# Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty

Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha

**Abstract**—People often have erroneous intuitions about the results of uncertain processes, such as scientific experiments. Many uncertainty visualizations assume considerable statistical knowledge, but have been shown to prompt erroneous conclusions even when users possess this knowledge. Active learning approaches been shown to improve statistical reasoning, but are rarely applied in visualizing uncertainty in scientific reports. We present a controlled study to evaluate the impact of an interactive, graphical uncertainty prediction technique for communicating uncertainty in experiment results. Using our technique, users sketch their prediction of the uncertainty in experimental effects prior to viewing the true sampling distribution from an experiment. We find that having a user graphically predict the possible effects from experiment replications is an effective way to improve one's ability to make predictions about replications of new experiments. Additionally, visualizing uncertainty as a set of discrete outcomes, as opposed to a continuous probability distribution, can improve recall of a sampling distribution from a single experiment. Our work has implications for various applications where it is important to elicit peoples' estimates of probability distributions and to communicate uncertainty effectively.

**Index Terms**—Graphical prediction, interactive uncertainty visualization, replication crisis, probability distribution.

## 1 Introduction

There is an increasing interest in using experimental results to *estimate* effects in many scientific fields, including the size of the effect and how reliable it is. This movement runs counter to a more conventional focus on simply *detecting* effects (e.g., through tests for statistical significance), which can encourage overinterpretation of spurious or practically insignificant effects. Visualizations and other data representations are important for helping authors and readers alike to move toward estimation. In particular, a "replication crisis" occurring in many scientific fields [34, 52, 54] suggests the need for new ways to help users think through *replication uncertainty*—the expected distribution of effect sizes if an experiment were run again.

As an example, imagine a typical results report from a controlled experiment. The author describes an observed effect, say a mean *reduction* of 19 mm Hg in the systolic blood pressure of a sample of 10 heart attack survivors who were given a new drug, compared to 10 heart attack survivors who were not. The author also describes uncertainty around the effect, say a *confidence interval* from 8 to 30 mm Hg around the mean effect. Understanding replication uncertainty means being able to reason about how often potential replications of the study would see an effect of at least the same size, half the size, etc. Whether the audience consists of lay people presented with study results by the media, or scientists and other experts consulting findings in scholarly publications, accounting for replication uncertainty is a critical part of interpreting science [34, 52, 54].

Unfortunately, many typical uncertainty representations, like error bars, make it easy to ignore or misinterpret uncertainty [4, 36]. For example, many scientists misinterpret a 95% CI as indicating a region in which the mean of a replication is expected to fall 95% of the time [30]. As an alternative to requiring explicit training on statistical rules, asking a person to represent or predict information can be powerful ways to improve statistical reasoning through active learning [6, 57]. For example, asking people to graph statistical information like the risk associated with a disease in a discrete (frequency) format can lead to more accurate probability inferences [15, 49, 60]. 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 [19]. Though typically associated with educational contexts, active learning strategies are used to elicit user predictions about statistical models characterizing uncertain processes in recent interactive visualizations in the media [1, 9, 32, 38]. The act of predicting, for example by predict the party majority in voting outcomes for various states [38], may prompt deeper consideration of assumptions affecting the outcomes and the meaning of the visualized data [41].

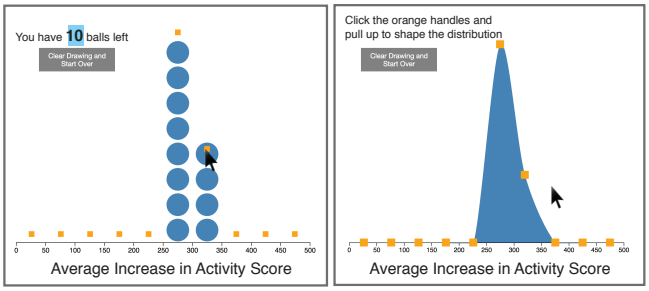



Fig. 1. Discrete and continuous elicitation interface used by participants in our study to predict replication uncertainty.

In this paper, we ask, Can *graphically predicting* uncertainty in scientific experiment results improve one's ability to recall the reliability of those findings, and to make predictions about the reliability of new experiments' results? We examine whether non-statisticians, who are most likely to misunderstand experimental uncertainty, can be helped by graphical prediction. Our primary contribution is a controlled study used to **show how outcomes related to a user's awareness of replication uncertainty are impacted by being asked to predict the uncertainty first using an interactive visualization**. We find that users who graphically predict replication uncertainty and see their prediction against the true sampling distribution in one experiment can more accurately complete a *transfer task* in which they must estimate replication uncertainty for a new experiment.

Our second contribution is to **identify the impact of different visual representations of probability on how well a user can recall and estimate how uncertainty impacts reported effects**. Considerable prior work shows that frequency formats for probability informa-


- *Jessica Hullman is with the University of Washington 1. E-mail: jhullman@uw.edu.*
- *Matthew Kay is with University of Michigan 2. E-mail: mjskay@umich.edu.*
- *Yea-Seul Kim is with University of Washington 3. E-mail: yeaseul1@uw.edu.*
- *Samana Shrestha is with Vassar College 4. E-mail: sashrestha@vassar.edu.*


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

tion, but outside of classic Bayesian reasoning problems [8, 21, 22, 45, 60], few studies examine the effects of using frequency formats to visualize uncertainty. We evaluate the difference between discrete-outcome visualizations (Fig. 1 left) versus continuous visualizations (Fig. 1 right) of a probability distribution. We find that while discrete outcome visualizations do not necessarily improve a user's ability to estimate uncertainty in the effect of a new study, users who interact with visualizations comprised of a small number of outcomes (20) can better recall the uncertainty in a reported effect from an experiment. This suggests that discrete visualizations of probability distributions with limited numbers of outcomes may provide a useful format for remembering statistical information among non-experts.

To enable the empirical contributions of our study, we first conducted a design space exploration of interactive visualization interfaces for predicting uncertainty. Our design space exploration develops and evaluates twelve interfaces that allow users to "draw" probability distributions, including both discrete and continuous visualization approaches, extending prior work in probability elicitation [28, 51]. We summarize high level guidelines for graphical prediction interfaces for distributions from our results.

Our results demonstrate new possibilities for communicating uncertainty in experimental science more effectively. They suggest the power of graphical prediction and discrete-outcome visualizations for *interactive uncertainty visualization approaches* that incorporate the user's prior knowledge to improve lay understanding of the reliability of statistical results. This is a crucial enterprise for any interested in improving public understanding of—and trust in—science.

Our results also pave the way for future research exploring the potential for interactive uncertainty visualizations to improve expert reasoning about experimental uncertainty, goals of the transparent statistics [39] and RepliCHI movements [66, 67]. Finally, our results have impliciations for graphical probability elicitation for Bayesian data analysis and other applications.

## 2 Background & Hypothesis Development

We summarize probability distributions that can be used to characterize replication uncertainty. To generate hypotheses, we survey research in two related areas: (1) teaching statistical reasoning, and (2) interacting with uncertainty visualizations.

### 2.1 Statistical Background: Four Distributions

Inference from experimental data involves understanding subtle differences between characterizations of uncertainty. Four different probability distributions can be used to describe the uncertainty in the mean observed effect from an experiment. First, by running an experiment, a scientist seeks to infer the **population distribution**, the true distribution of the values that a variable can take on in a population of individuals (Fig. 2.1). The **true sampling distribution**, the true distribution of a statistic obtained through drawing all possible samples of a given size from the population, models expected uncertainty due to the sampling process (Fig. 2.3).

With perfect knowledge of the world, we could exactly specify the population (true) mean and true sampling distribution. The **observed sampling distribution** (Fig. 2.6) and the and the **replication prediction distribution** (Fig. 2.7) represent our best guesses for the true mean and true sampling distribution respectively, based on the *imperfect knowledge* that can be obtained through experimentation. The observed sampling distribution assumes that the sample mean is representative of the true mean, while the replication prediction distribution accounts for the fact that the sample mean will not be equal to the true mean. Fig. 2 describes the typical usage of these distributions in experimental science.

### 2.2 Tools for Teaching Statistical Reasoning

How to teach people to make inferences about various types of probability distributions, as well as randomness and sampling in general [24, 23] is a primary focus in statistics education. One approach to teaching statistical reasoning claims that explicitly training people on statistical rules, like the law of large numbers, can improve their inferences [20, 50]. However, reformers of statistical pedagogy have argued that encouraging active reasoning is more beneficial than rule memorization [23], and proposed alternative techniques focused on active learning through analysis and simulation [5, 47, 59].

Sampling distributions (Fig. 2.3, Fig. 2.6), are notoriously difficult for people to reason about. Researchers have documented common misinterpretations among students (e.g., *the sampling distribution should look like the population distribution*) [11]. Others have studied errors made by experts when interpreting statistical representations based on sampling distributions [4, 30]. One active learning approach to teaching statistics advocates using simulations to improve understanding of sampling distributions [11, 46]: e.g. letting a user specify a sample size and population distribution and observe the sampling distribution [18, 19, 58]. The implication is that actively interacting with the process that produces the distribution improves students' abilities to reason about distributions in general (i.e., *transfer effect*).

*Prediction* may be a particularly beneficial form of interaction for understanding uncertainty. DelMas et al., for example, find that a simulation was not enough to leave students with accurate conceptions of sampling distributions [19]. Based on a belief that contradictory evidence is required to change one's beliefs [53], the researchers developed a prediction-based activity. Students first made estimates about population distributions in pre-test problems, then ran a simulation on the same distributions. Students who did the pre-test plus simulation showed additional gains on posttest questions relative to those who did not first make guesses about the distributions. While not applied to understanding uncertainty, Kim et al. [41] find that predicting data prior to viewing a visualization, and viewing the gap between one's predictions and the visualized data, can improve one's short term recall of the data. Motivated by these results, we examine whether making predictions about uncertainty (i.e., probability distributions) using interactive visualization interfaces can enhance statistical reasoning, including the ability to transfer one's understanding of uncertainty in one setting to a new setting.

### 2.3 Visualization Interactions to Understand Probability

Prior work in visualizing uncertainty has tested various representations of probability distributions with novices, indicating various errors in reading the visualizations (e.g., [14, 33, 40, 65]). Common representations of probability information like error bars have also been shown to lead to erroneous predictions among experts who have been trained to read them [4, 30]. For example, researchers and other "experts" frequently assume that the boundaries of a 95% confidence interval are meaningful, such that the population mean would fall within the boundaries 95% of the time if the study were replicated [30]. In actuality, the confidence level 95% is restricted to describing confidence in the procedure (if the study were replicated, 95% of the time the confidence interval would contain the population mean). However, the majority of existing uncertainty visualization studies focus on how well users can *read* probability information rather than how well they can *reason* with the information (e.g., by making predictions about the implications of uncertainty in presented data or new data, which we refer to as a *transfer effect*).

Actively *constructing* visual representations may be a more powerful strategy for helping people understand uncertain processes. Natter and Berry [49] found that participants who were asked to portray the size of a risk on a bar chart were more accurate and satisfied in their probability estimates. Cosmides and Tooby [15] presented people with the base rate of a disease and false positive rate of a test 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. Sedlmeier and Gigerenzer [60] conducted a controlled study of tutorial programs for Bayesian problems. Programs that instructed users on how to create frequency representations of probability information led to better short-term and long-term Bayesian reasoning compared to a rule training program. More recent studies indicate that graphical interfaces are advantageous for eliciting a person's subjective probability distribution [28, 27, 62]. We evaluate whether having a user "draw" their

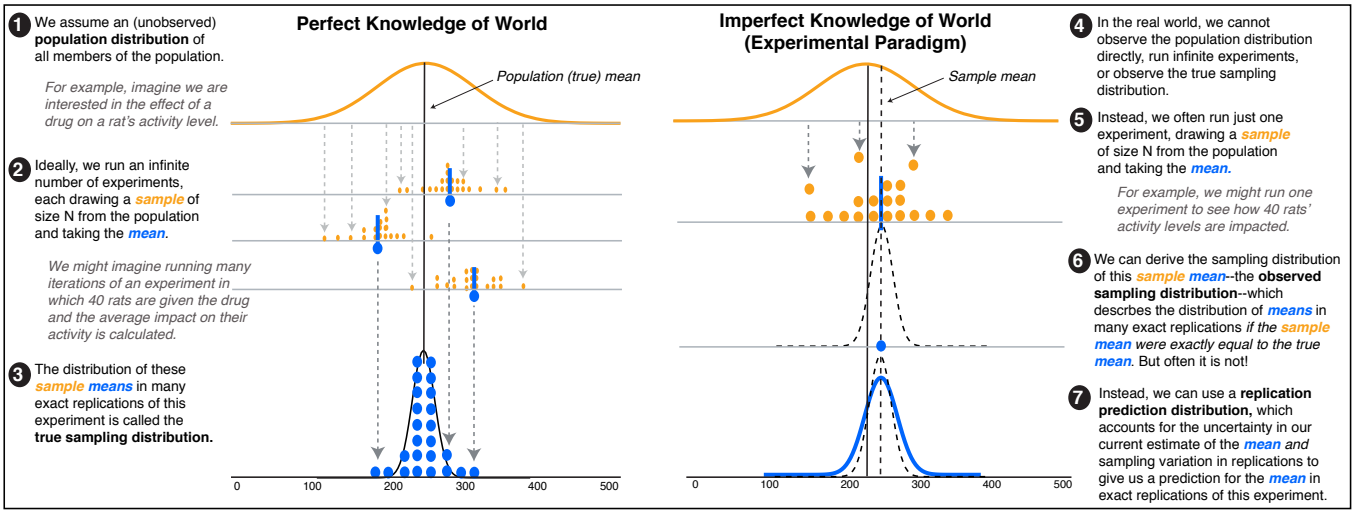



Fig. 2. A depiction of distributions relevant to replication uncertainty, including those based on perfect knowledge of the world (left) and those derived from samples obtained in experimentation (right).

estimate of a probability distribution using an interactive visualization interface helps them understand uncertainty in experiment results. As a preliminary design space exploration, we contribute a number of discrete outcome and continuous visualization interfaces for prediction.

Other visualization interactions that are thought to help uncertainty comprehension do not require that the user construct a visualization. For instance, a large body of research indicates that simply interacting with probabilities presented in discrete, *frequency formats* (e.g., 7 out of 10) is easier than cognitively processing the same information in a probability format (e.g., 70%) [26, 29]). However, the vast majority of this work addresses classic Bayesian reasoning tasks, which involve a narrow range of tasks (e.g., identifying false positives for a test given conditional probabilities, e.g., [8, 21, 22, 45, 60]). Some recent work applies this "discrete visualization advantage" hypothesis to the presentation of univariate probability distributions, finding that discrete-outcome visualizations can lead to better inferences particularly when small numbers of outcomes are used [33, 40]. Again, however, these works focus on how well users can read probabilities, rather than how well they can transfer probabilistic reasoning to new data sets. We examine whether discrete visualizations of probability distributions help users understand uncertainty in experiment results in terms of recall and transfer ability.

## 2.4 Formulating Study Conditions & Hypotheses

Several prior studies point to the benefits of active reasoning tasks for understanding statistical concepts. In particular, active reasoning in the form of *prediction tasks* and tasks that require *constructing visualizations* appears promising. Hence we test how well users who complete a *graphical prediction task* in which they graphically specify a probability distribution can recall and predict uncertainty in experiment results compared to users who do not use graphical prediction.

The prior work also suggests that presenting discrete-outcome visualizations can help people express and reason about probability. However, discrete formats also reduce precision in communicating a distribution. We are interested in whether a discrete visualization that summarizes a distribution using frequency can have benefits for recalling or reasoning about uncertainty in potential replications of an experiment. Hence we vary whether a *discrete-outcome* versus a *continuous* visualization of a probability distribution is used by subjects in our study. Crossing graphical prediction with visualization results in four conditions:

- **Graphical Prediction - Discrete**: The user predicts the sampling distribution for a presented experiment using a discrete visualization prior to seeing the true sampling distribution.
- **Graphical Prediction - Continuous**: The user predicts the sampling distribution for a presented experiment using a continuous visualization prior to seeing the true sampling distribution.
- **Baseline - Discrete**: The user views the true sampling distribution using a discrete visualization.
- **Baseline - Continuous**: The user views the true sampling distribution using a continuous visualization.

Graphical prediction represents a form of *implicit* training on how to interpret the sampling distribution that is eventually visualized for the user, since the user is not formally trained on how the sampling distribution relates to the (observed) sample data. Instead, the act of prediction is likely to draw their attention to where their initial guess was wrong, prompting active reasoning. However, some researchers have advocated *explicitly* training users on the rules related to statistical processes like sampling distributions [20, 50]. We include conditions in which users complete a rule training task on sampling distributions as a means of comparing an alternative interactive training task to graphical prediction:

- **Rule Training - Discrete**: The user is presented with information about sampling distributions and a hypothetical distribution. She is asked to calculate the sampling distribution standard deviation using an interactive form. She then views the true sampling distribution using a discrete visualization.
- **Rule Training - Continuous**: The user completes the same training task. She then views the true sampling distribution using a continuous visualization.

### 2.4.1 Hypotheses

Informed by the prior work, we consider how our study conditions will affect several proxies for a user's awareness of uncertainty: the user's ability to *recall* the uncertainty in a presented experiment's results (which is likely to reflect related factors like their attention [17] and depth of processing [48] of the uncertainty) and the user's ability to *transfer* what they inferred about replication uncertainty to make predictions about a new experiment.

- **H1 (Graphical prediction effect)**: Graphically predicting what will happen if an experiment is replicated prior to seeing the true sampling distribution will lead to more accurate *recall* of that distribution and improved accuracy in estimating the replication uncertainty for a new experiment in a *transfer task*.
- **H2 (Discrete visualization effect)**: Viewing the true sampling distribution for an experiment using a discrete representation will lead to more accurate *recall* of that distribution and improved accuracy in estimating the replication uncertainty for a new experiment in a *transfer task*.
- **H3 (Rules training effect)**: Completing an explicit instructional task about sampling distributions will lead to more accurate *recall* of that distribution and improved accuracy in estimating the replication uncertainty for a new experiment in a *transfer task*.

## 3 Preliminary Study: Interface Selection

Only a handful of prior works demonstrate interfaces that enable a user to sketch a distribution (e.g., [27, 28]. We therefore conducted a design space exploration of graphical prediction interfaces in order to select the discrete and continuous interface for our study.

### 3.1 Design Development and Iteration

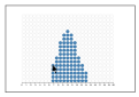
We aimed to develop designs that exhibited three properties we believed would make graphical prediction interfaces effective and accessible to novices: they should *require little training*, *be expressive enough to capture users' intuitions about probability*, and *encourage accurate reasoning about probability*.

These properties focused our design space exploration. The goal of requiring *little training* led us to create interfaces that rely on direct manipulation to reduce abstractness. The goal of *expressiveness* is motivated by delMas's proposal that evidence contradictory to one's beliefs is needed to motivate a change in beliefs when learning [53]. If a user can express their honest best guess, they may be more likely to benefit from seeing their prediction against a true distribution than if the interface forces symmetry and other normative properties of common distribution types (e.g., Gaussian). Finally, the need for an interface to *encourage accurate reasoning about probability* inspired our development of multiple prediction interfaces that use discrete outcome visualization, based on evidence that frequency formats can improve statistical reasoning [26, 31]

To further organize our design space exploration, we first brainstormed *interaction techniques* (clicking, brushing, dragging to select outcomes), *selection scopes for an interaction* (apply to the specific outcome that was interacted with, or apply to all outcomes below, i.e., filling down a column of a discrete-outcome histogram), and *default states for outcomes* (outcomes appear only when interactions are triggered, or outcomes appear by default but their positions must be modified). We combined these three factors where it was feasible (e.g., continuous interfaces lack outcomes and therefore have only one selection scope and default state). We also varied the number of outcomes used in discrete-outcome interfaces, to explore the trade-off between fewer outcomes, which require fewer interactions to build a distribution and have been shown in prior work to be easier for participants to read [40] and relate to [7], and more outcomes, which increase expressiveness.

#### 3.1.1 Discrete Outcome Visualization Interfaces

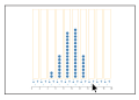
**W** e developed two types of *paint-outcomes-by-dragging* interfaces. Both types first present the user with a grid of empty circles representing outcomes. A standard paint-by-dragging interface allows users to fill in circles with color by dragging the mouse over each. For all discrete interfaces (with the exception of the rolling-balls interfaces), a warning message appeared to the left of the drawing area when the user reached the total number of outcomes. All subsequently added outcomes turn red until the total number being used is within the limit.



**A** s a more efficient version of the paint-by-dragging interaction, we also created 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 below it in the column, with either 20 or 100 outcomes.

**W** e developed a *pull-up* interface that allows the user to drag up handles that are equally spaced along the *x*-axis to add outcomes. The user is given either 20 or 100 total outcomes, with 10 or 20 handles along the *x*-axis.

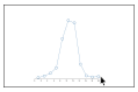
**R** ather than requiring the user to add or fill outcomes, the *rolling-balls* interface first presents a uniform distribution of outcomes. The user drags the 20 (shown left) or 100 outcomes between bins to form the distribution.



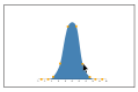
**F** inally, we also implemented two versions of one of the few distribution sketching interfaces evaluated in prior work: the *balls-and-bins* interface evaluated in [28]. We created a 20 ball version and a 50 ball version to investigate the impact of the number of outcomes. The user clicks an up arrow (△) or down arrow (▽) below each of 10 bins to either add one or remove one outcome (ball) at a time to that bin.

We used a 50 ball version rather than a 100 ball version as in [28], because early users found it to require too much clicking.

#### 3.1.2 Continuous Visualization Interfaces



**W** e created several continuous probability interfaces to explore a second potential trade-off, between the more familiar representation of a probability distributions using a continuous format (e.g., a density plot) and the reasoning benefits of discrete formats. 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 a probability density function (pdf). As the line is created, 11 handles equally-spaced along the *x*-dimension are added to the line. The user can later adjust the shape of the curve by dragging the handles.



**W** e developed a *continuous-pull-up* interface, in which a user is presented with 10 equally spaced handles along the *x*-axis. The user creates a pdf by dragging up the handles. The interface smooths the curves the user draws using cardinal spline fitting.

We did not label *y*-axes with probability values in the interfaces, based on evidence that thinking about relative probabilities is easier for people than thinking about absolute probabilities [51].

### 3.2 Evaluations with Users

We evaluated the 12 graphical prediction interfaces first by asking volunteers in our labs to try the interfaces using think-aloud protocols. After further design iteration we deployed an online survey on Mechanical Turk (MTurk). MTurk samples can provide comparable quality to university or other online samples but tend to be more demographically diverse [10]. We recruited 80 U.S. based workers with approval ratings of 95% or above to help ensure quality [63]. Each worker was paid $2.00 for their participation.

Our goal was to identify general patterns in preferences and how well people could use different interfaces, to inform the choice of discrete and continuous interface for our study. Each participant was asked to use six total interfaces to reproduce the same reference distribution (a normal probability density function with with $\mu$=10 and $\sigma$=2). Each assigned interface appeared on a separate screen, with the reference distribution pictures to the left (Fig. 3). The participant was asked to replicate the reference distribution as best she could using each interface. Interfaces were randomly assigned and counterbalanced, but to aid comparison, each participant was always assigned both the high resolution (e.g., 50 or 100 outcome) and low resolution (20 outcome) version of the same technique (e.g., balls-and-bins, rolling-balls, etc.). We also paired the two continuous probability interfaces.

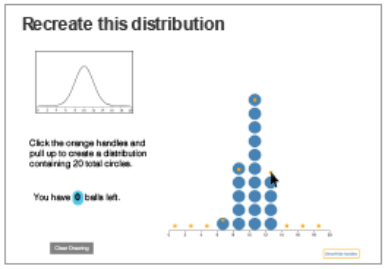



Fig. 3. Interface used in evaluating graphical prediction techniques (pull-up 20 outcomes shown).

In addition to gathering text feedback, we also quantitatively measured participants' performance with the interfaces by capturing **response time**, how long (in sec) the user spent replicating the distribution; **accuracy**, how close the user's drawing of the distribution was to the reference distribution as log Kullback-Leibler (KL) divergence, an information theoretic measure of the similarity between two distributions [42][1]; and **satisfaction**, how satisfied the user was with using the interface to replicate the distribution (obtained using a 100 pt

[1]We used a discrete version of the reference distribution for scoring; we show in supplemental material that this calculation is not distinguishable from reasonable alternatives.

slider from *Not satisfied* to *Very satisfied* after completion of all six task screens).

#### 3.2.1 Results

As our design space study was conducted primarily to inform the designs used in our reasoning study, we report high-level results here only and refer the reader to the supplemental material for further details [2]. The trends we report are based on participants' text comments and the three dependent measures we collected.

**Continuous probability interfaces outperform discrete**. Overall, participants were reliably more satisfied with the continuous interfaces compared to discrete interfaces we tested (increase of 20/100 points, 95% PI: [8,33]). They may also be faster in many cases (taking 0.71× the time, 95% PI: [0.48×, 1.02×]) and more accurate (-0.52 change in log(KL), 95% PI: [-1.22, 0.23]). [3]

**Less outcomes reduces time to draw distribution**. For discrete interfaces, we tend to observe lower response times with fewer outcomes (50- and 100-outcome interfaces take 1.6× the time of 20-outcome interfaces, 95% PI: [1.2×, 2.1×]).

**More outcomes does not lead to higher accuracy**. We assumed that more outcomes would enable more accurate replications of the reference distribution; however, our results do not support this. We see comparable accuracy with more outcomes (change in log(KL) of -0.11, 95% PI: [-0.68, 0.53]).

**Pull-up discrete interfaces lead to higher accuracy**. Both pull-up discrete interfaces performed relatively well in terms of response times, satisfaction, and accuracy. To compare the pull-up interfaces to the other discrete interfaces, we ran a mixed effects model implemented in glmer2stan [43] for each of the three measures. Mixed effects models are commonly used to account for repeated measures from the same subject, which is specified as a random effect. We specified interface type and order as fixed effects. We specified the balls-and-bins 50 outcomes interface as the reference group. We find that the pull-up 20 outcomes interface performs similarly to the balls-and-bins 50 from the prior work, both of which are the most accurate discrete interfaces (though not reliably so; difference of log(KL) of -0.04, 95% PI: [-0.37, 0.29]).

These results suggest that the continuous pull-up interface is the more effective of the continuous interfaces; we therefore select this as the continuous interface in our study. For our discrete interface, while the pull-up interface performs comparably to the balls-and-bins 50 interface, the discrete pull-up interface more closely mirrors the interaction of the continuous pull up interface. We therefore select the pull-up 20 interface in order to have the fewest differences between conditions (apart from the discrete or continuous representation).

## 4 Study Design: Reasoning About Uncertainty

We conduct a study to test how well users who complete a *graphical prediction task* can recall and estimate uncertainty in experiment results compared to users who do not use graphical prediction (H1). We vary whether a *discrete-outcome* versus a *continuous* visualization of a probability distribution is used by participants in our study (H2). We also include a *rules training* condition as an alternative form of explicit instruction on sampling distributions (H3). Fig. 4 shows the six conditions.

| | | Training | | |
|---|---|---|---|---|
| | | None | Implicit | Explicit |
| Representation | Discrete | Discrete None | Discrete Predict | Discrete Rules |
| | Continuous | Continuous None | Continuous Predict | Continuous Rules |

Fig. 4. Study conditions.

Our study presents all participants with a description of a single instance of a hypothetical experiment, in the form of sample statistics (mean, standard deviation, number of observations). All participants see the same instance of the hypothetical experiment, which we sample from a larger set of experimental results that we simulated (Fig. 5).

### 4.1 Measuring Understanding of Replication Uncertainty

As a first measure of participants' awareness of uncertainty, we consider how accurately the user of a experimental report can recall a reported effect with uncertainty. Recall of the effect and uncertainty is likely to impact how they incorporate the findings into future judgments (e.g., of related studies, or in daily life). Specifically, we examine how accurately a user can recall a presented sampling distribution. A **graphical recall task** asks participants 'What would happen if this experiment were replicated many times?' Re-creating the *true sampling distribution* that she was shown is the most accurate response a participant could give. All participants use an interface that matches the visualization format with which they viewed the true sampling distribution (discrete or continuous).

We also inquire about the same distribution by asking participants to complete a **text recall task** consisting of text probability questions. Accurately answering text probability questions provides evidence of the strength of a participant's mental representation of the true sampling distribution, as translating information between modalities requires one to appropriately abstract properties of the original representation [16].

Another measure of participants' understanding lies in their ability to transfer what they have learned about uncertainty to new situations. We designed a **transfer task** that describes an instance of a second experiment in a different domain and asks participants 'What would happen if this study were replicated many times?' All participants use a graphical interface (either discrete or continuous, depending on their condition) to predict this distribution. This task is best described as *near case transfer* based on the similar format for the presentation of the second experiment (e.g., short narrative description with sample statistics) [56]. However, participants will perform better on this task when they abstract the relationship between sample statistics and a sampling distribution from the first experiment presentation, so the transfer task also represents one proxy for general learning about distributions. The *replication prediction distribution* (Fig. 2) represents the best prediction a participant could make in the transfer task given only a sample mean and standard deviation; because the sample statistics are unlikely to exactly represent the population parameters, the replication prediction distribution takes into account the fact that the true population mean will not be exactly the sample mean, and so is wider than the *observed (or sample-based) sampling distribution*, which would not be a well-calibrated prediction. We derive the replication prediction distribution after the method described by Spence *et al.* [64] for producing predictive intervals for replications. This distribution is equivalent to a Bayesian replication prediction distribution for the sample mean in an exact replication, if an uninformed prior were used for the analysis of the first experiment.

The study procedure is shown in Fig. 5. We interleave the presentations of experimental results and tasks to make sure that the recall task is sufficiently challenging: participants read about the second experiment in between viewing the true sampling distribution for the first experiment and doing the recall task. Those in the Baseline condition read the experiment information and, as a "placebo" interaction, retype the sample statistics prior to viewing the true sampling distribution. To control for differences between participants' levels of prior familiarity with interactive visualization, we include an initial training screen for all participants which provides basic instructions on how to draw a distribution with the graphical interface that they are assigned. Finally, we also utilize the Berlin numeracy test so that we can account for the effect of participants' a priori statistical literacy levels [13]. All experimental stimuli are available in supplemental materials.

### 4.2 Study Stimuli: Experiment Domains

Our study presented participants with information about two fictional scientific experiments from two domains: animal behavior (A) and

[2] Supplemental material is also at: https:github.com/jhullmanuw/imagining_replications_infovis2017 DOI: 10.5281/zenodo.836886

[3] We report percentile intervals, or quantile credibility intervals, a Bayesian analogue to a confidence interval [44].

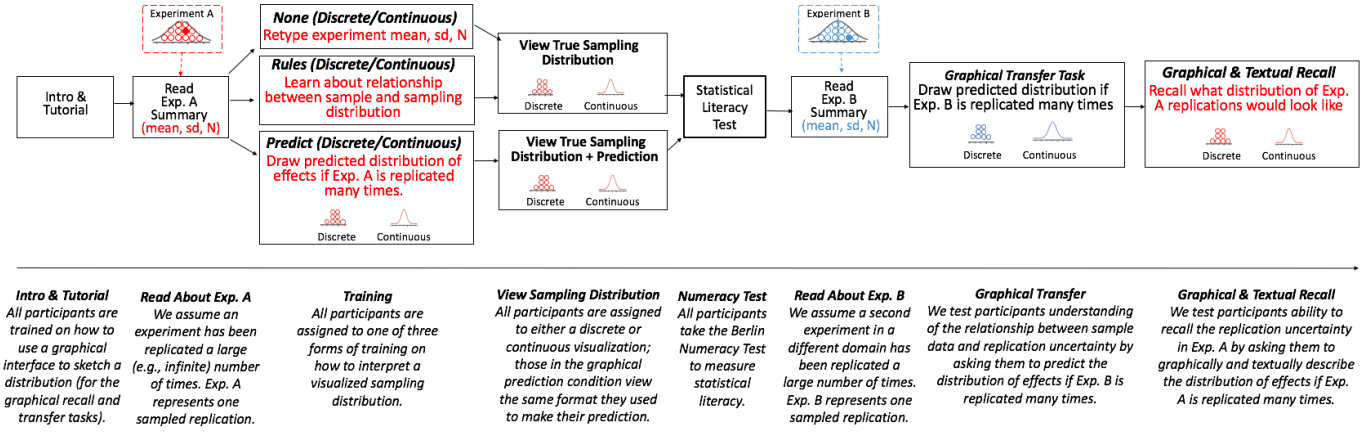



Fig. 5. Depiction of study procedure. Participants are assigned to one of three types of training in how to read a visualization of a sampling distribution (Baseline, explicit training on rules for deriving a sampling distribution, and implicit training via graphical prediction), and one of two visualization formats (discrete and continuous). All participants are later tested on their understanding of replication uncertainty in a transfer and a graphical and text recall task.

computer science (B). Both experiments were based on actual experiments ([12] and [3], respectively). For the rat activity experiment (A), 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 (B), participants were presented with the mean difference in the productivity of 8 engineers who used two programming languages consecutively in a within-subjects design.

Both experiments described elaborate domain-specific measures (of activity in rats, and of productivity in software engineers). We intentionally chose these measures to reduce the chance that participants would apply prior knowledge in any of the tasks.[4]

### 4.3 Participant Population

Because we wanted to study how our interventions affect a general sample of non-statisticians, we posted the study as a single HIT on Amazon's Mechanical Turk, open to U.S. workers with an approval rating of 95% or above. Workers could only participate once. 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 numeracy test questions and 8 text recall questions and an additional $0.50 each for the recall and transfer tasks if the elicited distribution had small (0.30 pr less) KL divergence with the appropriate correct distribution. The average bonus earned was $0.80.

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

## 5 Study Results: Reasoning About Uncertainty

### 5.1 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 ($\sigma$=688s). Time to completion per condition ranged from 1222s (continuous predict) to 1554s (discrete predict). The average worker got 2 out of 4 answers correct on the Berlin numeracy test ($\sigma$=1.4; range: 1.78-2.1), with a distribution comparable to prior statistical literacy benchmarks on AMT [13]. Full demographics are reported in supplemental material.

[4]Future work might examine how prior knowledge impacts replication predictions relative to (say) a Bayesian norm, but we wished to leave out the impact of expert knowledge in this initial work.

### 5.2 Analysis Methods

We analyze the data using several approaches. As a primary modeling approach for both the recall and transfer tasks in which participants draw distributions, we use Bayesian implementations of mixed effects linear regressions in the *rethinking* package for R. To quantify the difference between the participant's response and the correct distribution in the recall and transfer tasks, we again use KL divergence, which accounts for differences in the location and shape of the distributions. Rather than running only a single model to examine the *mean error* (in log KL divergence) of a participant's distribution relative to the correct distribution, we leverage the flexibility of the Bayesian implementation to run two-part models that differentiate mean error ($\beta$; overall how accurate are participants' response distributions by condition?) and dispersion ($\gamma$; is there more variance between participants' accuracy in some conditions?). Both mean effect and dispersion are important for understanding the potential for the different conditions we tested to improve statistical reasoning: a large effect is desirable (larger $\beta$), but so is a reliable effect across participants (lower $\gamma$).

The first submodel we run for the recall and transfer tasks regresses the mean effect ($\beta$ coefficients) in KL divergence on dummy variables denoting discrete visualization, graphical prediction task, and rules training. We include the score on the Berlin numeracy test and the interaction between discrete visualization and graphical prediction. We mean center the Berlin numeracy test score to improve interpretability.

The second submodel address variance levels in conditions by regressing the dispersion ($\gamma$ coefficients) of each effect ($\beta$ coefficients) in log space on the same set of variables. This submodel allows us to examine whether some conditions result in more varied behavior between participants.

For both submodels, we use the same discretized reference distribution for both the discrete and continuous conditions to calculate KLD. This choice avoids penalizing users of discrete visualizations for the lower amount of precision a discrete interface affords. However, this choice may bias results in favor of discrete visualization users. We confirmed all main effects that we report are robust to multiple alternative KLD calculation methods, including changes to the format of the of the reference distribution and response distributions (see supplemental material).

Using Bayesian models also enables us to build in prior expectations for the effects we examine. We build in weakly-informed priors of mean effects and dispersion for each condition. We specify identical Gaussian priors centered on 0 for each effect ($\beta$; standard deviation of 5), and for each dispersion ($\gamma$; standard deviation of 2.5).

We report results as the distribution of posterior estimates of each effect for each submodel (Fig. 7A and Fig. 9A; violin plots depict these distributions). All effects are relative to a participant in the Baseline Continuous condition with an average score on the Berlin numeracy

test. We avoid reporting $p$-values and instead interpret the posterior estimates by looking at the degree to which 95% Percentile Intervals (reported in text) overlap with 0 (indicating the possibility of no effect). We also use the posterior estimates to derive the expected cell means (i.e., the expected mean and standard deviation of the log KL error for each condition) and plot these values in Fig. 7B and Fig. 9B violin plots.

Because specific differences in KL divergence can be difficult to understand on an intuitive level, Fig 6 depicts how log KL is affected by location and variance. We also present distribution plots of the bias in the mean and the standard deviation of participant's response distributions in the recall (Fig. 7C and D density plots) and transfer tasks (Fig. 9C and D density plots). These plots provide a more concrete look at how much participants in different conditions overestimated versus underestimated the mean and standard deviation of the correct distribution.

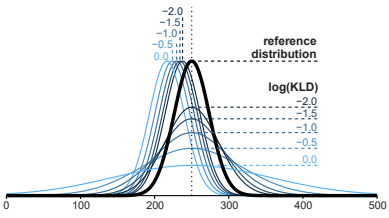



Fig. 6. To aid interpretation of effect sizes in this paper, this plot shows log(KL divergence) for distributions with varying means and SDs relative to a reference distribution. Distributions closer to the reference distribution have lower log(KLD).

To analyze the text recall results, where each participant answers multiple questions, we use a mixed effects model in glmer2stan [43]. We regress the absolute error in participants' responses (reported as probabilities) on the same set of dummy variables (fixed effects) – discrete visualization, graphical prediction task, and rules training–as well as the mean-centered Berlin numeracy test score and the interaction between discrete and predict as fixed effects. We include subject and question number as random effects. We plot posterior estimates of effects in Fig. 8A and expected means by condition in Fig. 8B.

### 5.2.1 Graphical Recall Task

We compare the graphically recalled distributions from participants to the true sampling distribution for the first hypothetical experiment involving rat activity levels. The true sampling distribution conveys the population parameters; a participant cannot be more accurate than to provide this distribution when asked about the distribution of replication effects. We are interested in the mean of each posterior distribution for $\beta$, indicating the normative error for that condition, as well as the variance ($\gamma$), indicating how much participants differed from one another in their error rates in that condition. Fig. 7A presents the distributions of posterior estimates for the mean effect ($\beta$) and dispersion ($\gamma$). Fig. 7B presents the distributions of expected values of the effect and dispersion by condition. To determine whether apparent effects are reliable, we look to whether the 95% PIs for the estimates (not pictured) overlap with 0.

We see a *clear improvement in KL divergence from being in a discrete condition* ($\beta$: -0.59, 95% PI: [-1.04, -0.19]), in line with H2. We see *little effect of being in a graphical prediction condition, or being in a rules training condition*, in contrast to H1 and H3. Getting a 1 point higher score on the Berlin numeracy test correlates with a small but reliable improvement to KL divergence ($\beta$: -0.11, 95% PI: [-0.18, -0.05]). We observe a highly variable interaction effect from being in both a discrete and graphical prediction condition ($\beta$: -0.41, 95% PI: [-0.83, 0]), suggesting that for some, predicting may be more beneficial combined with a discrete-outcome visualization.

Being in a discrete, predict or rules condition results in higher estimated variance ($\beta$: 0.95, 0.23, 0.59; 95% PI: [0.74, 1.14], [0.03, 0.42], [0.41, 0.76], respectively). Being in the discrete predict condition lowers variance ($\beta$: -0.42, 95% PI: [-0.77, -0.15]); however, the practical implications of this reduction in variance are questionable given the higher variance from being in either discrete or predict.

To gain further insight into how participants' response distributions

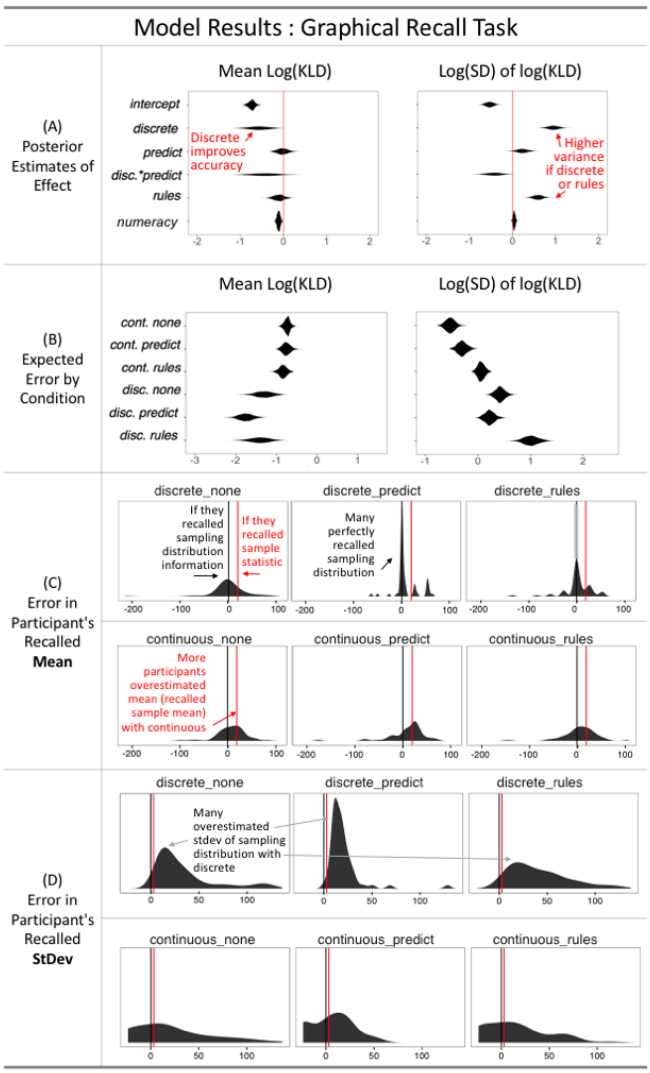



Fig. 7. Results of the graphical recall task. Violin plots depict the distributions of posterior estimates of effects (A) and expected effects by condition (B), where error is measured as log KL divergence. Density plots (C, D) compare the means (C) and standard deviations (D) of participants' recalled distributions to those of the true sampling distribution (black line at 0) and sample distribution (red line).

differed from the true sampling distribution, we examine the average difference between the means and standard deviations of participants' response distributions and the true sampling distribution. Fig. 7C and D depicts 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 population (true) mean. Expected absolute error for the mean of a participant's response distribution ranged from 13.3 to 17.4 ($\sigma$: 20.1-32.8) for discrete conditions, and 21.3 to 26.4 ($\sigma$: 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 true sampling distribution: 19.7% of 61 participants. A number of participants (18.6% of 59) in the discrete rules training condition and discrete no predict condition (24.6% of 57) also perfectly recalled the true sampling distribution, though outliers flatten these

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 (11.9% of 177) recalled the distribution with one or two outcomes misplaced. These results 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*.

On the other hand, despite undergoing a training task that explicitly provided formulas for calculating the standard deviation of the sampling distribution and for allocating density given this value, participants in the rules training conditions do as poorly or worse than other conditions in accurately recalling the standard deviation.

### 5.2.2 Text Recall Task

As expected, the text recall task resulted in noisier estimates than graphical recall. Mean absolute error in participants' responses to the text probability questions ranged from 14.7 to 34.8 ($\sigma$=20.8-31.8).

Examining the posterior estimates of mean effect and dispersion (Fig. 8 bottom), *only the Berlin numeracy test score reliably predicts lower error* (-4.19, 95% PI: [-5.33, -3.05]). Being in a discrete, predict, rules, or a discrete*predict condition may also reduce error, but these estimates are not reliable.

### 5.2.3 Graphical Transfer Task

To score participants' responses to the graphical transfer task, we calculate KLD of their response relative to a replication prediction distribution, as this distribution infers the population parameters given the sample statistics that are provided, but allows for the fact that those statistics may not capture the population parameters. We construct this distribution using the method outlined by Spence *et al.* [64]. Because the replications are assumed to have the same sample size, given the mean ($M_1$), standard error ($SE_1$), and sample size ($N$) of the first study, the replication prediction distribution for the mean in a replication ($M_2$) is $M_2 \sim \text{StudentT}(M_1, SE_1\sqrt{2}, N-1)$.

From the results (Fig. 9A) we observe that *only the graphical prediction task and score on the Berlin numeracy test reliably reduce KL divergence* (-0.31 95% PI: [-0.64, -0.03] and -0.23 95% PI: [-0.31, -0.15], respectively). Using a discrete visualization, on the other hand, results in worse performance (0.34 95% PI: [0.07, 0.64]). We see the same pattern of estimates for sigma as observed in the graphical recall task, with discrete, predict, and rules increasing variance, while discrete and predict offsets the increase by lowering variance (Fig. 9B).

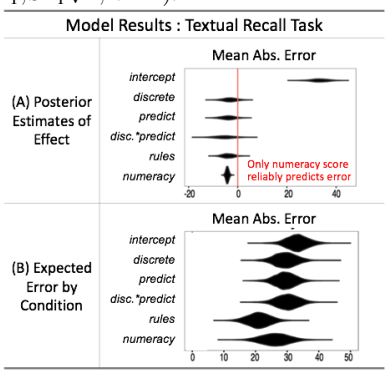



Fig. 8. Results of the text recall task. Violin plots depict the distributions of posterior estimates of effects (A) and expected effects by condition (B). All text questions asked participants to report probabilities; error represents the deviation of these probabilities from the true probabilities according to the true sampling distribution.

To better understand which of the relevant distributions (Fig. 2) participants' predictions most resemble, we ran identical Bayesian regressions with dependent measures of KL divergence relative to the true sampling distribution for the programming languages study, and to the observed (data) distribution of the study. We observe slightly greater improvement in KL from prediction and discrete representations against both alternative reference distributions, suggesting that participants are not differentiating sampling and replication prediction distributions (results available in supplemental material).

Fig. 9 also depicts density plots of the signed difference between the means (Fig. 9C) and standard deviations (Fig. 9D) of the participants' predicted distributions and the Spence *et al.* replication prediction distribution. Across conditions, most 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 (i.e., the reported sample mean) while other participants did not. It is notable that errors tend to be in the same direction. Upon examining the data and interface more closely, we suspect 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 replication prediction distribution by about 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 graphical prediction conditions show a tendency to underestimate the standard deviation. The standard deviation of the replication prediction distribution is greater than that of the observed sampling distribution by a factor of at least $\sqrt{2}$. Hence, participants in the graphical prediction 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.

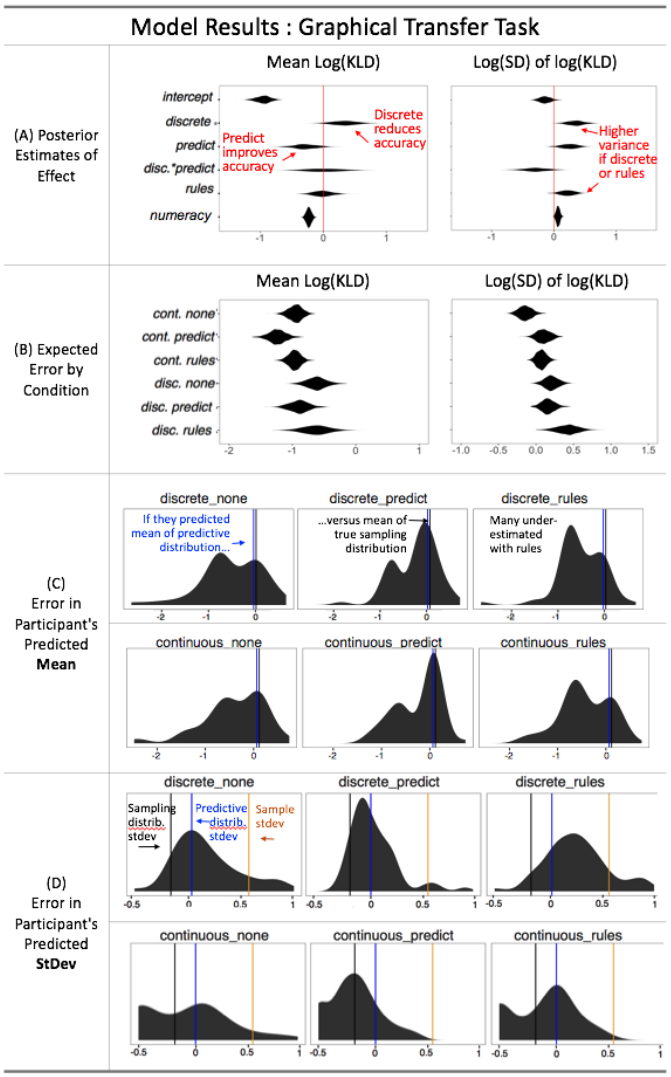



Fig. 9. Results of the graphical transfer task. Violin plots depict the distributions of posterior estimates of effects (A) and expected effects by condition (B), where error is measured as log KL divergence. Density plots (C, D) compare the means (C) and standard deviations (D) of participants' predicted distributions in the transfer task to those of the replication prediction distribution (blue line at 0) as well as the true sampling distribution (black line) and sample (orange line).

## 6 Discussion and Future Work

Our design space exploration and informal evaluation of graphical prediction interfaces indicated that **participants tend to prefer and produce accurate distributions more quickly with continuous visualizations of probability distributions**. However, our controlled experiment on statistical reasoning instead points to advantages of discrete-outcome visualizations for recall of a distribution, adding to the body of work indicating reasoning benefits of discrete visualizations for Bayesian reasoning [26, 31, 60], specifying population distributions [28], and probability extraction [33, 40]. A discrete-outcome visualization improved participants' graphical recall of the population sampling distribution, providing partial support for H2. We suspect that **discrete-outcome visualizations are advantageous for recall because they reduce the distribution to a small number of outcomes that can be remembered via shape**. To our knowledge, this property of discrete-outcome visualizations has not been discussed in the literature. However, the fact that better recall is not also seen on the text recall task suggests that this memorability effect may be 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., that each outcome represents 5% of replications).

Discrete visualizations did not help, however, with predicting replication uncertainty of a new study in a new domain. Instead, using discrete visualizations led to worse accuracy in our participants' estimates. It is possible that the high variance in performance with discrete visualizations that we observed both in the recall and transfer tasks helps explain this result: some participants may not have understood the meaning of the individual outcomes.

Our study results provide partial evidence for H1: **graphical prediction *before* seeing the true replication uncertainty in one study leads to more accurate predictions of replication uncertainty of a new study in a new domain**. Again, however, the implications of this result are complex. While reliable, the effect was variable between participants; some participants may not have benefited. Additionally, we saw no clear comparative advantage for prediction on *recalling* the true sampling distribution. Only discrete-outcome visualizations appear to help with recall. It is possible that the memory advantage of discrete-outcome visualizations dominates any effect of prediction on recall. We do observe that **graphical prediction, when combined with discrete-outcome visualization, reduces variance in users' performance for recall and transfer tasks**, though the practical implications of this effect are difficult to reason about. It is possible that the prediction task focuses participants on the meaning of the representation, so that they benefit slightly more from a discrete format.

We find no evidence that explicit training on sampling distributions improves recall nor transfer (H3). It is possible that more thorough training, with personalized feedback and multiple practice problems, is needed to see improvements.

#### 6.0.1 Limitations

Our study examined whether there are benefits to drawing a distribution and viewing discrete-outcome visualizations in the form of short term recall improvements and more accurate predictions of the replication prediction distribution for a new task. Future work should explore longer term recall benefits, and corroborate the rather variable prediction effect through replications.

Additionally, our transfer task design may have emphasized the sample statistics over the experimental description by preventing participants from returning to the description as they made their prediction. Future work should explore how graphical prediction and discrete visualizations impact new predictions that require more qualitative assessment of experimental features, as well as other "far case" transfer tasks [56].

While we accounted for general statistical reasoning ability (as measured by the Berlin test) in our analysis, our work did not examine specific differences in how effective graphical prediction tasks or discrete-outcome visualizations are for helping people of different prior experience levels to make more accurate judgments. Additionally, spatial abilities, which we did not measure, are known to be correlated with mathematical ability (e.g., [2]) and may influence use of a visualization interface for probability distributions.

### 6.1 Future Work: Graphical Prediction Applications

#### 6.1.1 For Engaging Visualization Users with Uncertainty

Our use of an MTurk sample suggests that even outside of educational contexts, graphical prediction may benefit non-expert users of probability representations by engaging them to think more carefully about uncertainty. For example, media reports on scientific results often reference readers' prior knowledge to engage their interest [61]. However, cuing readers to think about their own everyday contexts can motivate the use of heuristics over analytical reasoning [37] as cited in [61], making readers more likely to overlook uncertainty in reported studies. Asking users to make a prediction may provide a means of leveraging the natural curiosity stimulated by engaging one's prior knowledge while emphasizing uncertainty. Simulations of uncertain processes have also appeared in popular interactives as a means of making complex sampling processes more understandable [35, 38, 55]. The focusing effect of graphical prediction may help simulation users understand what samples represent and how uncertain processes produce probability distributions.

To help realize these applications, future work should explore the larger design space of interactions with uncertainty representations. For example, even oft-misunderstood representations such as error bars representing confidence intervals may be better understood if users are first given a chance to predict their length. Uncertainty-related predictions could also be "gamified," such that users receive real or hypothetical rewards based on the accuracy of their predictions.

#### 6.1.2 For Improving Scientific Reliability in HCI and Beyond

Future work should examine whether graphical prediction of uncertainty can benefit users who have statistics experience, such as HCI researchers, in line with the goals of the RepliCHI [66, 67] and transparent statistics movements [39]. We suspect that many researchers could benefit from graphical predictions as a way to focus more attention on uncertainty in their own and others' studies. For example, readers of scientific publications could use interactive visualizations to test their statistical understanding as they read about reported effects. Because making a prediction requires some prior knowledge, graphical predictions may help nudge scientists towards considering the weight of evidence in the literature (e.g. through meta-analysis); Ioninidis has argued that scientists' overinterpretation of the significance of single studies has contributed to the replication crisis [34].

#### 6.1.3 For Elicitation from Experts and Others

How to best elicit priors from experts, such as for Bayesian analyses where prior expectations are critical [25], remains an open problem for which few graphical interfaces have been evaluated [51]. The results of our interface study can inform further development of interactive prior elicitation mechanisms. Our reasoning study with non-experts suggests graphical prediction interfaces could also have value for evaluating statistical literacy. For example, whether a user constructs a symmetric distribution, the moments of their distribution, and their error over multiple prediction exercises could provide educators with valuable information about probability distribution literacy.

## 7 Conclusion

We evaluated a novel graphical prediction technique that may help people grasp uncertainty in experiment replications. Graphically predicting the replicability of an experimental effect led to more accurate predictions of the replication uncertainty for a new study in a different domain. We also found new benefits of discrete visualizations of probability: for improving recall of a probability distribution. Our results motivate new applications in presenting uncertainty in ways that work toward helping the general public better understand—and know when to trust—scientific experiments.

## REFERENCES


[1] G. Aisch, A. Cox, and K. Quealy. You draw it: How family income predicts children's college chances, 2015.

[2] M. T. Battista. Spatial visualization and gender differences in high school geometry. *Journal for research in mathematics education*, pages 47–60, 1990.

[3] C. A. Behrens. Measuring the productivity of computer systems development activities with function points. *IEEE Transactions on Software Engineering*, 9(6):648, 1983.

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

[5] D. Ben-Zvi and J. B. Garfield. *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 2004.

[6] C. C. Bonwell and J. A. Eison. *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports.* ERIC, 1991.

[7] G. L. Brase. Which statistical formats facilitate what decisions? the perception and influence of different statistical information formats. *Journal of Behavioral Decision Making*, 15(5):381–401, 2002.

[8] G. L. Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3):369–381, 2009.

[9] S. Carter, M. Ericson, D. Leonhardt, B. Marsh, and K. Quealy. Budget puzzle: You fix the budget, 2015.

[10] K. Casler, L. Bickel, and E. Hackett. Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6):2156–2160, 2013.

[11] B. Chance, R. del Mas, and J. Garfield. Reasoning about sampling distribitions. In *The challenge of developing statistical literacy, reasoning and thinking*, pages 295–323. Springer, 2004.

[12] P. Clarke, D. S. Fu, A. Jakubovic, and H. C. Fibiger. Evidence that mesolimbic dopaminergic activation underlies the locomotor stimulant action of nicotine in rats. *Journal of Pharmacology and Experimental Therapeutics*, 246(2):701–708, 1988.

[13] E. T. Cokely, M. Galesic, E. Schulz, S. Ghazal, and R. Garcia-Retamero. Measuring risk literacy: The berlin numeracy test. *Judgment and Decision Making*, 7(1):25, 2012.

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

[15] L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

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

[17] F. I. Craik, M. Naveh-Benjamin, G. Ishaik, and N. D. Anderson. Divided attention during encoding and retrieval: differential control effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1744, 2000.

[18] G. Cumming and N. Thomason. Statplay: Multimedia for statistical understanding, in pereira-mendoza (ed. In *Proceedings of the Fifth International Conference on Teaching Statistics, ISI*. Citeseer, 1998.

[19] R. C. delMas, J. Garfield, and B. Chance. A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, 7(3), 1999.

[20] G. T. Fong, D. H. Krantz, and R. E. Nisbett. The effects of statistical training on thinking about everyday problems. *Cognitive psychology*, 18(3):253–292, 1986.

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

[22] R. Garcia-Retamero and U. Hoffrage. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, 83:27–33, 2013.

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

[24] J. B. Garfield and I. Gal. Assessment and statistics education: Current challenges and directions. *International Statistical Review*, 67(1):1–12, 1999.

[25] A. Gelman and D. Weakliem. Of beauty, sex, and power. *American Scientist*, 97, 2009.

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

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

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

[29] R. Hastie and R. M. Dawes. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage, 2010.

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

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

[32] J. Huang, A. Sun, and F. Fessenden. Who needs a gps? a new york geography quiz, 2015.

[33] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11), 2015.

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

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

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

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

[38] J. Katz, W. Andrews, and J. Bowers. Elections 2014: Make your own senate forecast, 2014.

[39] M. Kay, S. Haroz, S. Guha, and P. Dragicevic. Special interest group on transparent statistics in hci. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1081–1084. ACM, 2016.

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

[41] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017.

[42] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[43] R. McElreath. glmer2stan (r package), 2014.

[44] R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*, volume 122. CRC Press, 2016.

[45] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2536–2545, 2012.

[46] J. D. Mills. Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1):1–20, 2002.

[47] D. S. Moore. New pedagogy and new content: The case of statistics. *International statistical review*, 65(2):123–137, 1997.

[48] M. Moscovitch and F. I. Craik. Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior*, 15(4):447–458, 1976.

[49] H. M. Natter and D. C. Berry. Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology*, 19(1):123–135, 2005.

[50] R. E. Nisbett, D. H. Krantz, C. Jepson, and Z. Kunda. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4):339, 1983.

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

[52] H. Pashler and E.-J. Wagenmakers. Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530, 2012.

[53] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education*, 66(2):211–227, 1982.

[54] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[55] K. Quealy and A. Cox. The first g.o.p. debate: Who's in, who's out and the role of chance. *The New York Times*, July 2015.

[56] D. Schunk. *Learning Theories: An Educational Perspective*. Pearson, 2004.

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

[58] C. J. Schwarz and J. Sutherland. 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.

[59] P. Sedlmeier. *Improving statistical reasoning: Theoretical models and practical implications*. Psychology Press, 1999.

[60] P. Sedlmeier and G. Gigerenzer. Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3):380, 2001.

[61] P. Shah, A. Michal, A. Ibrahim, R. Rhodes, and F. Rodriguez. Chapter seven-what makes everyday scientific reasoning so challenging? *Psychology of Learning and Motivation*, 66:251–299, 2017.

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

[63] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 275–284. ACM, 2011.

[64] J. R. Spence and D. J. Stanley. Prediction interval: What to expect when you're expecting a replication. *PloS one*, 11(9):e0162874, 2016.

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

[66] M. L. Wilson, W. Mackay, E. Chi, M. Bernstein, D. Russell, and H. Thimbleby. Replichi-chi should be replicating and validating results more: discuss. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 463–466. ACM, 2011.

[67] M. L. Wilson, P. Resnick, D. Coyle, and E. H. Chi. Replichi: the workshop. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 3159–3162. ACM, 2013.