Imagining Replications: Graphical Elicitation & Discrete Visualizations Improve Reasoning About Experimental Uncertainty

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

People often have erroneous intuitions about the results of uncertain processes, such as scientific experiments. We present a graphical elicitation technique for improving people's statistical inferences about uncertain data. We contribute a design space survey to develop and evaluate the effectiveness of 12 discrete and continuous interfaces for graphically eliciting subjects' predictions about probability distributions. We further conduct a controlled study in which participants use the best performing discrete and continuous elicitation interfaces to estimate the uncertainty around an observed experimental effect, then view their distribution against the true sampling distribution. We find that while using discrete probability representations is sufficient to improve inferences about future replications of a single experiment, graphically estimating the possible effects from experiment replications is a more effective way to improve one's ability to make predictions about replications of new experiments. Our work has implications for probability elicitation and uncertainty communication.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces; Graphical User Interfaces

Author Keywords

Graphical elicitation, information visualization, uncertainty, sampling distribution.

INTRODUCTION

Statistical information is increasingly encountered in everyday life, from media reporting on scientific studies or political elections to tracking one's health. It is natural for people to want to make predictions about future events from such data. However, in making predictions, people often overlook sample size information and its relation to sample variance [51]. People also find it hard to understand how the distribution of a sample statistic, like a mean, is related to the underlying data distribution [5, 8]. Not understanding these concepts can result in non-optimal statistical reasoning, including overconfidence in uncertain estimates [35].

For example, imagine a typical results report from a controlled experiment. The author describes an observed effect and uncertainty around the effect. To reason about whether the result is reliable (e.g., likely to replicate under similar circumstances) requires understanding the likely distribution of effect sizes upon repeating the experimental process. Typical uncertainty representations, such as error bars and confidence envelopes, make it easy to ignore or misinterpret reported uncertainty [4, 28]. Concerns around the replicability of experimental results in multiple scientific fields [26, 42, 44] suggest that scientific expertise may not prevent people from overlooking the implications of uncertainty in statistical evidence. Consequently, statistical reformers call for the development of tools to provide cognitive evidence that leads people to develop accurate interpretations of statistical phenomena [14].

Asking a person to represent or predict information can be powerful ways to provide cognitive evidence through active learning [6, 46]. For example, asking people to graphically reproduce statistical information like the risk associated with a disease in a discrete (frequency) format can lead to more accurate probability inferences [12, 38, 49]. Alternatively, asking learners to make predictions about a data set, such as in pre-test, may increase their ability to learn from subsequent representations of that data [16]. Though typically associated with educational contexts, active learning strategies are used to elicit user predictions about statistical models characterizing uncertain processes in several recent interactives by the New York Times [1, 7, 24, 29]. For example, "Make Your Own

Senate Forecast" asks readers to predict the party majority in voting outcomes for various states. The potential of such interactions to improve data comprehension is promising, as the act of predicting may prompt deeper consideration of assumptions affecting the outcomes and the meaning of the visualized data.

We study how asking users to graphically estimate the implications of uncertainty with either discrete or continuous probability representations impacts statistical reasoning. Our goal is to see whether positive effects of graphical construction tasks and discrete probability formats identified in prior work on classic Bayesian reasoning problems [12, 48] and population distributions [22] also apply when participants are asked to make inferences from sampling distributions around reported statistics. Our specific goal is to see whether graphical estimation can improve inferences about one commonly misinterpreted uncertain process: experimental results.

Our contributions based on these goals are to:

1. Evaluate the design space of graphical elicitation interfaces for probability distributions. We develop discrete-outcome and continuous probability framed interfaces, extending prior work in probability elicitation [22, 40]. We find that **continuous probability interfaces result in faster, more precise and more satisfying drawings of a reference distribution.**
2. Evaluate how having users graphically estimate the outcomes of replicating a scientific experiment impacts their statistical reasoning. We find through controlled study that **users who graphically estimate the distribution of possible effects and see their prediction against the ground-truth distribution more accurately predict the outcome of replicating a new study.**
3. Identify how using a discrete versus a continuous visualization to view and/or estimate a probability distribution impacts users' statistical reasoning. We **observe an advantage of discrete representations for recalling a sampling distribution**, suggesting that discrete visualizations with limited numbers of outcomes may provide a useful format for remembering statistical information among non-experts.

Our results inspire future applications of graphical probability elicitation for Bayesian data analysis and the transparent statistics [30] and RepliCHI movements [53, 54].

RELATED WORK

We summarize related work in teaching statistical reasoning and the benefits of constructing probability representations.

Tools for Teaching Statistical Reasoning

Statistical reasoning involves interpreting data, graphical representations, and statistical summaries. Underlying accurate statistical reasoning is a solid understanding of concepts related to uncertainty, such as understanding and being able to make inferences about distributions, randomness, and sampling [19, 18]. In the last two decades, reform movements in educational psychology and statistical pedagogy have criticized teaching approaches in statistics for focusing on rote memorization over active reasoning [18, 18]. Proposals to improve educational approaches to statistical reasoning advocated focusing more on data and less on theory [5, 10, 37, 48] and including active learning and technology through analysis and simulation [5, 37, 48].

Sampling distributions (distributions of sample statistics) are notoriously difficult for people to form accurate intuitions about. Researchers have documented common misinterpretations among students (e.g., the sampling distribution should look like the population distribution) [8] and errors made among experts when interpreting statistical representations based on sampling distributions [4]. Many researchers have advocated using simulations to improve understanding of sampling distributions [8, 36], for example by enabling a user to specify a sample size and population distribution and observe the sampling distribution [15, 16, 47]. Cumming refers to such simulations as "cognitive evidence" necessary for learners to form accurate intuitions about statistical concepts. We study an orthogonal technique to simulation by focusing on how estimating the implications of statistical information can inspire learning by drawing a participant's attention to the gap between their thinking and a normative answer.

Our interest in graphical prediction is motivated in part by work by delMas et al., who found that a simulation was not enough to leave students with accurate conceptions of sampling distributions [16]. Based on a belief that contradictory evidence is required to change one's beliefs [43], the researchers developed an activity in which students ran a simulation on preloaded population distributions from pretest problems, allowing them to view the sampling distribution that they had considered before the simulation. Students who did the activity showed additional gains on posttest questions, suggesting that reflecting on one's earlier predictions may enhance statistical reasoning.

Constructing Representations to Understand Probability

Approaches to visualizing uncertainty in data include a variety of representations for distributional information that have been tested with novices (e.g., [11, 25, 31, 50]). However, visualizations of uncertainty may not be beneficial if users do not adequately understand the underlying constructs. In fact, common representations of distributional information like error bars have been shown to lead to erroneous predictions even among experts [4]. Or, people may find the uncertainty information complex and choose to ignore it [28]. JH: maybe add something about discrete, low numbers of outcomes

Actively constructing graphical representations can help people understand probability and risk. Natter and Berry [38] found that participants who completed a reflective task (i.e., portraying the size of a risk on a bar chart or answering a reflective question) were more accurate and satsified in their probability estimates. Cosmides and Tooby [12] presented people with the base rate of a disease and false positive rate of a test for the disease in order to study Bayesian reasoning. Participants who used a graphical display to fill in the information more accurately estimated how many people had the disease than those who viewed filled-in graphs. Sedlemeier and Gigerenzer [49] conducted a controlled study of tutorial programs for Bayesian problems. Programs that instructed

users on how to create frequency representations (motivated by research indicating that cognitive processing of probabilities is easier given a frequency format [20]) led to better immediate Bayesian reasoning and accuracy five weeks later compared to a rule training program.

More recent studies indicate that graphical interfaces are advantageous for eliciting a person's subjective probability distribution [22, 21]. Goldstein and Rothschild [22] find that a graphical method in which subjects construct a probability distribution out of discrete outcomes leads to responses that better capture the ground truth distribution for a set of viewed stimuli than verbally asking for properties like the median, extremes, and fractiles. We conduct a design space survey to evaluate what properties make a graphical elicitation interface effective. We also contribute a controlled study to examine how graphical prediction using such interfaces can be applied to improve understanding of experiment results.

STUDY 1: EVALUATING THE ELICITATION DESIGN SPACE

What makes a graphical interface for eliciting a probability distribution effective? The two-fold goal of our first study is 1) to develop a design space of graphical elicitation techniques for probability distributions that are suitable for novice as well as expert users, and 2) identify the most effective and usable interfaces for further study.

Part 1: Design Space Development and Iteration

We first developed many graphical elicitation interfaces for predicting a probability distribution. We began with sketching and paper prototyping followed by implementation and informal testing among the authors and others in their labs. Three design goals informed our process. Throughtout the process, we differentiated two types of interfaces that differed in how they framed the distribution: discrete outcome interfaces and continuous probability interfaces.

Design Goals

We are interested in finding ways to improve statistical reasoning among novices in statistics as well as experts. A graphical elicitation interface should therefore be **easy to use without training**. Each of the interfaces we designed rely on direct manipulation to reduce abstractness and complexity.

To reduce common misinterpretations a graphical elicitation interface should also **encourage accurate reasoning about probability**. We developed a number of designs that use discrete outcome framings, rather than continuous representations, based on the evidence that thinking in terms of frequencies helps people engage in more accurate statistical reasoning [20, 23].

The **precision** of the distributions that an interface supports creating is important: the degree to which a user can match an external or internal probability distribution using the interface will directly affect accuracy (for example, one cannot as precisely represent a probability distribution using 20 outcomes as one could using 100 outcomes). Related to precision is expressiveness, or how large the space of possible distributions that can be created with the interface is. It is desirable for an interface to allow users to construct a wide range of distribution shapes to support their natural predictions. We therefore intentionally chose not to develop interfaces that force symmetry or other normative properties of a sampling distribution.

Elicitation Interfaces

We developed two categories of interfaces to explore the space of trade-offs between our design goals: discrete outcome framed interfaces and continuous probability interfaces.

Discrete Outcome Interfaces: We developed 10 discrete outcome interfaces. To explore the trade-off between precision and accurate interpretation, we varied the number of outcomes that a user is given to construct the distribution with each of 6 different interaction techniques.

We implemented two versions of the *balls-and-bins* interface evaluated by Goldstein and Rothschild [22] for specifying a distribution: a 20 ball version and a 50 ball version (Fig. 1g). To construct a distribution, the user clicks an up arrow ($\triangle$) below each of 10 bins equally spaced along the x-axis to add one outcome (i.e., circle) at a time to that bin. To remove outcomes, the user clicks a down arrow ($\triangledown$) below the bin. Each bin holds up to 10 or 20 circles and the user is given 20 or 50 total balls to use (20 ball and 50 ball version respectively). We opted to create a 50 ball version rather than a 100 ball version as in [22] based on preliminary user testing with a prototype of the latter, which participants found to require too much clicking. When the total number of allotted outcomes has been reached, the interface informs the user with a message in a feedback panel to the left of the drawing area. The user can continue to add outcomes to their distribution, but all the outcomes turn red in color until the total number being used is within the allotted limit.

We developed two types of *paint-outcomes-by-dragging* interfaces. Both types first present the user with a grid of circles representing outcomes. A standard paint-by-dragging interface allows users to fill in circles with color by dragging the mouse over each (Fig. 1d). Preliminary user testing led us to also create a fill-down paint-by-dragging interface that allows users to drag over a circle in the grid to fill that circle and all circles directly below it in the grid column (Fig. 1e). This interaction is similar to a draggable distribution drawing interface in the OnlineStatBook [33], but with circles representing discrete outcomes as opposed to a barchart, and an added feature that lets users achieve the same effect by clicking on the circle as well as dragging. In each case, the user was given either 20 or 100 total outcomes to use (with a corresponding 10x10 or 20x18 grid. When the total number of allotted outcomes has been reached, the interface informs the user with a message in a feedback panel to the left of the drawing area. The user can continue to more outcomes to their drawing, but all the outcomes turn red until the total number being used is within the allotted limit.

We developed a *pull-up* interface that allows the user to drag up handles that are equally spaced along the x-axis in order to add outcomes (Fig. 1a). The user is given either 20 or 100 total outcomes, with 10 or 20 handles along the x-axis and per column limits of 12 or 20 outcomes. When the total number

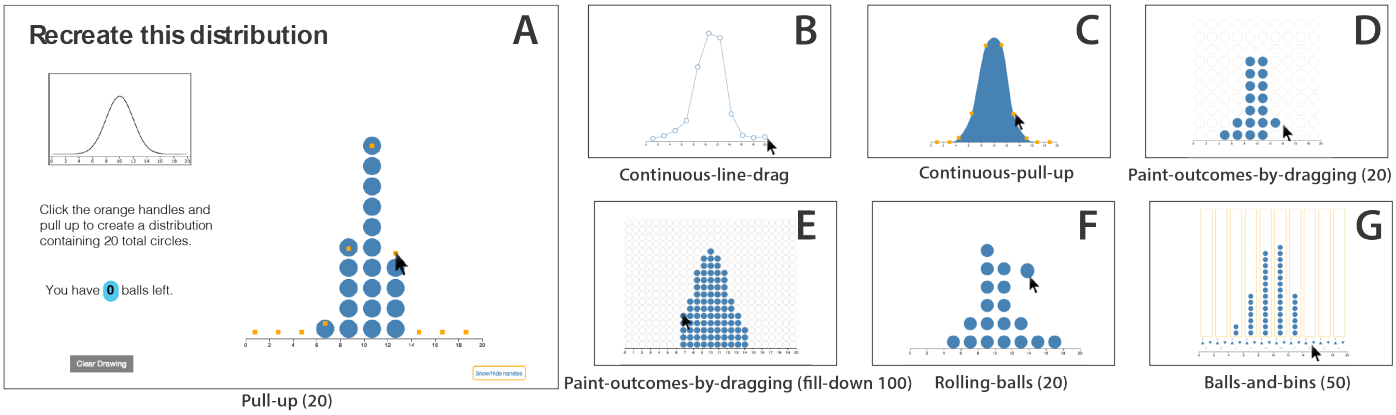

Figure 1. Evaluative study interface for pull-up 20 outcomes elicitation interface (left), with examples of six other interaction types we evaluated.

of allotted outcomes has been reached, the interface informs the user with a message in a feedback panel to the left of the drawing area. The user can continue to more outcomes to their drawing, but all the outcomes turn red until the total number being used is within the allotted limit.

Rather than requiring the user to add or fill outcomes, *rolling-balls* interfaces present a uniform distribution of outcomes for the user to drag into a distribution shape (Fig. 1f). The user was given either 20 or 100 total outcomes to use.

We did not label the *y*-axis with probability values for any of the interfaces, in keeping with the unlabeled axis in the reference distribution. We suspect that adding labels could negatively impact peoples' ability to specify the distributions, as thinking about relative probabilities tends to be easier for people than thinking about absolute probabilities [40].

Continuous Probability Interfaces: To further explore the trade-off between precision and accurate interpretation, we created several continuous representation interfaces to allow finer control over the probability distribution.

We developed a *continuous-line-drag* interface that allows a user to drag from the left to the right side of the axis to shape a line into the probability density function they desire (Fig. 1b). As the line is created, a total of 11 equally spaced handles along the *x*-dimension are added to the line. The user can adjust the shape of the curve by dragging the position of the handles after releasing the mouse to end the initial drawing.

We also developed a *continuous-pull-up* interface, in which a user is presented with 10 equally spaced handles along the *x*-axis (Fig. 1c). The user can create a probability density function by dragging up the handles.

We did not label the *y*-axis with probability values for either of the continuous interfaces.

Part 2: Evaluation Survey

We evaluated the effectiveness of the 12 elicitation interfaces using a controlled study with participants from two populations: a general population consisting of workers on Amazon's Mechanical Turk, and a population we expected to have slightly more statistical experience, consisting of students and other individuals associated with a human-computer interaction mailing list at a large public university. Our evaluation focused on three dimensions that are likely to impact the effectiveness of an interface for supporting understanding of uncertainty: *ease of use*, or how easily a user could replicate a given distribution with the interface; *accuracy*, or how closely a user could replicate a distribution with the interface; and *satisfaction*, or how satisfied a user was with their ability to use the interface to create the distribution.

Study Design and Procedure

We designed a repeated measures between subjects study in which individuals used multiple interfaces to construct a reference distribution. Each participant was assigned to six of the total of 14 elicitation interfaces. The participant was first introduced to the study task, which was to replicate a distribution using a variety of graphical interfaces. The participant then completed six task screens, each of which presented a single elicitation interface with brief instructions for how to use the interface (Fig. 1a). A reference distribution (a probability density function for a normal distribution with μ=10 and σ=2) was presented to the left of the drawing area. The participant was instructed to use the drawing interface to replicate the reference distribution as best she could. After completing the six task screens, the participant was shown a rating screen with the names of the six interfaces, which could be clicked to reload the interface in the drawing pane. She used a slider from "Not satisfied" to "Very satisfied" to the right of each interface to rate her satisfaction with the interface for the task.

In assigning participants to combinations of six interfaces, we paired the interfaces such that a participant was always assigned both the high resolution (e.g., 50 or 100 outcome) and low resolution (20 outcome) version of the same interaction mechanism (e.g., balls-and-bins, rolling-balls, etc.) We designed this pairing to reduce the subject-specific variance in comparing between the high and low resolution version. We also paired the two continuous probability interfaces based on their similar framing of the distribution. We counterbalanced interface pairs across participants within both samples.

We recruited 30 participants by advertising the survey on a human computer interaction email list at a large university. Participants were optionally able to enter their email address

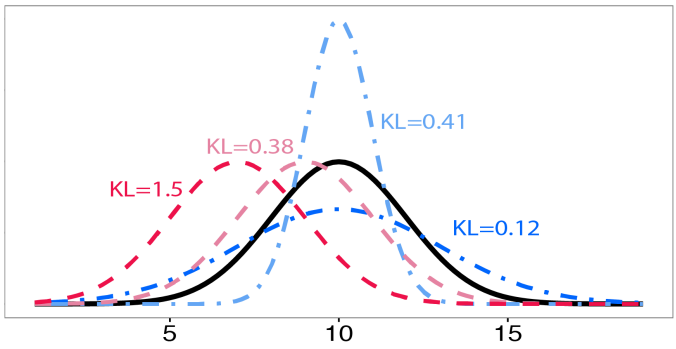

Figure 2. KL divergence for distributions with varying standard deviations (blue) and locations (pink) relative to a reference distribution (black).

Table 1. My caption

Interface	Ease-of-Use (Time in s)	Satisfaction Rating	Accuracy (KL Div.)
Balls-and-bins 20	85 (46)	40 (29)	0.22 (0.20)
Balls-and-bins 50	92 (10)	44 (30)	0.34 (0.74)
Fill-down paint-by-dragging 20	57 (43)	47 (30)	0.35 (0.29)
Fill-down paint-by-dragging 100	123 (91)	49 (35)	0.32 (0.28)
Paint-by-dragging 20	76 (84)	46 (29)	0.62 (1.5)
Paint-by-dragging 100	97 (70)	59 (31)	0.78 (1.4)
Pull-up 20	91 (83)	47 (30)	0.19 (0.22)
Pull-up 100	158 (164)	55 (33)	0.26 (0.40)
Rolling-balls 20	83 (61)	30 (25)	0.26 (0.25)
Rolling-balls 100	150 (108)	30 (29)	0.40 (0.23)
Continuous-line-drag	69 (50)	56 (30)	0.21 (0.18)
Continuous-pull-up	68 (57)	65 (28)	0.14 (0.25)

on the introductory page to the survey to be entered into a lottery for a $50 gift card for their participation. We recruited 80 additional participants from the Master's workers pool on Amazon's Mechanical Turk. Each worker was paid $1.50 for their participation.

Dependent Measures

We operationalized ease-of-use, accuracy, and user satisfaction as follows:

- *Ease-of-use*: Total time spent using the interface.
- *Accuracy*: Kullback-Leibler (KL) divergence between the user's generated distribution and the reference distribution. KL divergence is an information theoretic measure of similarity between two distributions P and Q, which measures, roughly, the amount of information lost if Q is used to approximate P.
- *Satisfaction*: Satisfaction slider response value.

To calculate accuracy as KL divergence to the reference distribution for the discrete interfaces, we created a discretized version of the reference distribution given the width of the interval used for responses. For example, the interval size is 1 if the interfaces centered participants' inputs at 1, 2, 3, etc. and the interval size is 2 if the interface centered participants' inputs at 2, 4, 6 etc. We report the KL divergence for each response distribution relative to the reference distribution (note that KL divergence is not symmetric).

Results

Prior to comparing the interfaces along the three dimensions of interest, we checked for differences between the two study samples for any of the three dimensions. The MTurk sample took slightly longer per interface (13 seconds or 14%) than the university sample (μ_{MTurk}: 102s, σ_{MTurk}: 96s; μ_{Univ}: 89s, σ_{Univ}=88s). Mean KL divergence and mean satisfaction ratings were not distinguishable (μ_{MTurk}: 0.33, σ_{MTurk}: 0.39, μ_{Univ}: 0.32, σ_{Univ}=0.68, and μ_{MTurk}: 46, σ_{MTurk}: 47, μ_{Univ}: 31, σ_{Univ}= 32, respectively). We report below on the results pooled across both samples (full results are available in Supplemental Materials).

Table. **??** presents the results, with darker colors indicating better performance. Overall, the continuous probability interfaces outperformed the discrete interfaces, resulting in relatively low values on all three dimensions. The continuous-pull-up interface slightly outperformed the continuous-line-draw interface, receiving a higher average satisfaction rating (65.3 vs 56.1) and lower KL divergence (0.15 vs 0.20). We test the continuous-pull-up interface as our continuous probability interface in our statistical reasoning study.

Overall, the results indicate a trade-off between accuracy and ease-of-use: the less time it took participants to replicate the reference distribution, the greater the KL divergence. Interestingly, KL divergence is not predictably lower when the resolution (number of outcomes) of the interface is high. Instead, certain interactions, such as clicking below columns to add outcomes (balls-and-bins) and dragging up columns of outcomes (pull-up) lead to greater precision, even with smaller numbers of outcomes.

Among discrete probability interfaces, both the balls-and-bins and and pull-up interfaces performed well across the three measures, with the exception of the time it took participants to replicate the reference distribution with the pull-up 100 outcomes interface (158s, relative to 98s across all interfaces). To compare the discrete interfaces while accounting for the repeated measures design, we used separate mixed effects models implemented in glmerstan [**?**] for each dependent measure. We specified interface type, order, and study sample (MTurk or university) as fixed effects and subject as a random effect, using a maximal random effects structure to allow for varying effects of order per subject [2]. We specified the balls-and-bins 50 outcomes interface as the reference group, so that the estimated effect for each interface represents the difference in the corresponding dependent measure relative to the interface that is closest to the state-of-the-art from prior work. This analysis revealed that only the pull-up 20 interface outperformed the balls-and-bins 50 interface on all three measures. We therefore test the pull-up 20 outcomes interface in study 2. Full results are available in supplemental material.

STUDY 2: REASONING ABOUT UNCERTAIN DATA

In our second study, a general sample of crowdworkers from Amazon's Mechanical Turk make predictions about the implications of uncertainty in reported experiment results.

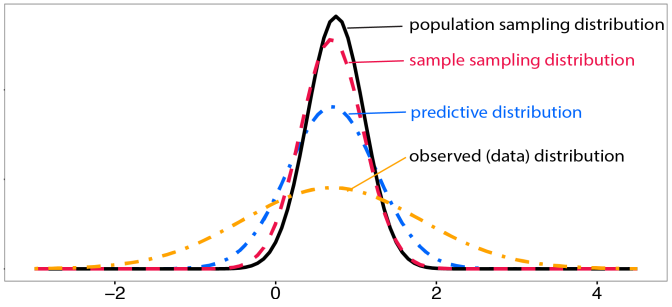

Figure 3. Placeholder figure.

Study Objectives

Prior work shows that having people graphically reiterate probability information can improve their ability to make statistical inferences, such as in classic Bayesian reasoning problems [12, 38, 49]. Prompting people to think about the answers to statistical reasoning questions before revealing them may also contribute to improved statistical reasoning [16]. Motivated by these results, we investigate the impact of a simple graphical probability elicitation task on statistical reasoning about sampling distributions.

Specifically, we are interested in whether having a user graphically estimate what will happen if an experiment is replicated prior to seeing the true (population) sampling distribution can improve her ability to make inferences about the likely effects of replicating that experiment and new experiments.

Inference from experimental data involves understanding subtle nuances between probability distributions. For example, three different sampling distributions can be used to describe the uncertainty in experimental results, in addition to the observed data distribution (Fig 3). The *observed sampling distribution* represents the sampling distribution that is constructed by setting assuming the sample mean and standard deviation represent the population mean and standard deviation. The *population sampling distribution* is the sampling distribution given the true population mean and standard deviation (which is often unknowable in practice). The *predictive distribution* is the sampling distribution that can be calculated by modeling the uncertainty around the population mean and standard deviation based on sampling error.

We draw on these distinctions in evaluating participants' understanding in our study. We present participants with the population sampling distribution for a fictional experiment that describes several sample statistics (mean, standard deviation, and number of observations). We are first interested in how well participants' can later recall the population sampling distribution using a graphical elicitation interface. In a **graphical recall task**, subjects are asked to think about 'What would happen if this study were replicated many times?' Specifying the population sampling distribution that she was shown is the most accurate response a participant could give.

As a second recall task, we inquire about the same distribution by asking participants for textual probability estimates. We expect this **textual recall task** to result in greater variability in responses than graphical elicitation based on recent work indicating improved accuracy when subjects graphically depict a population distribution [22]. However, being able to accurately answer textual questions provides evidence of the strength of a participant's mental representation of the population sampling distribution, as translating information between modalities requires one to appropriately abstract properties of the original representation [13].

In addition to better comprehending the properties of the presented distribution, another measure of participants' understanding lies in their ability to transfer what they have learned to new situations. We therefore measure all participants' accuracy on an **graphical prediction task** in which they must predict the sampling distribution for a new study, in a new domain. Participants are again given the sample mean, standard deviation, and number of observations. Participants' predictions should match the predictive distribution, indicating that the participant has accounted for the fallibility of the sample statistics as a representation of the population parameters.

Hypotheses

Our first hypothesis states that:

H1 (Graphical elicitation effect): Graphically eliciting estimates of what will happen if a study is replicated prior to presenting the true distribution will lead to more accurate probabilistic inferences about the given study and a new study.

Using discrete representations of probability has proven to be a robust strategy for improving statistical reasoning, such as in Bayesian reasoning problems [20, 23, 41, **?**], eliciting population distributions [22], and reading probability values [25, 31]. We vary whether participants use discrete and continuous representations of a probability distribution in our study to better understand 1) how the probability framing that is used affects graphical elicitation, and 2) how effects of graphical elicitation compare to those of simply using discrete representations of probability. We hypothesize that:

H2 (Discrete framing effect): Viewing the true distribution using a discrete probability representation will lead to more accurate probabilistic inferences about the given study and a new study than viewing the true distribution using a continuous probability representation.

Finally, graphical elicitation represents an implicit way of instructing participants about the implications of uncertainty in experiment results: rather than being taught the relationship between sample size and properties of the sampling distribution, they are expected to infer relevant relationships by comparing their prediction to the true distribution. However, training people directly on rules related to sampling distributions can be a viable way of improving statistical reasoning [17, 39]. As an additional benchmark for understanding the impact of graphical prediction on statistical reasoning, we include one discrete and one continuous probability framing condition in which participants complete a sampling distribution training task. We hypothesize that:

H3 (Rules training effect): Completing an instructional task about sampling distributions lead to more accurate probabilistic inferences about the given study and a new study.

Study Design

We describe the study design, procedure, and stimuli.

Study conditions

We conducted a between subjects study as a single HIT on Amazon's Mechanical Turk among U.S. workers with an approval rating of 95% or above. We used a three factor design in which participants were randomly assigned to either a **discrete or continuous representation**, and to either a **estimate elicitation or no-elicitation** condition. The **sampling distribution rules training** manipulation was a nested factor in the no-elicitation conditions. The design results in six total conditions: discrete elicitation (*D_E*), discrete no-elicitation (*D_NE*), discrete rules training (*D_R*), continuous elicitation (*C_E*), continuous no-elicitation (*C_NE*), continuous rules training (*C_R*). Based on the results of study 1, we selected the pull-up 20 interface as our discrete interface and the continuous-line-drag interface as our continuous interface.

Study Procedure

The study procedure was as follows (steps in parentheses were not completed by all conditions):

Part 1: Read Experiment Summary: After an introductory screen, participants condition watched a brief tutorial video on how to draw a distribution with either the discrete or continuous interface. All participants were then presented with a summary of a scientific experiment about the impact of a stimulant on rats' activity. The experiment summary described the experiment design, the mean observed difference between treatments, the observed standard deviation of the difference, and the sample size. To familiarize participants with distributions and variance, the experiment description also defined a distribution (i.e., a set of possible data outcomes from an experiment) and standard deviation (i.e., a quantity such that approximately 70% percent of rats will be within +-1 standard deviation of the average and approximately 95% will be within +-2 standard deviations). A figure depicting the 68-95-99.7 rule for normal distributions [52] was shown to further illustrate the concept of standard deviation.

Part 2a: (Rules Training): After reading the experiment summary, participants in the rules training condition completed a training screen on sampling distributions. We designed the screen to align with principles for effective statistical training, including clarity of the sampling process and role of chance [17] as cited in [48]. In the exercise, participants were provided with a definition of the sampling distribution and description of its relationship to the observed data distribution (specifically, that the standard error, or standard deviation of the sampling distribution, is equal to the standard deviation of the observed data distribution divided by the square root of the sample size). Participants were given the mean, standard deviation, and sample size (N) for an example observed distribution. Participants entered the latter two quantities into an interactive form, which then calculated the standard error of the sampling distribution for them.

Part 2b: (Graphical Elicitation): After reading the experiment summary, participants in the elicitation conditions were presented with a graphical elicitation interface below the experiment summary. Participants were asked to estimate what the set of possible increases in average activity score with the stimulant and without would look like if the study were replicated many times. To control for time and engagement across conditions, all other participants were asked to retype the mean difference and standard deviation of the difference in activity level while the elicitation participants were drawing.

Part 3: View Population Sampling Distribution: All participants then viewed a visualization of the population sampling distribution, the sampling distribution of the population difference given the population standard deviation. We opted to show the sampling distribution based on the population parameters because the typical goal of an experiment is to make inferences about unobserved population parameters.

Part 4: Statistical Literacy Test: All participants completed the Berlin numeracy test, a test of statistical and risk literacy consisting of four word problems []. The literacy test served as a distractor before the participants' were tested on their understanding of the experiment data.

Part 5: Make Predictions For New Experiment: All participants were then presented with a second experiment description from a different domain. After they had read the experiment description, all participants completed the graphical prediction task. For consistency, participants in the no-elicitation continuous condition used the continuous graphical elicitation interface for the prediction task, and participants in the no-elicitation discrete condition used the discrete graphical elicitation interface.

Part 6: Graphical & Textual Recall Experiment 1: After the prediction task, all participants completed the graphical and textual recall tasks. First, participants were asked to draw the distribution of expected effect sizes if the first (rats) study were replicated many times. After they had completed an initial drawing, participants were presented with probability questions asking for the probability of different effect sizes if the study were replicated many times. The questions appeared on the same screen as the graphical drawing interface, and participants were encouraged to adjust their initial drawing if they felt it necessary. All participants then completed a set of demographic questions.

Study Stimuli: Experiment Domains

Participants were presented with information about two fictional scientific experiments spanning two domains: animal behavior and computer science. Both experiments were based on descriptions of actual experiments: a study of the impact of stimulants on rats' activity [9] and of the productivity of software engineers based on the programming language used [3]. For the rat activity experiment, which was the first experiment presented and the focus of the recall tasks, participants were given the mean increase in activity and standard deviation of the increase among 40 rats given a stimulant versus a placebo in a within-subjects design. In the engineer productivity experiment, which was the focus of the graphical prediction task, participants were presented with the mean difference in the productivity of 8 engineers who used two programming

languages in a within-subjects design. The full experiment descriptions are available in Supplemental Material.

Both experiments described relatively elaborate domain-specific measures (of activity in rats, and of productivity in software engineers). We intentionally chose these measures to reduce the likelihood that participants would apply prior knowledge in any of the tasks. We also anonymized the specific treatments in both studies (i.e., the specific stimulant in the rat activity experiment and programming languages in the engineer productivity experiment).

Participant Population

We posted the study as a single HIT on Amazon's Mechanical Turk, open to workers in the United States with an approval rating of 95% or above. The reward for the HIT was $2.50. Participants were eligible for a total bonus of $2.20, including $0.10 per question within 10% of the true answer for the four Berlin Literacy test questions and 8 textual recall questions and an additional $0.50 for the graphical recall task and graphical prediction task if the elicited distribution was close to the true distribution (measured as a difference in KL divergence of less than 0.30; see Fig. **??** for context). The average bonus earned was $0.80.

We advertised the HIT for 60 workers per condition. We determined the sample size using a prospective power analysis in which we used pilot results from a mixed effects model identical to that reported for the text recall task below to simulate our study design with varying sample sizes. We chose the lowest sample size that provided at least 80% power.

Results

Data Preliminaries

362 workers completed the HIT. Counts per condition were between 59 and 64 as a result of workers completing the HIT after it timed out (6 workers) or pressing the back button, which resulted in failures to record data (4 workers dropped).

Workers completed the HIT in an average of 1404s (σ=688s). Time to completion per condition ranged from 1222s (continuous predict condition) to 1554s (discrete predict condition). The average worker got 2 out of 4 answers correct on the Berlin literacy test (σ=1.4) with no clear differences between treatments (range: 1.78-2.1).

Core Results

Graphical Recall Task: We compare the graphically recalled distributions from participants to the population sampling distribution for the rats experiment. As an initial comparison of participants' response distributions to the population sampling distribution, we examine the average difference between the mean of participants' response distributions and the population mean, and between the standard deviation of the participants' response distributions and the standard deviation of the population sampling distribution. Figs. 4a and b depict density plots for both forms of error by condition. Though less visible for the discrete predict condition, which had greater variance than both other discrete conditions, more participants in the discrete conditions produced distributions with means very close to the true mean. Mean absolute error for the mean of a participant's response distribution ranged from 13.3 to 17.4 (σ: 20.1-32.8) for discrete conditions, and 21.3 to 26.4 (σ: 19.1-25.8) for continuous conditions. The density plots indicate a slight tendency to overestimate the mean among participants in continuous conditions. All participants tended to overestimate variance when recalling the true distribution.

The density plots for the discrete predict group are noticeably more peaked than those in other conditions. This is partially due to a relatively large proportion of participants in this condition that perfectly recalled the population sampling distribution: 12 out of 61 participants. A number of participants (11 out of 59) in the discrete rules training condition and discrete no predict condition (14 out of 57) also perfectly recalled the population sampling distribution, though multiple outliers in these conditions flatten the density curves for both measures. Four other participants across the discrete conditions perfectly recalled the shape of the distribution, but incorrectly positioned the distribution. Many others (21 out of 177) recalled the distribution within 5-10% (i.e., misplacing one or two outcomes). These frequencies suggest that discrete representations that use a small number of outcomes have an advantage for facilitating the encoding of distributional information to memory via shape.

Interestingly, despite undergoing a training task that explicitly provided formulas for calculating the standard deviation of the sampling distribution (standard error) from sample statistics and for allocating density given the standard error, participants in the rules training conditions do as poorly or worse than other conditions in producing distributions that accurately recall the standard deviation.

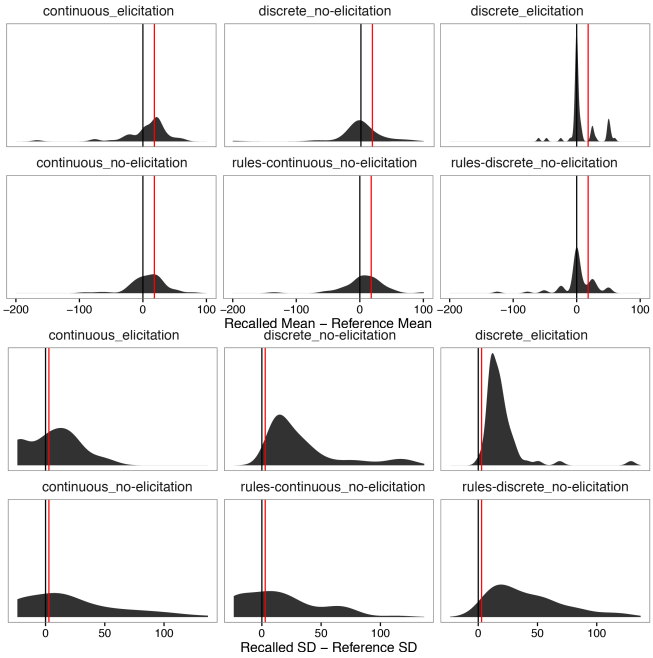

Figure 4. Density plots of the signed error in the mean (top) and standard deviation (bottom) or participants' recalled distributions. The red line indicates expected error if participants recalled properties of the sample sampling distribution.

As a measure of the similarity of participants' response distributions that captures differences in both the location and shape of the distributions, we use KL divergence. As in study 1, we use the same discretized reference distribution for both the discrete and continuous condition participants. We are interested in the mean (location) of each posterior distribution, indicating the normative error for that condition, as well as the variance, indicating how much participants differed from one another in their error rates in that condition.

Specifically, we regress discrete, elicitation, rules, Berlin literacy test score (out of 4) and the interaction between discrete and predict on KL divergence. We define a submodel for the mean and dispersion (standard deviation in log space) of KL divergence. We use a Bayesian model, and build in conservative estimates of the effect and variance in KL for each condition as prior distributions. We specify identical wide Gaussian priors for the effect from each covariate, centered on 0 with a standard deviation of 10. We similarly specify identical Gaussian priors on the standard deviation (in log space) for each effect, centered on 0 with a standard deviation of 5.

We report results as posterior estimates of the effect for each covariate with 95% credibility intervals (the Bayesian analog to a confidence interval) [32] (Fig. 5).

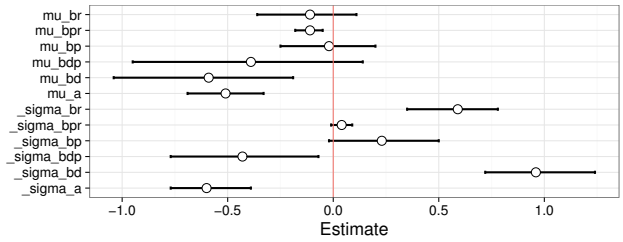

Figure 5. Estimated effects from regressing discrete, elicitation, rules, Berlin literacy score, and predict*elicitation on KL divergence for the graphical recall task. Estimates for σ are modeled in log space.

We see a clear improvement in KL divergence from being in a discrete condition (-0.59, 95% CI: -1.04 - -0.19), in line with H2. We see little main effect of being in an elicitation condition, or being in a rules training condition, in contrast to H1 and H3. Getting a 1 point higher score on the Berlin literacy test correlates with a small but reliable improvement to KL divergence (-0.11, 95% CI: -0.18 - -0.05. We also observe an interaction effect (albeit with high variance) from being in both a discrete and elicitation condition, suggesting that estimating first may be more beneficial if combined with a discrete visualization for some participants.

Being in a discrete, elicitation or rules condition results in higher estimated variance. An exception occurs for the discrete elicitation condition, which has a negative coefficient for variance. This result, combined with the patterns for this condition in the density plots above, suggests that while the benefits of discrete visualizations can vary considerably across participants, elicitation tasks can reduce this variability. It may be that the elicitation task serves to focus participants on the meaning of the representation, so that they are better prepared to benefit from the discrete visualization. Interestingly, the rules training task, while clearly relevant to the task of comprehending a sampling distribution, does not appear to have the same focusing effect.

Textual Recall Task: Mean absolute error in participants' responses to the textual probability questions ranged from 14.7 to 34.8 (σ=20.8-31.8). We model the absolute error in participants' responses using a mixed effects models implemented in glmerstan [**?**]. We specified discrete, elicitation, Berlin literacy test score, rules, and the interaction between discrete and elicitation as fixed effects and subject and question number as random effects, using a maximal random effects structure to allow for varying effects of the literacy score by subject and by question [2].

We report results as posterior estimates of the effect for each covariate with 95% credibility intervals (Fig. 6).

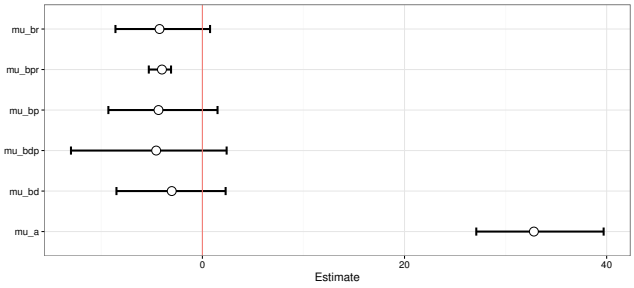

Figure 6. Estimated effects from regressing discrete, elicitation, rules, Berlin literacy score, and predict*elicitation on KL divergence for the textual recall task.

Examining the coefficients indicates that only the Berlin literacy test score reliably predicts lower error (-4.0, CI: -5.4 - -3.1). Being in a discrete, an elicitation, a rules, or a discrete*prediction condition all reduce error but less reliably (-3.0 CI: -8.5 - 2.3; -4.3 CI: -9.3 - 1.6; -4.2 CI: -8.6 - 0.75; -4.6 CI: -13.0 - 2.4).

Graphical Prediction Task: The most accurate prediction a participant could make when asked to estimate the distribution of effects if a new study is replicated many times aligns with the predictive distribution: the sampling distribution for the sample mean and standard deviation (as the population mean and standard deviation cannot be known), adjusted to allow for uncertainty in the location of the mean. To score participants' responses to the graphical prediction task, we constructed the predictive distribution using the Leek method [34]. The resulting distribution is centered on the sample mean, with a standard deviation equivalent to the standard error given the sample standard deviation times the square root of 2 (Fig. 3).

Figs. 7a and b depict density plots of the signed difference between the means and standard deviations of the participants' predicted distributions and the mean and standard deviation of the predictive distribution. Across conditions, the majority of participants underestimated the mean. The density plots indicate bimodality in responses, where some participants in each condition correctly identified the best location on which to center their predicted distribution given the experiment summary (i.e., the reported sample mean) while other participants were less accurate. It is interesting that errors tend to be in

the same direction: upon examining the task interface and data more closely, we hypothesized that many participants may have chosen to locate their predicted distribution near the center of the x-axis range, which ranged from -2.5 to 2.5. Doing so would cause these participants to underestimate the mean of the predictive distribution by approximately 0.7, in line with the pattern in the density plots.

The density plots of signed differences in standard deviation show different patterns by condition. Participants in both elicitation conditions show a tendency to underestimate the standard deviation. The standard deviation of the predictive distribution is equal to the standard deviation of the sample sampling distribution multiplied by the square root of 2 [34]. Hence, participants in the elicitation condition show a bias toward underestimating the standard deviation in the direction that would be expected if their estimates were closer to the sampling distribution for the new study.

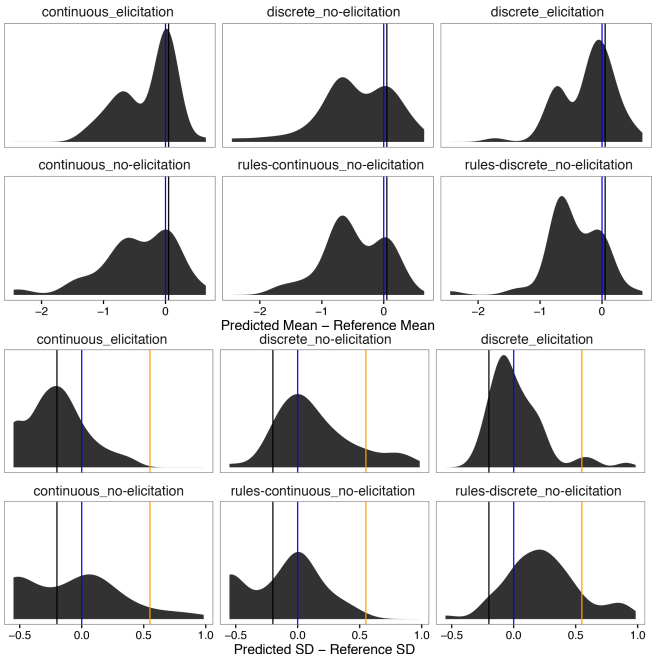

Figure 7. Density plots of the signed error in the mean (top) and standard deviation (bottom) or participants' recalled distributions. The blue line indicates expected error if participants' predictions better matched properties of the predictive distribution (the best possible guess). The black line indicates expected error if participants' predictions better matched properties of the population sampling distribution. The orange line indicates expected error if participants' predictions better matched properties of the data distribution.

We measure the accuracy of participants' predicted distributions for the second experiment description using KL divergence relative to the predictive distribution. We run a similar Bayesian regression to that used for the graphical recall task, using the same fixed and random effects and again using a submodels for the mean and dispersion (standard deviation in log space) of KL divergence. Fig. 8 depicts the estimated effect of each covariate with a 95% credible interval.

From the results we observe that only the estimate elicitation task and score on the Berlin literacy test reliably reduce KL divergence (-0.31 CI: -0.64 - 0.00 and -0.23 CI: -0.31 - -0.15,

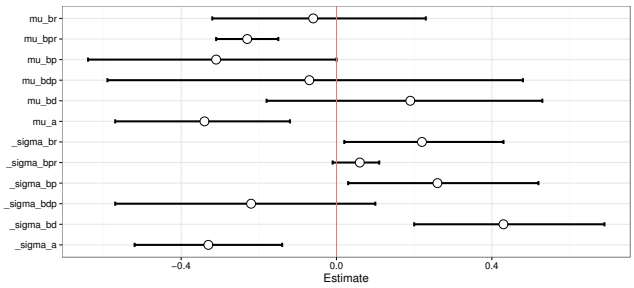

Figure 8. Estimated effects from regressing predict, discrete, rules and predictdiscrete on KL divergence for the graphical prediction task. Estimates for σ are modeled in log space.

respectively). We see a similarly pattern in the estimates for sigma as observed in the graphical recall task, with variance expected to be lower for participants in the discrete*elicitation condition (-0.22 CI: -0.57 - 0.10).

From these results, it is not clear whether participants in the elicitation conditions are in fact estimating the sampling distribution for the study, then adjusting that distribution to account for uncertainty in the population mean. We therefore run identical Bayesian regressions with dependent measures of KL divergence relative to the population sampling distribution for the programming languages study, and to the observed (data) distribution of the study. We observe slightly greater improvement in KL from elicitation and discrete representations, suggesting that participants are not differentiating sampling and predictive distributions (full results available in supplemental material).

DISCUSSION

Our study results provide partial evidence for H1: completing the graphical elicitation task helps participants to make more accurate predictions of the distribution of effects upon replication of a new study in a new domain. However, we do not observe a comparable effect on graphically or textually recalling the probability of different sizes of effects for the study for which participants were shown the population sampling distribution. It is possible that the elicitation task did not result in deeper processing of the populations sampling distribution because it imposed a greater cognitive load than simply viewing the distribution. Or participants may have been distracted by their own estimates. JH: analyze overlap between participants prediction and recall for elicitation conditions

A discrete probability visualization improved participants' recall of the population sampling distribution, providing partial support for H2's prediction that discrete representations help with recall. We suspect that by reducing the distribution to a small number of outcomes that can be remembered via shape, the discrete visualizations present the distribution in a way that is easier to later recall. To our knowledge, this property of discrete visualizations has not been discussed in the literature. However, the fact that better recall is not also seen on the textual recall task may provide evidence that this memorability effect is somewhat superficial, not extending to tasks that require modality translation. We suspect that some participants were able to accurately remember the shape and location of the

distribution but without necessarily understanding its meaning (e.g., each outcome represents 5% of the replications).

We find little support for H3's prediction that providing participants with explicit training on sampling distributions would improve their recall and predictions about a new study. Instead, we see neither a positive nor negative effect of training on performance for any of our three tasks. It is possible that more thorough training, with personalized feedback and multiple practice problems, is needed to see improvements in our statistical reasoning tasks from a rules-based training.

Limitations

Future Work: Applying These Results

Our results suggest a promising alternative to conventional ways of visualizing uncertainty, which tend to inspire passive viewing and often separate the presentation of uncertainty from the data it describes. Though probability distributions and in particular sampling distributions are considered difficult for many novice statisticians [16, 5, 19, 18], our results show that a simple elicitation task, when implemented using an interactive graphical interface, can help focus users on the important properties of a distribution, improving their ability to make predictions about replications of new studies. Our use of a general MTurk sample suggest that even in non-educational or scientific environments that appeal to broad sets of users, elicitation may help users to understand and engage with standard representations of probability distributions.

Simulation of random samples is an orthogonal technique to elicitation that has been found effective for improving reasoning about sampling distributions in educational contexts [16, **?**]. Recently, graphics designers at the New York Times and elsewhere have created discrete-outcome based animated and interactive simulations to support understanding of statistical models [27, 29, 45]. Given our finding that graphical elicitation tasks reduce variance in users' ability to effectively use discrete visualizations, incorporating graphical elicitation of users' estimates for statistical models in interactive simulations could help focus users on what samples represent, and increase reflection on how uncertain processes produce probability distributions.

Future work should whether our elicitation approach could also benefit users who have statistics experience, such as HCI or psychology researchers. If so, our results pave the way for developing new representations of uncertain data aimed at research contexts. For example, interactive visualizations could allow readers of a scientific publication to test their existing knowledge of statistical inference as they read about reported effects and potentially improve at understanding the implications of uncertainty for reported effects, and potential for successful replications, over time.

An important application of our results is for eliciting users' prior expectations for analyses. Bayesian data analysis allows an analyst to specify a prior JH: Finish

Another application of graphical elicitation benefits not just the user, but the owner of the data, who may want to elicit readers' expectations of an uncertain process. For example, graphical elicitation could be used to gather information on naive or expert intuition about the likely result of an experiment, providing an empirical means of evaluating how "surprising" a reported effect is.

CONCLUSION

REFERENCES

1. Gregor Aisch, Amanda Cox, and Kevin Quealy. 2015. You Draw It: How Family Income Predicts ChildrenâĂŹs College Chances. (2015). http://nyti.ms/1BqOX3h
2. Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278.
3. Charles A. Behrens. 1983. Measuring the productivity of computer systems development activities with function points. *IEEE Transactions on Software Engineering* 9, 6 (1983), 648.
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. Dani Ben-Zvi and Joan B Garfield. 2004. *The challenge of developing statistical literacy, reasoning and thinking*. Springer.
6. Charles C Bonwell and James A Eison. 1991. *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports.* ERIC.
7. Shan Carter, Matthew Ericson, David Leonhardt, Bill Marsh, and Kevin Quealy. 2015. Budget Puzzle: You Fix the Budget. (2015). http://www.nytimes.com/interactive/2010/11/13/weekinreview/deficits-graphic.html?_r=0
8. 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.
9. PB Clarke, DAVINA S Fu, ALEXANDER Jakubovic, and HANS C Fibiger. 1988. Evidence that mesolimbic dopaminergic activation underlies the locomotor stimulant action of nicotine in rats. *Journal of Pharmacology and Experimental Therapeutics* 246, 2 (1988), 701–708.
10. George Cobb. 1992. Teaching statistics. *Heeding the call for change: Suggestions for curricular action* 22 (1992), 3–43.
11. M. Correll and M. Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *Visualization and Computer Graphics, IEEE Transactions on* 20, 12 (Dec 2014), 2142–2151. DOI: http://dx.doi.org/10.1109/TVCG.2014.2346298
12. Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *cognition* 58, 1 (1996), 1–73.

13. Richard Cox. 1999. Representation construction, externalised cognition and individual differences. *Learning and instruction* 9, 4 (1999), 343–363.
14. Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
15. Geoff Cumming and Neil Thomason. 1998. Statplay: Multimedia for statistical understanding, in Pereira-Mendoza (ed. In *Proceedings of the Fifth International Conference on Teaching Statistics, ISI*. Citeseer.
16. Robert C delMas, Joan Garfield, and Beth Chance. 1999. A model of classroom research in action: Developing simulation activities to improve studentsâĂŹ statistical reasoning. *Journal of Statistics Education* 7, 3 (1999).
17. Geoffrey T Fong, David H Krantz, and Richard E Nisbett. 1986. The effects of statistical training on thinking about everyday problems. *Cognitive psychology* 18, 3 (1986), 253–292.
18. Joan Garfield. 2002. The challenge of developing statistical reasoning. *Journal of Statistics Education* 10, 3 (2002), 58–69.
19. Joan B Garfield and Iddo Gal. 1999. Assessment and statistics education: Current challenges and directions. *International Statistical Review* 67, 1 (1999), 1–12.
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 and Decision Making* 9, 1 (2014), 1.
23. Ulrich Hoffrage and Gerd Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic medicine* 73, 5 (1998), 538–40.
24. Jon Huang, Albert Sun, and Ford Fessenden. 2015. Who Needs a GPS? A New York Geography Quiz. (2015). http://www.nytimes.com/interactive/2015/03/09/nyregion/nyc-taxi-quiz.html
25. 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).
26. John PA Ioannidis. 2005. Why most published research findings are false. *PLoS Med* 2, 8 (2005), e124.
27. Neil Irwin and Kevin Quealy. 2014. How Not to Be Misled by the Jobs Report. *The New York Times* (May 2014). http://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html
28. Susan Joslyn and Jared LeClerc. 2013. Decisions with uncertainty: the glass half full. *Current Directions in Psychological Science* 22, 4 (2013), 308–315.
29. Josh Katz, Wilson Andrews, and Jeremy Bowers. 2014. Elections 2014: Make Your Own Senate Forecast. (2014). http://nyti.ms/1plfIyv
30. Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016a. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1081–1084.
31. Matthew Kay, Tara Kola, Jessica Hullman, and Sean A Munson. 2016b. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. *Proc. CHI 2016* (2016).
32. John Kruschke. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
33. David M. et al Lane. 2007. Online Statistics Education: A Multimedia Course of Study. (2007). http://onlinestatbook.com/
34. Jeffrey T Leek, Prasad Patil, and Roger D Peng. 2015. A glass half full interpretation of the replicability of psychological science. *arXiv preprint arXiv:1509.08968* (2015).
35. Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. 1977. Calibration of probabilities: The state of the art. In *Decision making and change in human affairs*. Springer, 275–324.
36. Jamie D Mills. 2002. Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education* 10, 1 (2002), 1–20.
37. David S Moore. 1997. New pedagogy and new content: The case of statistics. *International statistical review* 65, 2 (1997), 123–137.
38. Hedwig M Natter and Dianne C Berry. 2005. Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology* 19, 1 (2005), 123–135.
39. Richard E Nisbett, David H Krantz, Christopher Jepson, and Ziva Kunda. 1983. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review* 90, 4 (1983), 339.
40. Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.
41. A. Ottley, E. Peck, L. Harrison, D. Afergan, C. Ziemkiewicz, H. Taylor, P. Han, and R. Chang. 2015. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *Visualization and Computer Graphics, IEEE Transactions on* (2015).

42. Harold Pashler and Eric-Jan Wagenmakers. 2012. EditorsâĂŹ introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science* 7, 6 (2012), 528–530.

43. George J Posner, Kenneth A Strike, Peter W Hewson, and William A Gertzog. 1982. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education* 66, 2 (1982), 211–227.

44. Florian Prinz, Thomas Schlange, and Khusru Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery* 10, 9 (2011), 712–712.

45. Kevin Quealy and Amanda Cox. 2015. The First G.O.P. Debate: WhoâĂŹs In, WhoâĂŹs Out and the Role of Chance. *The New York Times* (July 2015). http://www.nytimes.com/interactive/2015/07/21/upshot/election-2015-the-first-gop-debate-and-the-role-of-chance.html

46. Daniel L Schwartz and Taylor Martin. 2004. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction* 22, 2 (2004), 129–184.

47. Carl James Schwarz and Jason Sutherland. 1997. An on-line workshop using a simple capture-recapture experiment to illustrate the concepts of a sampling distribution. *Journal of Statistics Education* 5, 1 (1997).

48. Peter Sedlmeier. 1999. *Improving statistical reasoning: Theoretical models and practical implications*. Psychology Press.

49. Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* 130, 3 (2001), 380.

50. Susanne Tak, Alexander Toet, and Jan van Erp. 2014. The Perception of Visual Uncertainty Representation by Non-Experts. *Visualization and Computer Graphics, IEEE Transactions on* 20, 6 (2014), 935–943.

51. Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin* 76, 2 (1971), 105.

52. Wikipedia. 2016. 68âĂŞ95âĂŞ99.7 rule. (2016). https://en.wikipedia.org/wiki/68-95-99.7_rule

53. Max L Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. RepliCHI-CHI should be replicating and validating results more: discuss. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 463–466.

54. Max LL Wilson, Paul Resnick, David Coyle, and Ed H Chi. 2013. RepliCHI: the workshop. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 3159–3162.