believe that they would have an advantage over Blorg (or vice versa) if they chose the standard boulder. Both screens required that they checked a box to confirm that they understood the information presented to them.

On the third screen they saw text with statistics about the special boulder, along with a visualization that summarized the information about the standard and special boulder, as per the condition they had been assigned to. Below this they were asked for their willingness to pay for the special boulder. Participants responded by moving a slider that ranged from 0 to 250 Ice Dollars and was initialized at the 250 Ice Dollar mark to encourage people to respond with the most they were willing to pay. They were required to move the slider from its default value before they could submit a response.

After they submitted their response to this question, participants were asked to estimate the probability of superiority for both the standard and special boulders. Estimating the probability of superiority for the standard boulder served as an attention check, as it should have been clear from the previous screens that the probability of winning with the standard boulder was 50%. Participants responded to each question by typing a whole number between 0 and 100, inclusive, for each question. This concluded the experiment.

Results

Having collected responses from all participants, we conducted the analyses specified in our pre-registration plan. We first removed participants who failed the attention check, indicated by an answer other than 50 out of 100 to the question about the probability of winning with the standard boulder.[2] This left us with 1,743 participants.

H1. **Willingness to pay.** We computed the mean willingness to pay for the special boulder in each of the four conditions along with the standard deviation in willingness to pay and the corresponding standard errors. The results are shown in Figure 2. The left facet corresponds to the standard setting where the captions matched the information in the visualization people were shown (e.g., those shown the 95% CI visualization saw text about 95% CIs), whereas the right shows results when the text contained extra information (e.g., information about 95% CIs and 95% PIs was shown in captions regardless of the visualization that was shown). In the left facet, we see that participants were willing to pay substantially more on average (80 Ice Dollars compared to 50) for the treatment when the visualization emphasized inferential uncertainty (95% CIs)

[2]Repeating the analysis below and including these participants yields similar results.

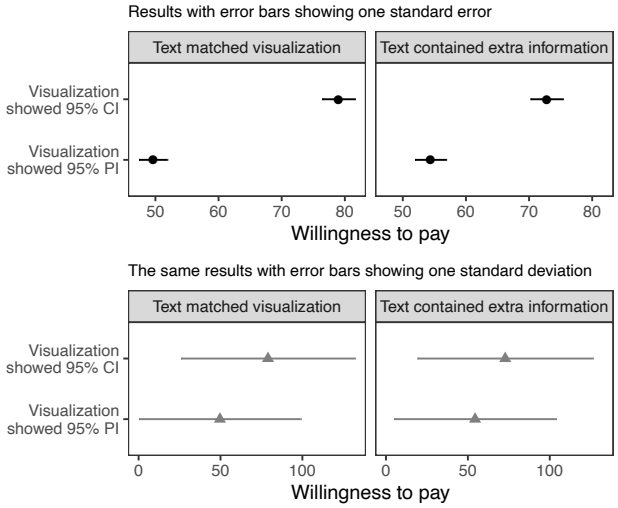


Figure 2: Results from our first experiment, visualized two different ways. In the top row, the black circles show mean willingness to pay for the treatment with error bars that correspond to one standard error in estimating the mean. In the bottom row, the grey triangles also show mean willingness to pay for the treatment, but have error bars that show one standard deviation in individual responses. Participants were willing to pay more on average for the treatment when the visualization emphasized inferential uncertainty (95% CIs) compared to outcome uncertainty (95% PIs). This difference persists even when information about both 95% CIs and 95% PIs is presented in the text accompanying the visualizations (right panels). Based on these results, we expect readers who are shown the plot in the top row to perceive this effect to be larger than readers who are shown the plot in the bottom row.

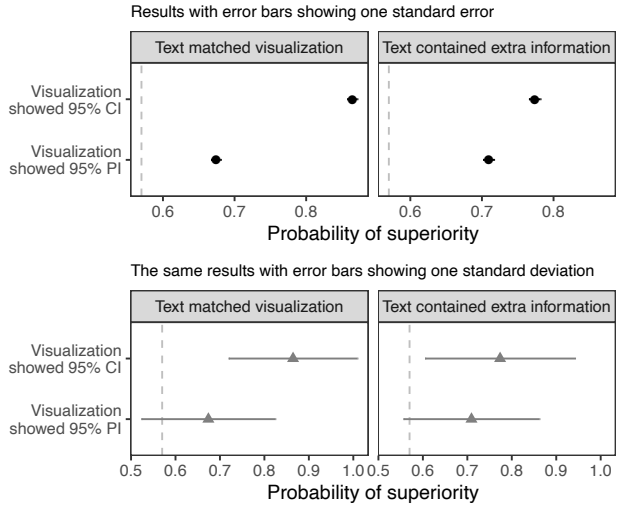


Figure 3: A summary of the estimated probability of superiority for the treatment in all four conditions of our first experiment. On average, participants estimated the treatment was more likely to be effective over a control condition when the visualization emphasized statistical significance (95% CIs) compared to effect sizes (95% PIs).

compared to outcome uncertainty (95% PIs). On the right we see that this effect persists even when participants are given extra information in the text alongside each figure, although the difference between groups is somewhat smaller (72 versus 54 Ice Dollars).

In the spirit of our experiment, we present visualizations of these results with two different types of error bars in the top and bottom rows of Figure 2: in the top row the black circles show one standard error in estimating the mean willingness to pay, whereas error bars in the grey triangles on the bottom row show one standard deviation in individual responses. The former most directly communicates that our results are statistically significant (two-sided t-tests: $t(861) = -8.57), p < 0.001$ for matching text, $t(836) = -5.17, p < 0.001$ for extra text), whereas the latter affords the reader an estimate of the standardized effect size we observed (Cohen's d = 0.57 for matching text, Cohen's d = 0.36 for extra text).

These are reasonably large effect sizes, but based on the results of our experiment, we expect that readers of this article would estimate them to be even larger if they were shown only the more conventional black circle error bars in the top row instead of the grey triangle error bars in the bottom. And while we could have chosen to visualize our results using only the figure in the bottom row, we show both formats here and throughout the rest of the paper for two reasons. First, showing both versions underscores the point that how we visually present uncertainty in results can change the way readers perceive reported effects. Second, showing the bottom row alone could make it more difficult to assess whether observed differences in mean willingness to pay are systematic or specific to the responses we analyzed, at least without further information about the number of participants involved in the experiment. We discuss this further in Section 5.

In all conditions participants were willing to pay substantially more for the special boulder than would be expected if they were risk-neutral and simply maximizing their expected payoff. The risk-neutral willingness to pay is 17.5 Ice Dollars, calculated as the difference between the expected value of winning when competing with the special boulder ($250 \times .57$) as compared to competing without it ($250 \times .50$). Taking expected value maximization as a norm, the PI group was closer to the normative answer by around 20 to 30 Ice Dollars.

H2. **Probability of superiority.** We find similar results for participants' estimates of the probability of winning if they use the special boulder when playing an opponent who uses the standard boulder, with an even larger difference between conditions than above. As shown in the left facet of Figure 3, participants who saw 95% CIs thought that they would win 86% of the time that they used the special boulder on average, whereas those who were shown 95% PIs reported an average probability of superiority of 67%. These are both higher than the true value of 57%, but the latter is much closer than the former, which implies near certainty in winning with the special boulder. These differences are statistically significant (two-sided t-tests: $t(892) = -19.38, p < 0.001$ for matching

text, $t(836) = -5.80, p < 0.001$ for extra text) with relatively large effect sizes (Cohen's d = 1.29 for matching text, Cohen's d = 0.40 for extra text).

This experiment makes clear that the type of error bars that authors use to visualize uncertainty in their results can have a substantial impact on how effective readers perceive interventions to be. Specifically, focusing on inferential uncertainty by showing standard errors in visualizations leads laypeople to overestimate the effectiveness of treatments compared to showing standard deviations, at least in the setting we have studied here. In our next experiment we extend these results to include other visualizations and an additional effect size, and collect further information on how readers perceive the distributions depicted by these visualizations.

EXPERIMENT 2

Our second experiment serves several purposes. First, having established that conventional representations using either standard errors (as in Figure 1a) or standard deviations (as in Figure 1b) both have some shortcomings, we use Experiment 2 to explore two alternative visualizations. The first is shown in Figure 1c, where error bars depict 95% CIs, but the vertical axis has been rescaled to accommodate 95% PIs. This relatively simple modification is similar to intentionally extending the y-axis range to include zero—as visualization experts sometimes advise—but has the benefit of conveying information about both inferential uncertainty, through the error bars, and outcome uncertainty, through the axis range.

The second alternative visualization we consider in Experiment 2 is HOPs, which show uncertainty through a series of animated frames that depict samples from underlying distributions [26]. HOPs are an attractive candidate for our task because they encode the probability of superiority directly via frequency, allowing readers to estimate the probability of superiority by counting (or estimating using ensemble processing [1]) the fraction of frames that samples from one distribution dominate samples from another. In this sense they are optimized for our task, providing a useful comparison point. While HOPs can be used to show samples from any distribution, here we use HOPs for only individual outcomes and investigate how participants perceive effect sizes when shown HOPs compared to other formats. A static representation of frames from one of the HOPs used for this experiment is shown in Figure 1d.

We also used this experiment to elicit alternative and more fine-grained measurements of the distributions and effect sizes that participants perceived from different visualizations. Specifically, after participants saw a visualization, we asked them to use the Distribution Builder[3] tool [21, 22] to specify a full distribution of outcomes they could expect under each condition. This allowed us to impute alternative measurements for responses collected in our previous experiment and offered more fine-grained information than can be captured by these responses alone. Take, for instance, the probability of superiority measurements from our previous experiment, where participants were asked for an integer between 0 and 100 for the number of times they expected to outperform their opponent if they used the special boulder. Many participants responded with numbers such as 75, which could be an artifact of people being anchored on certain salient numbers (e.g., multiples of 5). Using the Distribution Builder offered an alternative means of eliciting this information that is unlikely to be subject to such effects. It also allowed us to measure people's subjective beliefs about how much variation there was in individual outcomes by computing the standard deviation of the distributions they provided.

Finally, we included two different effect sizes in this experiment as a robustness check. We considered the same "small" effect size as in the previous experiment, but added a "large" [15] effect size corresponding to a Cohen's d of 1.0. This allowed us to check if our previous results were specific to the stimuli we chose or to round number or ceiling effects. For instance, it could be the case that when asked for a probability of winning in an uncertain setting, people gravitate towards certain salient numbers (e.g., "95%") instead of using the full range of possible values.

Before running this experiment, we formulated and preregistered the following hypotheses:

H3. **Willingness to pay.** Participants will exhibit different willingness to pay for the same treatment based on the visualization they see.

H4. **Stated probability of superiority.** Participants will report different estimates for the probability that undergoing the treatment provides a benefit over a control condition based on the visualization they see.

H5. **Implied standard deviation.** Participants will report distributions with different standard deviations based on the visualization they see.

H6. **Implied probability of superiority.** Participants will report distributions that imply different estimates for the probability that undergoing the treatment provides a benefit over a control condition based on the visualization they see.

We tested each of these hypotheses at two levels of granularity. First, we compared visualizations that emphasize statistical significance (Figure 1a and Figure 1c) to those that focus on effect sizes (Figure 1b and Figure 1d). Second, we conducted a more fine-grained analysis which compared specific pairs of visualizations to each other: 95% CIs to 95% CIs with a rescaled axis, 95% CIs with a rescaled axis to 95% PIs, 95% CIs with a rescaled axis to animated samples (HOPs), and HOPs to 95% PIs. We implemented all of these tests as two-way ANOVAs with planned comparisons.

Experimental Design

We used a 2 x 4 between-subjects design that varied whether people were presented with a small or large effect size and which of the four visualizations they saw. Participants were randomly assigned to an effect size corresponding to either a Cohen's d of approximately 0.25 or 1.0 and independently randomly assigned to see one of the four visualizations in Figure 1 (95% CIs, 95% PIs, 95% CI rescaled, or HOPs). This

[3] https://quentinandre.github.io/DistributionBuilder/

created the eight conditions in our experiment. In contrast to our previous experiment, we did not manipulate the captions alongside each visualization, but instead followed the convention that the text next to a figure matches the information shown in the figure.

Data & Stimuli

We constructed stimuli following the same procedure as in the previous experiment, but expanded this to include the alternative visualizations and the large effect size. We re-used data from the previous experiment for the small effect size and added the two alternative visualizations. The 95% CI visualization was rendered with a vertical axis that matched the 95% PI visualization in both its range and the placement and labeling of tick marks. The HOPs visualization showed 940 different frames at a rate of 2.5 frames per second, where each frame contained a random sample from the special and standard boulders. (Empty frames where samples fell outside the 95% PI covered by the vertical axis were removed.) This corresponded to a total playing time of 6.3 minutes prior to the animation automatically looping back to the first frame. These stimuli are shown in Figures 1c and Figures 1d. The latter is represented in static form in this document but was shown as an auto-played animated GIF in our experiment.

For the large effect size, which corresponds to a Cohen's d of approximately 1.0 and probability of superiority of 76%, we followed the same procedure but shifted the mean of the special boulder from 104 meters to 116 meters, keeping the standard deviation at 15.3 meters, as was the case for the small effect size. We used captions that matched the information shown in each figure (as opposed to including extra information as in two conditions of the previous experiment).

Participants

We recruited 2,400 participants from Amazon's Mechanical Turk for the experiment. We set our sample size so that we had approximately 80% power to detect a minimum difference of 5 percentage points in reported probability of superiority between conditions at a 5% significance level. All participants were located in the U.S. with Mechanical Turk approval ratings of 97% or above. Participants were randomly assigned to one of eight conditions (95% CIs w/ small effect, n = 299; 95% CIs rescaled w/ small effect, n = 314; 95% PIs w/ small effect, n = 300; HOPs w/ small effect, n = 278; 95% CIs w/ large effect, n = 293; 95% CIs rescaled w/ large effect, n = 308; 95% PIs w/ large effect, n = 300; HOPs w/ large effect, n = 308). Each participant received a flat payment of $1.00.

Procedure

The procedure for this experiment was identical to the previous experiment with one addition. Once participants had read the instructions, seen statistics of the standard and special boulders, provided their willingness to pay for the special boulder, and estimated the probability of winning with each boulder, they saw one additional screen. This screen did not show any of the statistics about the boulders or the visualizations, but instead contained two copies of the Distribution Builder interface: one for the standard boulder and one for the special boulder. Each of these was configured to use the available page width, with the standard boulder interface shown first and the special boulder interface shown below it. In each interface participants placed 100 balls into 20 equally spaced bins using the plus and minus buttons in each bin to indicate how many out of 100 slides they thought would land at each distance. Once they had placed 100 balls in each interface they could submit their responses and complete the experiment.

Results

As per our pre-registration plan, we again removed participants who failed the attention check by responding with something other than 50 out of 100 for the stated probability of winning with the standard boulder.[4] This left us with 1,830 participants.

H3. **Willingness to pay.** The analysis for this experiment was similar to the previous one, but with more conditions and comparisons between them. Figure 4 shows the results for willingness to pay for each condition, where facets indicate the effect size condition participants were assigned to. As before, we show two types of error bars: the black circles on the top have error bars corresponding to one standard error in estimating the mean willingness to pay, whereas the grey triangles on bottom have error bars that cover one standard deviation in individual responses. Looking at the small effect size, we see similar results to the first experiment: participants were willing to pay substantially more on average when shown 95% CIs (88 Ice Dollars) compared to 95% PIs (52 Ice Dollars). Looking at the alternative visualizations, we see that rescaling the axis on the 95% CI visualizations helps somewhat, with mean willingness to pay falling in between these two extremes (70 Ice Dollars). HOPs performed similarly to 95% PIs (mean willingness to pay of 49 Ice Dollars).

We see similar, but less stark, differences between visualization conditions for the large effect size. Comparing results for the same visualization at different effect sizes, we find that participants were more sensitive to differences in the underlying effect sizes when shown 95% PIs and HOPs compared to either of the 95% CI visualizations. This further supports the idea that it is difficult to infer effect sizes from visualizations that focus on statistical significance. Running the two-way ANOVA specified in our pre-registration plan shows statistical significance for all planned comparisons between visualizations ($t(1825) < -3.5, p < 0.001$) with the exception of 95% PIs versus HOPs, for which there is no statistically significant difference in willingness to pay.

With the small effect size, as in Experiment 1, participants in all conditions had a willingness to pay that exceeded the normative value of 17.5 dollars, especially in the CI conditions. With the large effect size, the normative value was 65 dollars, calculated as the difference between the expected value with the special boulder ($250 \times .76$) and that with the standard boulder ($250 \times .5$). Here participants in the PI and HOPs conditions came quite close to the normative answer, while those in the CI conditions exceed it.

H4. **Stated probability of superiority.** As shown in Figure 5, results for stated probability of superiority follow a similar pattern, with even larger and more consistent differences across

[4]Repeating the analysis below and including these participants yields similar results.

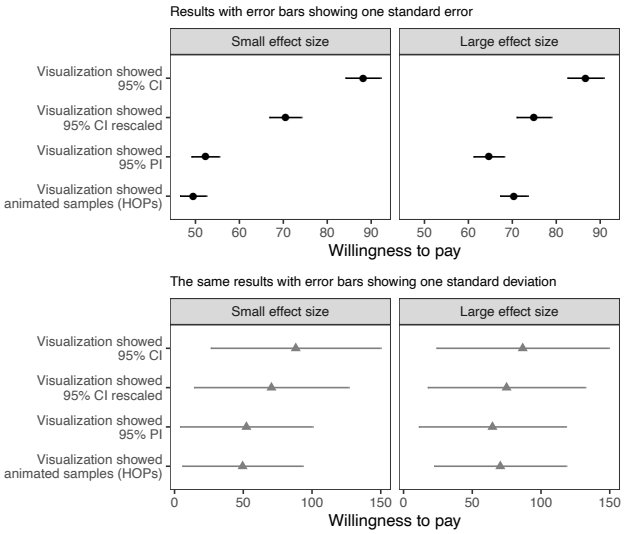


Figure 4: A summary of the willingness to pay for the treatment in our second experiment. Across both small and large underlying effect sizes, participants were on average willing to pay the most for the treatment when shown 95% CIs and least when shown 95% PIs or HOPs. 95% CIs with a rescaled axis fell between these two extremes.

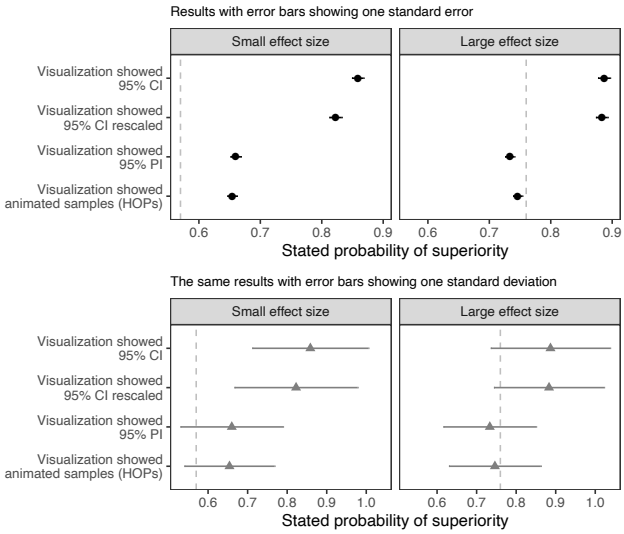


Figure 5: The estimated probability of superiority for our second experiment, which shows similar trends to willingness to pay. Dashed lines show the true underlying probability of superiority for the small and large effect size conditions.

visualization conditions. Participants estimated the probability of winning with the special boulder to be substantially higher when shown 95% CIs compared to 95% PIs or HOPS, and were less sensitive to changes in the underlying effect size when error bars showed 95% CIs. Comparing average probability of superiority between participants who saw 95% CIs between the small and large effect sizes shows an average increase of only 3 percentage points, whereas participants who saw 95% PIs or HOPs show a 10 percentage point difference between effect sizes. Looking at the distribution of responses, we note that the mode of the distribution for CIs was close to 1.0, even with the small effect size. A ceiling effect limits the amount the mean can move in the large effect size.

As in our previous experiment, we see that participants overestimated the probability of superiority for the small effect size relative to its true value of 57%. The large effect size, however, was more or less accurately perceived in the 95% PI and HOPs conditions relative to its true value of 76%, but was still overestimated in both of the 95% CI conditions.

As above, a two-way ANOVA shows statistical significance for all planned comparisons between visualizations ($t(1825) = -2.24, p = 0.025$ for 95% CIs rescaled versus 95% CIs; $t(1825) < -16.8, p < 0.001$ for others) with the exception of 95% PIs versus HOPs, where there is no statistically significant difference in stated probability of superiority.

H5. **Implied standard deviation.** Our last two analyses leverage responses collected through the Distribution Builder interface, which are summarized in Figure 6. This figure contains all 366,000 data points gathered through the Distribution Builder, where we have grouped participants by the effect size and visualization they saw and summed the counts in each bin across participants. The aggregate histogram for the standard boulder in each condition is shown in blue, the histogram for the special boulder is in red. The solid lines show the true distributions for the standard and special boulders for comparison, and the dotted vertical lines show the true means of each distribution. Looking across visualizations shows a clear pattern: the means of each distribution are well estimated across all conditions, but participants who were shown 95% CI visualizations (the right two columns) perceived both distributions to be more concentrated than those who saw 95% PIs or HOPs (the left two columns).

These aggregate histograms do, however, hide some variability across responses that individual participants gave: some distributions were symmetric and bell-shaped, others not. To better capture this variation across individuals, we first computed the implied standard deviation (SD) for each of the distributions each participant provided, resulting in 3,660 estimates (two for each of the 1,830 participants). We then grouped these implied standard deviations by visualization condition and effect size and looked at this measure across participants. This analysis confirms that despite variability across participants, the same pattern as in the aggregate histogram holds: for the small effect size, distributions for the 95% PIs and HOPs conditions have an average implied SD of approximately 17.5 Ice Dollars, which is much closer to the true value of 15.3 Ice Dollars than the average implied SD of approximately 11 Ice Dollars for the 95% CI conditions.

A two-way ANOVA on implied standard deviation across participants shows similar results to willingness to pay and stated probability of superiority. All planned comparisons between visualizations are statistically significant ($t(3655) = 2.1, p = 0.04$ for 95% CIs rescaled versus 95% CIs, $t(3655) >$

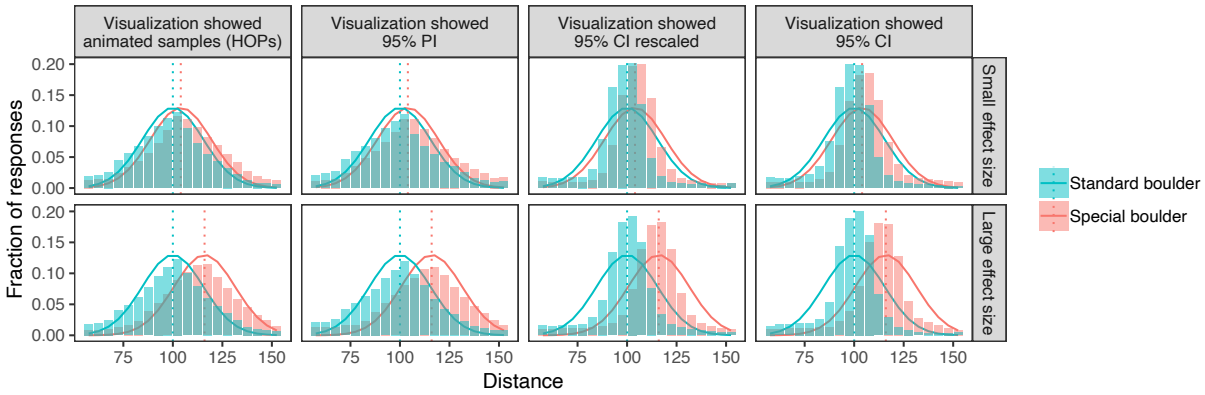


Figure 6: An aggregated view of all 366,000 data points collected through the Distribution Builder interface in our second experiment. Solid lines show the true distributions from which the stimuli were generated. The true means, indicated by the dotted vertical lines, were well-estimated in all conditions, but participants who saw either of the 95% CI visualizations significantly underestimated the variance of each distribution, as shown in the right two columns of the figure.

20, $p < 0.001$ otherwise) with the exception of 95% CIs versus HOPs, where there is no statistically significant difference.

H6. **Implied probability of superiority.** Finally, as shown in Figure 7, we find that the implied probability of superiority from the histograms elicited through the Distribution Builder match the stated probabilities of superiority from earlier in the experiment quite closely, and show nearly identical results across conditions. We see the same bias towards the middle of the response range as in the stated probability of superiority, but in each case responses from 95% PIs and HOPs are both closer to the truth than the 95% CI visualizations.

A two-way ANOVA on implied probability of superiority across participants shows no statistically significant difference between 95% PIs and HOPs or between 95% CIs rescaled and conventional 95% CIs. All other planned comparisons are statistically significant ($t(1825) < -11, p < 0.001$).

We note that a prior study from the literature [26] found differences in accuracy in estimating probability of superiority between HOPs and 95% PIs, which we do not observe here. Looking into the data from both studies, we noticed that in the present study, very few participants stated probabilities of superiority in the wrong direction (i.e., less than .5 for the superior option), while it happened more often in the prior study, impacting accuracy. This is even the case for participants who did not pass the attention check in this study. We suspect this difference may be primarily due to participants in the present study being asked to state probability of superiority twice: first for the standard boulder and then for the special boulder. Responding to an easier first question with an answer of .5 may have made it clear in participants' minds that the second probability must be greater than 0.5.

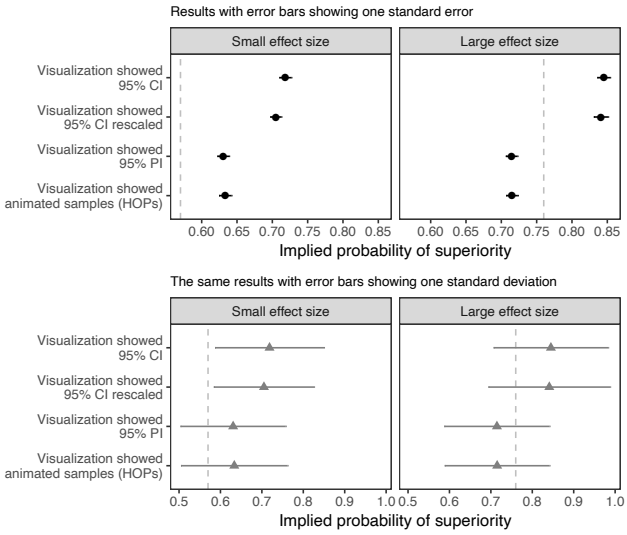


Figure 7: The probability of superiority implied by the histograms elicited through the Distribution Builder in our second experiment, which shows similar trends to the stated probability of superiority for this and our previous experiment. Dashed lines show the true underlying probability of superiority for the small and large effect size conditions.

DISCUSSION

Responsible authors know that it is important to communicate uncertainty when making statistical claims. Ideally, authors who use an estimation approach should include both outcome and inferential uncertainty in reporting a study. However, scientists face choices about how to graphically present results, and these choices influence how people perceive the effects they investigate. This article investigated four ways of presenting scientific results (95% confidence intervals, 95% prediction intervals, rescaled 95% confidence intervals, and HOPs) and their effects on readers' willingness to pay for a treatment, perceived probability of superiority of a treatment, and beliefs about the distributions of outcomes under the treatment. Across all measures, we find that visualizations oriented towards inferential uncertainty (95% CIs) led to the greatest deviation from normatively correct answers. People presented with 95% CIs overstated willingness to pay by the greatest amount, overstated probability of superiority by the greatest amount, and understated the variability in outcomes

by the greatest amount. Presenting 95% CIs on rescaled axes improved answers somewhat, but responses closest to the normatively correct answers across all measures were attained by people presented with visualizations that encoded information about variation in individual outcomes (95% PIs or HOPs).

These results are important because scientists often display their results using 95% CIs, the least accurate format we tested. Many scientists employ even narrower intervals of plus or minus one standard error which could lead to even greater inaccuracies than those measured here. Another important finding is that misperceptions were largest for visualizations depicting smaller effect sizes (Cohen's d of around .25), which is concerning because smaller effects tend to be the most commonly studied effects [10] Because scientists often use CIs and publish small effects, many scientific results are likely exaggerated in the minds of some readers. Speculating, the negative consequences of biased beliefs could be excessive faith in the effectiveness of policy interventions or the reinforcement of stereotypes (e.g., the false conclusion that statistically significant differences in means between groups implies that all the members of one group are superior to those in another, when in actuality the distributions overlap substantially due to high degrees of variation within groups). Investigating whether visualizing inferential uncertainty impacts a reader's tendency to engage in dichotomous thinking about effects (e.g., [6]) is worth exploring in future work.

While the results in this work are rather consistent across studies, our experiments were limited to a certain set of tasks, stimuli and participants. The questions participants in our studies answered depended on an accurate perception of the distribution of outcomes under the treatment. Future studies might examine how viewing outcome uncertainty, for example with PIs, influences people's perceptions of sampling distributions. We also investigated only four visualizations out of a large space including boxplots, histograms, densities, violin plots, and beyond. It may be the case that some of these alternatives perform even better under the scenarios we tested in terms of alignment with normative answers. It would be interesting to compare versions of these visualizations that orient readers towards inferential uncertainty with versions that emphasize outcome uncertainty. It would also be interesting to consider whether transforming the information shown in charts—such as by plotting the distribution of the difference between the two conditions, as some researchers have proposed [37]—reliably impacts perceptions for the better or worse.

In addition, we only studied laypeople, who likely read about scientific studies in the press and are important consumers of statistical information, but may have limited experience with error bars. Despite this, the patterns they created with the Distribution Builder for the most part followed symmetrical bell shaped patterns at the individual level and, as Figure 6 shows, on the aggregate level. It is an open question as to whether these misperceptions of results extend beyond lay readers to expert readers such as scientists. Based on our own perceptions, we think they might. For example, Figure 2 shows a effect size with a Cohen's d of .57, which is rather large.[5] To our eyes, the effect seems larger in the top left panel than in the bottom left panel, even though we know that it isn't. Perhaps this is because we are so accustomed to looking at confidence intervals that it is hard to interpret error bars differently. It would be worth testing whether other researchers have similar perceptions. Regardless of whether experts show the same behavior as laypeople, we believe that our results have implications for how scientists should communicate their own work to laypeople, for example in the popular press.

We see a few promising directions for future research. The first would be finding an ideal visualization for conveying both inferential and outcome uncertainty. While PIs and HOPs led to greater accuracy than than CIs in our tasks, potentially valuable information about inferential uncertainty might be lost if CIs were universally replaced by PIs or HOPs. Modifying PIs or HOPs to display inferential uncertainty as well as outcome uncertainty may prevent some readers from overestimating effect size, but would require axes to be scaled such that small differences in means are hard to read, potentially leading to greater error in some inferences. One alternative is to show both types of plots as we have done in most of the figures in this paper, however, doubling the number of plots for all papers seems excessive and might become burdensome for both authors and readers. An innovative representation that does justice to effect size and sample mean differences would be welcome, and would require experiments that assess participants on both concepts.

Second, we note the literature is largely silent on how people process the visualizations we studied to arrive at intuitive estimates of probability of superiority and related measures. For HOPs, a plausible mechanism would be frequency encoding [19], which would make a concrete prediction, namely that the frequency of superiority experienced during the time in which the participant is looking at the visualization would be reflected their response. For error-bar based representations, there are no frequencies to encode but eye tracking and think-aloud measures collected during responses or during the use of the Distribution Builder could give insight into the process [36].

Lastly, while this paper studied willingness to pay and perceptions of differences between distributions, our work would have greater practical relevance if it studied how different visualizations impacted decision making about policies. It would be particularly interesting to identify the types of decisions for which each visualization—even those that performed poorly here—is well suited. That said, given the misperceptions that we document and the reality that many published effects are small, it does seem to be the case that putting less emphasis on statistical significance in visualizations and more emphasis on effect sizes would benefit scientists and their audiences.

ACKNOWLEDGMENTS

We thank William Cai and Jennifer Allen for their help in programming and running the experiments described in this paper.

[5] To put it in perspective, about 75% of the studies in [10] had smaller effect sizes than a Cohen's d of .57 (i.e., correlation of .274).

REFERENCES

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

[2] American Psychological Association and others. 2001. *Publication manual (5th edition)*. American Psychological Association Washington, DC.

[3] Nicholas J Barrowman and Ransom A Myers. 2003. Raindrop plots: a new way to display collections of likelihoods and distributions. *The American Statistician* 57, 4 (2003), 268–274.

[4] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4 (2005), 389.

[5] Melanie L Bell, Mallorie H Fiero, Haryana M Dhillon, Victoria J Bray, and Janette L Vardy. 2017. Statistical controversies in cancer research: using standardized effect size graphs to enhance interpretability of cancer-related clinical trials with patient-reported outcomes. *Annals of Oncology* 28, 8 (2017), 1730–1733.

[6] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. (2019).

[7] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 5 (2013), 365.

[8] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, and others. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637.

[9] Beth Chance, Robert del Mas, and Joan Garfield. 2004. Reasoning about sampling distribitions. In *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 295–323.

[10] Open Science Collaboration and others. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

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

[12] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

[13] Geoff Cumming, Fiona Fidler, Pav Kalinowski, and Jerry Lai. 2012. The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology* 64, 3 (2012), 138–146.

[14] Geoff Cumming and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist* 60, 2 (2005), 170.

[15] Peter Cummings. 2011. Arguments for and against standardized mean differences (effect sizes). *Archives of pediatrics & adolescent medicine* 165, 7 (2011), 592–596.

[16] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 65.

[17] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 144.

[18] Rocio Garcia-Retamero and Edward T Cokely. 2013. Communicating health risks with visual aids. *Current Directions in Psychological Science* 22, 5 (2013), 392–399.

[19] Gerd Gigerenzer. 1994. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability*. Wiley, 129–161.

[20] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review* 102, 4 (1995), 684.

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

[22] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment & Decision Making* 9, 1 (2014).

[23] Rink Hoekstra, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review* 21, 5 (2014), 1157–1164.

[24] Ulrich Hoffrage and Gerd Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic medicine* 73, 5 (1998), 538–540.

[25] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. 2018. Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 446–456.

[26] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[27] Harald Ibrekk and M Granger Morgan. 1987. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis* 7, 4 (1987), 519–529.

[28] Christopher H Jackson. 2008. Displaying uncertainty with shading. *The American Statistician* 62, 4 (2008), 340–347.

[29] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on visualization and computer graphics* (2018).

[30] Peter Kampstra and others. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. (2008).

[31] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. 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*. ACM, 5092–5103.

[32] Yea-Seul Kim, Logan Walls, Pete Krafft, and Jessica Hullman. 2019. A Bayesian Cognition Approach to Improve Data Visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[33] Martin Krzywinski and Naomi Altman. 2013. Points of significance: error bars. (2013).

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

[35] Nathaniel Schenker and Jane F Gentleman. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* 55, 3 (2001), 182–186.

[36] Michael Schulte-Mecklenbeck, Joseph G Johnson, Ulf Böckenholt, Daniel G Goldstein, J Edward Russo, Nicolette J Sullivan, and Martijn C Willemsen. 2017. Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science* 26, 5 (2017), 442–450.

[37] Transparent Statistics in Human–Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. (Feb 2019). DOI: http://dx.doi.org/10.5281/zenodo.1186169 (Available at https://transparentstats.github.io/guidelines).

[38] Leland Wilkinson. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American psychologist* 54, 8 (1999), 594.