Evaluating the Use of Uncertainty Visualisations for Imputations of Data Missing At Random in Scatterplots

Abhraneel Sarma, Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du, Eunyee Koh, and Matthew Kay

Abstract— Most real-world datasets contain missing values yet most exploratory data analysis (EDA) systems only support visualising data points with complete cases. This omission may potentially lead the user to biased analyses and insights. Imputation techniques can help estimate the value of a missing data point, but introduces additional uncertainty. In this work, we investigate the effects of visualising imputed values in charts using different ways of representing data imputations and imputation uncertainty—*no imputation*, *mean*, *95% confidence intervals*, *probability density plots*, *gradient intervals*, and *hypothetical outcome plots*. We focus on scatterplots, which is a commonly used chart type, and conduct a crowdsourced study with 202 participants. We measure users' bias and precision in performing two tasks—estimating average and detecting trend—and their self-reported confidence in performing these tasks. Our results suggest that, when estimating averages, uncertainty representations may reduce bias but at the cost of decreasing precision. When estimating trend, only hypothetical outcome plots may lead to a small probability of reducing bias while increasing precision. Participants in every uncertainty representation were less certain about their response when compared to the baseline. The findings point towards potential trade-offs in using uncertainty encodings for datasets with a large number of missing values. This paper and the associated analysis materials are available at: `https://osf.io/q4y5r/`

Index Terms—Uncertainty visualisations, missing values, data imputation, multivariate data

1 Introduction

Data quality issues are a persistent problem for visual analytic systems, and missing values are one of the most common causes of imperfect data [30]. The presence of missing data points can make it challenging for analysts to interpret and derive insights from the data, using visual analysis tools or otherwise. Fig. 1 shows an example of the impact missing values can have on inference—the incomplete dataset may lead to an analyst underestimating the trend in the data. Current visual analytic systems do not provide much support to users for handling missing values in their analysis. For instance, Tableau either indicates the number of dropped cases that are not represented in the graph (default) or allows the user to represent them as zero values. While excluding missing values during the analysis phase (also known as complete-case analysis) may be appropriate in certain conditions [33], it requires observations to be *missing completely at random*, which is not often the case. In fact, missing completely at random is a fairly strict but often unrealistic assumption about the data [33, 40]. If data is missing completely at random, dropping all observations with missing values from the incomplete dataset can be equivalent to a smaller but still complete dataset [33]. However, when assumptions for complete-case analysis do not hold, visualisation systems do not provide sufficient safeguards for users against the potential pitfalls of making erroneous inferences from datasets that contain missing values.

In statistics, a common approach for dealing with incomplete data is to impute missing values, which can result in more reliable inferences [33]. Recent work has studied how users interpret visualisations of incomplete datasets, when missing values are highlighted [2, 15, 41] or imputed [41]. These studies have found that highlighting the presence of missing values were preferred by participants and improved perceived data quality, while imputing missing values may lead to improvement in task performance. However, the mechanism under which missing values occurred in the datasets used as stimuli in prior work was *missing completely at random*. While prior work [41] did not find any meaningful effect of different visualisation methods, such as error bars, for representing imputed values on task performance, this may not stand true for other missingness mechanisms. Moreover, prior work [2, 15, 41] has exclusively focused on investigating the effects of missing data on very specific missingness pattern (univariate missingness where missing values are confined to a single variable), data types (time-series), and chart type (line graphs and bar graphs only), which leaves the effects of showing imputations for general multivariate datasets with distinct missingness patterns and chart types not known.

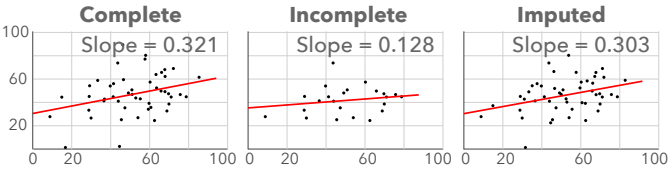


Fig. 1. The impact of missing values on trend estimation

However, imputation methods cannot precisely calculate the value of a missing observation, instead providing an estimate with some degree of uncertainty. Thus, how uncertainty is encoded in visualisation can be another important factor affecting participants' ability to make inferences from charts. Prior work has only used error bars to show the uncertainty of the imputed values [40,41] but common static representations such as error bars or confidence bands can be confusing or difficult to interpret [10, 21], leading to the proposal of numerous alternative uncertainty representations in recent visualisation literature such as hypothetical outcome plots [23, 27], probability density functions and its variants [10, 20, 25], which provide greater distributional information to the viewer. These techniques have been found to have varying degrees of effectiveness both in improving statistical reasoning as well as the quality of decisions in relevant tasks [23, 26, 27, 29]. However, these studies focused on decision making based on a single represented probability distribution or comparison between two represented probability distributions. Oftentimes, incomplete datasets contain multiple missing values which when represented with uncertainty information, will require users to process several probability distributions within a larger visualization at the same time. The effect of uncertainty representations in such scenarios remains not known from prior work.

In this study, we investigate the effect of six different ways of representing data imputations and imputation uncertainty (§3.2) on two visual analytics tasks—estimating average and detecting trend—with scatterplots (§3.3). The datasets used as stimuli were *missing at random* and *trivariate*, with missing values occurring in two of the dimensions (§3.1), which relaxes some of the constraints on missingness mecha-

- *Abhraneel Sarma and Matthew Kay are with Northwestern University. E-mail: abhraneel, mjskay@u.northwestern.edu.*
- *Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du and Eunyee Koh are with Adobe Research. E-mail: sguo, jhoffs, rrossi, fdu, eunyee@adobe.com.*

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

nism and data types from previous studies. We measure performance for each task using *bias*—presence of systematic errors—and *precision*—consistency in participants responses, and also record participants' self-reported confidence in their responses.

In the *average estimation* task, when compared to a baseline of no imputation, we find that all uncertainty representations reduced bias (desirable), but may also reduce precision (undesirable). Further, using *confidence intervals* to represent uncertainty may even result in unbiased, albeit somewhat more inconsistent, responses on average. In the *trend estimation* task, we find the effects to be inconsistent across the two proportions of missing data tested—showing only *mean* of imputations may have a small effect on reducing bias at low proportion of missing data, but may likely not reduce bias at higher proportion of missing data; on the other hand, while the *hypothetical outcome plots* may not reduce bias at lower proportion of missing data, it may likely reduce bias at higher proportion of missing data.

Our results highlight potential trade-offs that various uncertainty representations offer when visualising imputations of missing data in scatterplots, and the importance of taking tasks into account. Our primary contributions are as follows:

- Results of a **pre-registered empirical study** with 202 participants on the effect of six different representations of imputations and imputation uncertainty on scatterplots when performing two visual analytic tasks.
- A discussion on the **trade-offs of the six representations** on participants' bias and precision in performing the two visual analytic tasks.

2 Related Work

It is unclear how visualisation systems should represent datasets with missing values to support inference [51]. Visualisation systems such as Wrangler [28], xGobi [43,44], VIM [46], and MANET [48] made some initial attempts to support inspecting the presence of, and visualising imputations of missing values. xGobi [43,44] and MANET [48] placed significant emphasis on dealing with missing values during interactive graphical analysis. MANET [48] attempted to make users aware of the presence of missing values by making it easier to keep track of them—in charts such as scatterplots, where one of the two variables are missing, they visualise points for which information for one-dimension is available by plotting them along a special axis; in charts such as histograms, they augment the chart with an additional bar representing the frequency of missing values. xGobi [43, 44] provided users with functions to keep track of missing values, inspect missing value patterns, and impute missing values using different imputation methods. Graphical exploration of imputations methods were implemented to enable users to examine and compare precomputed imputations.

Despite their promise, neither of these two tools are commonly used today. Even though these solutions considered missing values during the EDA process, they did not investigate how imputed values should be represented to best support users in making inferences from the data. Most widely used visual analytics systems currently, such as Tableau or PowerBI, do not support imputation of missing data, and very few studies have investigated how analysts may benefit from representing imputed values and visualizing their uncertainty.

2.1 Theory of Missing Data

To ground the discussion of missing values and methods for treating them, we first provide a brief overview of the relevant statistical techniques and terminology. We consider tabular data structures where each row represents a unique observation, and each column represents a unique variable. Two factors which impact imputation and representation of values are missingness *mechanism* and missingness *pattern*.

Missingness Mechanism describes the relationship between the variables that are missing and the other variables in the dataset:

1. *Missing Completely at Random (MCAR):* the observations which are missing are a random subset of all observations [6]; that means missingness does not depend on the values of either the missing or observed data. In such cases, dropping all observations with missing values, also known as complete case analysis, is often appropriate and the result will be equivalent to the analysis being performed on a smaller dataset.
2. *Missing at Random (MAR):* there are "systemic differences between the missing and observed cases, but these can be entirely explained by other observed variables" [6]. In other words, missingness depends only on the observed variables in the dataset.
3. *Missing Not at Random (MNAR):* there are systemic differences between the missing and observed cases which cannot be entirely explained by the observed variables.

Prior work has primarily investigated data which is MCAR [40, 41]. However we omit this condition as it is a fairly strict assumption for most real-world data, and complete case analysis is often appropriate for such data, making imputations somewhat redundant. We also omit MNAR as such datasets require more complex imputation methods [33], and were deemed beyond the scope of the current study. In this work, we focus on data which is MAR.

Missingness Pattern describes which variables are missing and which are observed in a dataset [33], as well as indicating whether certain groups of variables have values missing together. Prior work [2, 15, 41] has focused on univariate missing data where missing values are confined to a single variable. We extend prior work by investigating missingness in two variables within a trivariate dataset. There can be different types of missingness patterns even though only two variables contain missing values (Fig. 2B and C). We restrict our study to the missingness pattern depicted in Fig. 2B and consider data with only one missing value per observation (see §3.1 and §3.2 for more details).

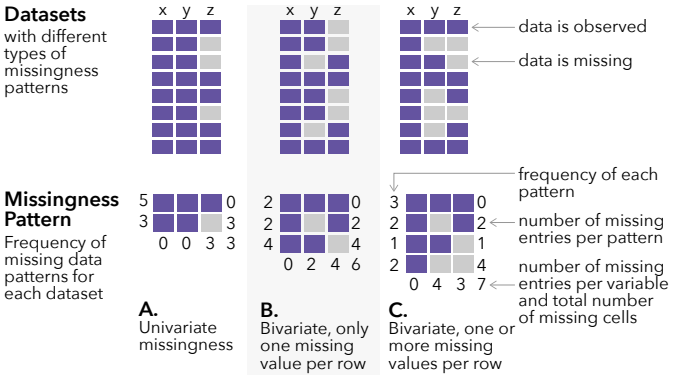


Fig. 2. Different types of missingness patterns

2.2 Evaluating visualisations of data with missing values

Recent work has investigated how different visual representations of data with missing values can impact the perception of data quality [15, 41], accuracy of decisions [15, 41], and confidence in decisions [2, 41]. Eaton et al. [15] explore the effect of zero-filling, omitting or highlighting missing values in a dataset. Andreasson and Riveiro explore omission, omission along with explanation, and "fuzziness"—but it is unclear whether any imputation was performed [2].

Song and Szafir [41] compare a wider range of approaches (highlighting or downplaying presence of missing values, annotating values with error bars, and omitting information). Complementary to the visual representations, missing data values were imputed using three different techniques (zero-filling, marginal mean, and linear interpolation). They evaluated participants' performance using two tasks—average and trend detection—which were operationalised as binary choice questions. They found that two of the imputation methods—linear interpolation and marginal means—improved accuracy, but did not find any effect of visualisation technique on accuracy. As the underlying dataset was *missing completely at random*, the presence of missing values should not have affected users' performance. Their results suggest that the

presence of misleading information such as zero-filling can have an adverse effect on the user. In a subsequent study, Song et al. [40] compared the effect of visualising imputations with error bars in a scatterplot to no imputation. They found that showing imputations with error bars had an effect on users' decision-making workflow but did not have a statistically significant effect on mental demand.

Most of these studies [2, 15, 41] looked at line and bar charts, depicting a very specific type of data (time-series); while one prior work did consider scatterplots [40] their focus was on the exploratory data analysis workflow and participants' mental demand, instead of participants performance in visualisation tasks. Although a few studies have used uncertainty representations to show imputed estimates [40,41], the design space was restricted to only error bars, leaving a large number of uncertainty visualisations, proposed in recent work, unexplored. Our study aims to fill this gap by investigating user performance of the two visual analytics tasks—average and trend estimation—in scatterplots with various types of uncertainty representations.

2.3 Communicating Imputation & Uncertainty Information

Different methods of conveying uncertainty can impact task performance. Error bars and box plots, two commonly used representations, encode summary statistics as marks, hiding distributional information from the user, and are designed with the goal of cognitive efficiency [8, 22, 32, 47]. However, they may not be ideal for many tasks [10, 26], may result in the misinterpretation of the statistic being encoded [5, 10, 21], and may even require greater cognitive effort compared to representations that provide distributional information [9]

Probability density function (PDF) plots, and its variants such as violin plots [20] and gradient plots [10, 25] allow presentation of univariate distributional information, with the design goal of providing complete information about the underlying probability distribution to the viewer. Hypothetical outcome plots (HOPs) [23, 27] use animations to convey distribution information, with each frame consisting of a draw from an underlying distribution. Uncertainty visualisations which communicate distributional information have been found to improve accuracy in statistical reasoning [23, 29], as well as performance in certain decision-making tasks [9, 17, 26, 27].

Most prior work on uncertainty visualisations requires users to consider one or two probability distributions for performing the statistical or decision-making task. Greis et al. [18] investigated how a viewer might integrate uncertainty from two sources in an average estimation task, and found that participants were able to weigh information more accurately when distributional information was provided. However, most incomplete datasets can have a large number of missing observations. An imputation method will estimate, with uncertainty, values for each missing observation. Thus, in such a scenario, users will have to consider and aggregate information from multiple distributions. In our study, we compare five types of uncertainty visualizations which encode either summary or distributional information, as well as a baseline of no imputation, to cover a range of previously-evaluated techniques.

3 Experiment Design

We design and conduct a pre-registered experiment to study the effect of different uncertainty encodings on users' performance in low-level visualisation tasks, and self-reported confidence in their responses, when inspecting an incomplete dataset. Our pre-registration can be found here: (`https://aspredicted.org/blind.php?x=137_X7N`). We use data with two quantitative variables (x, y) and one binary variable (z). Participants were asked to perform two visual analytic tasks—average and trend estimation—with scatterplots. The tasks were selected through an investigation of a full set of common visual analytics tasks (§3.3). We use scatterplots as it is an effective representation for a multivariate dataset with two quantitative variables, supports encoding of imputations for both quantitative variables at the same time, and is ideal for the two tasks we consider in the study. We simulate datasets where missing values were imputed (§3.4) and imputations were visualised using six distinct representations (§3.2). We also consider other factors which may impact the users' performance such as the missingness mechanism, patterns, and proportion (§3.1).

We conducted a pilot study to determine the size and direction of effects, if any exist, of the experimental variables. For the pilot, we used a real-world dataset, with either 20%, 30%, and 50% missing values, which are three levels of missingness proportions that have been considered in prior work on data imputation [14, 38, 41], and evaluated three representations—*baseline*, *mean* and *CI*—as a preliminary investigation on the effect of data imputation and uncertainty. We focused on a constrained design space of uncertainty visualisation to obtain more statistical power from a smaller sample of participants. In addition to the two tasks included in the final study, we also considered *Find extremum* and *Determine Range* tasks in the pilot study. We use the findings from our pilot study to refine the experimental design space, which we discuss in further detail below.

3.1 Manifestation of Missing Values in the Dataset

As discussed in §2.1, there are various ways in which missingness can manifest in a dataset, and consequently affect both how the data may be represented and how a user performs inference.

Missingness mechanism: The mechanism by which data is missing will impact how missing values may be imputed and the inferences that may be drawn. As discussed in § 2.1, we consider data which is MAR, with missingness correlated with the categorical variable (N), as imputing missing values may improve inference in such cases [33].

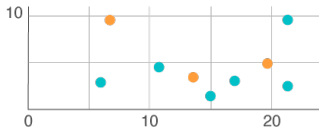
Missingness patterns: Whether missing values are only restricted to one variable (univariate missingness) or two or more variables (multivariate missingness) can affect how imputed estimates and corresponding uncertainty information can be encoded as marks in a graph. In our study, we allow missing values to occur in both variables (x and y), relaxing the univariate data settings of prior studies [2, 15, 41], but only permit missing values to occur in one variable per observation (Fig. 2B); this constraint allows us to focus on univariate representations of uncertainty (§3.2), which is commonly observed in existing uncertainty visualization studies and applications [17, 18, 26, 29]. If multiple variables for an observation were missing (Fig. 2C), we would require multidimensional representations of uncertainty to encode imputations. However, representing uncertainty of multidimensional distributions through a single mark poses a significant challenge as, in addition to communicating the uncertainty for each variable, we would also need to communicate the correlation between the missing variables. For example, an extension of confidence intervals for both x and y variables to represent 2D uncertainty can be misleading because it suppresses information on the correlation between the two variables; consequently two very different multivariate distributions can look identical in this representation. Thus, we deemed showing multi-dimensional uncertainty on a single mark to be outside the scope of current work, and focus on representing univariate missingness for each data point.

Proportion of missing values: We manipulate the proportion of missing data by creating incomplete datasets with 30% and 50% missing observations. We dropped the 20% missing proportion condition from the final study as results from our pilot study (§9 in the supplementary materials) and previous work [41], suggested that performance in tasks will likely not be impacted at lower proportions of missing data.

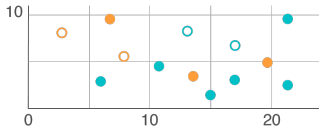
3.2 Visual Representations

The set of feasible representations that can be used to represent uncertainty is constrained by the encoding that has been used to map the variable—if we have uncertainty information for certain observations of the quantitative variable x, the encoding that is used to map x will constrain the available uncertainty representations. Consider for instance the scatterplot, where both x and y are mapped to position channels, and both x and y have missing values; uncertainty representations which use hue, such as Value Suppressing Uncertainty Palettes (VSUPs) [12] cannot be used as neither of the variables which contain uncertainty are being mapped to color. As we only consider uncertainty in the quantitative variables (x, y), we include four types of uncertainty visualisations which satisfy this constraint from recent work [18, 26, 29]. We consider two additional visual representations where we do not include imputations (baseline) or do not include uncertainty (mean).

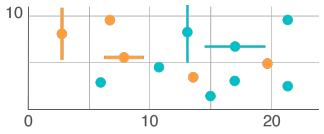
***Baseline**: No Imputation*
All complete cases are represented using a scatterplot with any missing observations not presented to the viewer. This technique is often the default in systems like Tableau, thus making it a reasonable baseline for comparison. We use filled circles as marks to represent each observed data value.



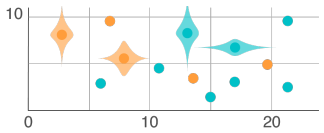
***Mean**: Mean Point Estimate*
Mean point estimates derived from our imputation method are represented using unfilled circle marks, while observed values are represented using filled circles similar to the baseline. No uncertainty information about the imputed estimate is encoded in this representation.



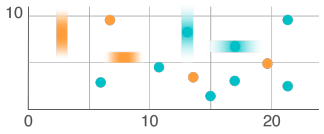
***CI**: Mean Point Estimate with 95% Confidence Intervals*
95% confidence intervals are one of the most commonly used encodings for communicating uncertainty. Since we have distributional information about each imputation, we estimate the 95% upper and lower bounds of this distribution and represent each imputed value using a point estimate and the corresponding 95% confidence interval.



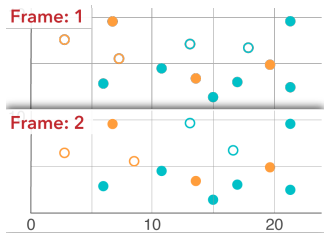
***Density**: Probability Density Plots*
Probability density function (PDF) plots, and other equivalents such as eyeball plots [42], violin plots [10] or raindrop plots [4], provide complete distributional information regarding a random value, using the orthogonal dimension to encode probability density. For instance, if the y value of a data point is imputed, the probability density of the possible y values is encoded as length along the x dimension. These plots allow a reader to understand the shape of the distribution, and support judgements about probability and intervals by comparing ratios of areas. Although these plots are commonly used, estimating probabilities from them may be difficult and may not be the most accurate [17, 24, 29].



***Gradient**: Gradient Plots Showing 95% Confidence Intervals*
Gradient plots use transparency to encode probability density [10, 25]. We map density values at regular intervals to α transparency, with values of higher density being mapped to low α (i.e. more opaque). We include *gradient* as it might be a more intuitive encoding of uncertainty [34] compared to *density* while still preserving distributional information.



***HOPs**: Hypothetical Outcome Plots of predicted imputed values*
HOPs [23] animate a set of draws from a distribution, which in this case, is the predictive distribution of an imputation. Since we are performing multiple imputations to estimate missing observations (see §3.4), *HOPs* allow us to present each imputed value, thereby communicating what different, hypothetically complete datasets might look like. We present 50 distinct frames, and use a frame rate of 2.5fps (each frame appears for 400ms) as per the recommendation by Kale et al. [27]



3.3 Tasks

There have been several proposed taxonomies of the various analysis tasks that visualisation users perform [1, 7, 13, 39]. Following prior work which investigates the performance of users on low-level tasks using scatterplots [31], we adopt Amar et al.'s [1] taxonomy which categorises elements of analysis into the following low-level tasks: *Read Value*, *Compare Value*, *Compute Derived Value*, *Find Extremum*, *Find Anomaly*, *Characterize Distribution*, *Determine Range*, *Cluster*, and *Correlation*. These tasks can be categorised into two groups [3, 31]: *individual* or *elementary* tasks are those referring to individual elements in the chart, and includes the read and compare value tasks; *summary* or *synoptic* tasks involve all or a subset of the elements in the chart, and include all remaining tasks.

The impact of imputation is likely going to vary based on the type of tasks. We use two criteria to decide whether to include a task in our study—first, whether presenting imputation and uncertainty information would impact the task; and second, whether a scatterplot is suited for performing the task. For individual tasks, two scenarios arise—in the first scenario, the target data point is not missing, which makes the task trivial and has been extensively studied in prior work; in the second, the target data point is missing, which would make the task impossible in the no imputation condition, and similar to probability estimation in the other conditions. This has been studied in prior work [23, 29] and it makes more sense to look at only the single distribution when performing this task. Thus, we exclude individual tasks as presenting imputations are not going to impact such tasks.

For *summary* tasks, imputations do play a role—when missing values are not imputed and hence omitted from the graph, the user would be performing tasks based on incomplete information, but may be unaware of this fact; when missing values are imputed, the user is supposedly acting on more complete information, but has to factor in the quality of imputations. Thus, we focus only on *summary* tasks and systematically review them to identify those that are likely to be affected by visualisation of imputed values. *Find Extremum* and *Determine Range* tasks are only non-trivial if an extremum is missing as otherwise, it will be no different then the analogous task for a dataset with no missing values; if an extremum is indeed missing, it will require the user to consider points at the tails of a distribution which are less likely to be accurately imputed, thereby diminishing the value of imputation. These two tasks were included in our pilot study, and as expected, we did not find a meaningful effect of imputation. *Characterize Distribution* task may be operationalised in different ways [39], and have subjective definitions. If the task is to understand the (joint or marginal) probability distribution of a variable, scatterplots may not even be best suited for the task. *Cluster* and *Find Anomaly* tasks have subjective definitions where ground truth is hard to specify and may introduce additional bias in the estimation. In this work, we try to focus on tasks that have clear ground-truth answers and can be performed based on objective value estimation. We thus omit these five tasks and focus on the remaining two tasks which are formulated as follows:

- *Compute Derived Value*: What is the average value of x?
- *Correlation:* Identify the trend line which best represents the relationship between x and y?

3.4 Stimuli Data Generation and Imputation

The results of our pilot study (§9 in the supplementary materials) indicated that there may be some effect of uncertainty encoding and proportion, but they were too noisy to determine, especially as the real world dataset used did not allow us to easily control for the difference between the incomplete and imputed datasets. This motivated us to use a simulated dataset—we first generate complete datasets and remove values to create an incomplete dataset that is MAR; we then impute the missing values and compute the uncertainty.

Generation: We sample 25 points from each of the two multivariate distributions: $D_1 = \text{MVRNorm}(\mu_1, \Sigma)$ and $D_2 = \text{MVRNorm}(\mu_2, \Sigma)$, where μ_1, μ_2 are 6-dimensional vectors, and Σ is a 6×6 covariance matrix. We use a group indicator variable, I to denote whether a sampled point belonged to D_1 or D_2. This results in a single dataset of $N = 50$ observations. (Code used to generate the stimuli is included in the supplementary materials).

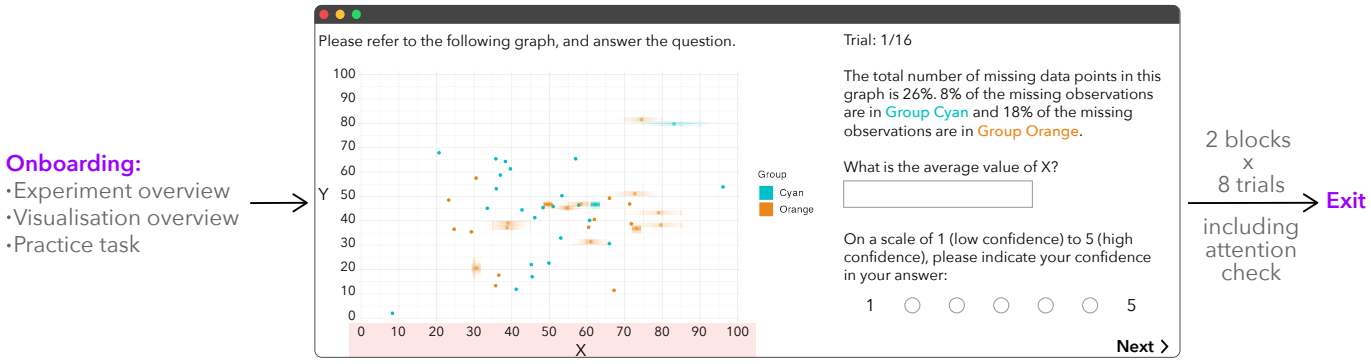


Fig. 3. Experiment overview and task interface

Generating Incomplete Datasets: We remove 30% or 50% of observations from the generated complete dataset under MAR using the `mice` R package [49] to obtain the incomplete datasets. Under MAR, missing data is correlated with a variable present in the dataset—in our case, missing values are correlated with the group indicator variable, I.

Imputation: We use the `mice` package to implement a multiple imputations procedure to fill in missing values for each simulated dataset using predictive mean matching. "Multiple imputation involves replacing each missing data point with $D \geq 2$ imputed values such that D complete data sets can be created from imputation procedure." [33] We estimate $D = 50$ imputations for each missing value.

Uncertainty calculation: Multiple imputations obtained from the previous step reflect the sampling variability associated with imputation, allowing us to quantify the imputation uncertainty for each missing data point. We consider $\bar{\theta}_i$ as the mean of the imputation for the $i-th$ missing data point, where N is the size of the dataset:

$$\bar{\theta}_i = \frac{1}{D}\sum_{n=1}^{N} \theta_{d,i}$$

We then define $\bar{s}_i$ as the estimated standard error of the mean imputed value, which is defined as follows:

$$\bar{s}_i = \frac{1}{\sqrt{D}} Var(\theta_{d,i})$$

Dataset Sampling: Treating each dimension as a unique variable, and taking two variables at a time, we obtain $\binom{6}{2} \times 500 \times 2 = 15000$ unique bivariate visualisations. We compute *average* and *trend* statistics for each of the complete, incomplete and imputed versions of the simulated datasets. The *trend* statistic is the estimated slope parameter obtained when regressing the y variable on to the x variable. From the 15000 candidate bivariate datasets, we sample 16 datasets to be used in our experiment using the following metric. We estimate the difference in the calculated statistics between the incomplete and imputed datasets—for average, we use the difference relative to the uncertainty (standard error) of the corresponding variable; for trend, we use the absolute difference. This difference, which we will refer to as Δ, indicates the difference in the solutions of our tasks for the incomplete and imputed datasets. If Δ is low, it implies that imputation is unlikely to improve task performance. We sample eight datasets for each task at two levels of Δ: 0.15 and 0.2. For *average estimation*, these values were selected because they correspond to small-to-medium standardised effect sizes (Cohen's d). For *trend estimation*, prior work [11] suggests that the margin of error in such tasks could lie in the range of 0.08-0.12; thus, we selected datasets with a slightly greater difference to be able to detect difference in participants' performance while accounting for potential margin of error. We use these datasets to create a chart for each visual representation being considered in our study.

3.5 Procedure

We employed a mixed design, using between-subjects treatment for *representation* (6), and a within-subjects treatments for *tasks* (2) and *proportion* (2). We define one block of trials as a single task repeated eight times. Proportion is randomly shuffled within each block of trials. Each participant completes 2 blocks of trials. Before participants begin the study they go through a tutorial which presents them with a graph similar to the ones they will subsequently be shown during the trials. The tutorial describes the graph, the presence of missing values and the proportion of missing values. For the test conditions, the tutorial informs the participants that an appropriate imputation method was used to estimate the missing values, and that these estimates contain uncertainty. For each of the uncertainty representation conditions, participants were given a brief description of how to interpret them. Following this description, participants are given a practice task and are provided feedback on their performance. Screenshots of the study interface can be found in the supplemental materials.

Fig. 3 provides a quick overview of how the experiment was set up. Participants in all conditions were informed of the proportion of missing values in the chart shown to them. For each task, we elicit responses using a direct report method [16] which involved either entering numeric values into a textbox for *average estimation* or adjusting a slider to modify the slope of a line for *trend estimation*. In addition, we elicit participants' self-reported confidence in their response using a five-item Likert-style question. As per our pre-registration, we recruited a total of 210 participants (35 participants for each between-subjects treatment), using the `prolific.co` platform, between Oct 10th and Oct 17th 2021. Each participant was paid $3 for completing the study. The median completion time was 12mins 13s, and the median wage was $14.76/hr. Participants who failed the attention check question were not allowed to complete the study. We excluded a further four participants whose responses for the trend estimation task was exactly the same value across all trials. Another four participants were excluded because they participated in the study multiple times, and when participating for the first time they did not complete the entire set of 16 trials. As a result, we were left with 202 participants. All participants were fluent in English, and resided in the United States; 49% of our participants self-identified as Female while one did not disclose, and all but one participants had completed a college degree. A detailed overview of the experimental setup can be found in §2 of the supplementary materials.

4 Analysis

Following similar work in the visualisation literature [35], we use *bias* and *precision* as measures to compare participants' performance. *Bias* (Fig. 4) indicates whether and how the mean response deviates from

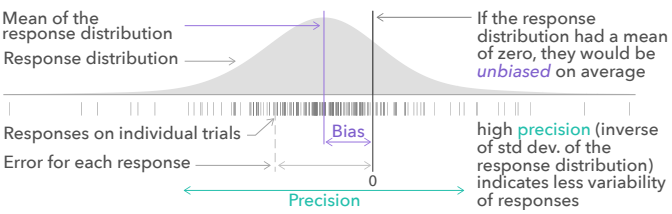


Fig. 4. Bias and precision measures used in the analysis (figure adapted from McColeman and Yang et al. [35]).

the actual value in a systematic manner. Even though each participants' response to a particular task may contain some error, if they are on average unbiased, the overall distribution of errors is generally expected to possess a mean at or close to zero. On the other hand, if participants have a tendency to consistently make mistakes in a certain direction, they are exhibiting signs of bias. *Precision* (Fig. 4) describes how consistent participants are in their responses relative to the actual value. If participants' responses are precise, their responses are very close to each other; if they are imprecise, then participants' responses may vary a great deal from one another. We use Bayesian hierarchical regression models for analysis as they provide complete distributional information for each parameter, and support probabilistic comparisons, as opposed to reporting dichotomous results from statistical significance tests.

4.1 Analysis of Performance in Tasks

We compare *bias* and *precision* across the between-subject (*representation* variable) and within-subject (*proportion* variable) conditions. We use a Bayesian hierarchical model to estimate the error distribution, using a student's t-distribution as the likelihood function. The mean and variance of the student's t-distribution are the bias and precision estimates for participants' responses.

Our regression model can be formulated using an extended form of the Wilkinson-Rogers-Pinheiro-Bates [37, 50] notation as follows:

1: $error \sim student\_t(\nu, \mu, \sigma)$
2: $\mu = representation * proportion +$
3: $adj\_trial\_id + block\_id + axis +$
4: $(proportion|participant\_id)$
5: $log(\sigma) = representation * proportion +$
6: $adj\_trial\_id + block\_id + axis +$
7: $(proportion|participant\_id)$

Here, *error* is the difference between a participant's response and the ground truth value; *representation* is a variable which indicates the uncertainty visualisation condition; *proportion* is a variable of two levels which indicates the proportion of missing data; *adj_trial_id* is an indicator variable for the trial number of the response, adjusted to be a value between -1 and 1; *block_id* indicates whether the response is from the first or second set of trials; *axis* indicates whether the *average estimation* task was performed for the variable mapped to the *x* or *y* axis. For *trend estimation* this predictor is omitted; and *participant_id* is a unique participant identifier.

Line 1: We define the likelihood of the distribution of the error in participants' responses as a `student_t` distribution. The `student_t` distribution is parameterised by ν (degree of freedom), which determines how fat the tails of the distribution are. The parameters μ (mean), and σ (scale) are our estimates for *bias* and *precision* respectively.

Line 2: We estimate the population-level effects of the experimental variables *representation* and *proportion* on μ (bias). These predictors determine the effect of different uncertainty representations and the proportion of missing data on *bias*.

Line 3: Our experiment consists of two blocks of eight trials each. We may reasonably expect that as participants progress through the trials, they may learn and find the task progressively easier. We may also expect fatigue to play a role. These two variables capture any potential learning or fatigue effects.

Line 4: Each participant will likely have a varying level of ability for accurately completing our tasks, which may result in errors being more correlated within one participant but less correlated across participants. A varying intercepts term for *participant_id* in our model allows us to account for such variation. Additionally, we may expect the effect of *proportion*, which is varied within subjects, to vary across participants, which we thus include using a varying slopes term. Varying slopes and intercepts provides us with improved estimates from repeated measures data by explicitly accounting for different sources of variation [36].

Lines 5-7: We use the same predictors described above to estimate for σ, as we believe that the same factors which may effect *bias* will also effect participants' *precision*.

4.2 Analysis of Confidence

We use a Bayesian hierarchical ordinal regression model to estimate a participant's probability of answering each item on the Likert-style question, and the effects of different experimental variables on this probability. Our regression model can be formulated as:

1: $R_i \sim Ordered(p)$
2: $logit(p_k) = representation * proportion +$
3: $adj\_trial\_id + block\_id +$
4: $(proportion|participant\_id)$

Line 1: The ordered distribution is a categorical distribution which takes a vector, $p = \{p_1, p_2, p_3, p_4\}$. The length of p depends on the number of discrete levels of the categorical response value. Each element in this vector indicates the cumulative probabilities for each response value, except for the maximum value which will always have a probability of 1 [36].

Lines 2-3: We estimate the effects of the experimental variables *representation* and *proportion* on p_k as population level effects. We include predictors for adj_trial_id and $block\_id$ to capture any effects of learning or fatigue as participants progress through the trial.

Line 4: We expect some variation between participants' reported confidence due to individual characteristics. We may also expect *proportion* to effect participants differently. Varying slopes and intercepts are included to account for these sources of variation.

5 Results

Our model estimates *bias* and *precision* for an *average* participant for each *representation* and *proportion*. We present results for an average participant to remove the effects of individual variation that is captured by our model. We marginalise over adj_trial_id, $block\_id$ and *axis* to average out the influence of these variables, as they do not appear to have a meaningful effect on our measures.

5.1 Average Estimation Task

For the *average estimation* task, the difference in estimates between the two levels of proportion are small and consistent for each *representation*. The results broken down by *proportion* can be found in §5.1-5.2 of the supplementary materials.

A. Bias in participants' responses (average estimation)

In all *representation* conditions except *CI*, an average participant is likely to be biased, and will consistently underestimate the average value of a set of points (Fig. 5A). Unsurprisingly, an average participant will exhibit the greatest magnitude of bias in the *no imputation* (baseline) condition. When compared to this baseline, all other visual representations appear to reduce bias (Fig. 5A). An average participant in the *HOPs* condition is expected to show the smallest reduction in bias, with an estimated 95% probability of being less biased compared to the baseline. The *CI* condition is expected to exhibit the greatest reduction in bias, perhaps even resulting in unbiased estimates on average.

B. Precision of participants' responses (average estimation)

An average participant is the most precise (most consistent in their responses) in the *no imputation* and *mean* conditions (Figure 5B). There is a small probability (60%) that the average participant might be slightly more precise in the *mean* condition compared to the baseline. The other uncertainty representation conditions—*CI*, *density*, *gradient*, and *HOPs*—are all more likely to result in less precise estimates for an average participant, with an estimated probability of at least 90%.

C. Self reported confidence (average estimation)

Our model analysing confidence provides us with a probability distribution for each item on the Likert-style question, for an average participant in each *representation* condition. Participants were somewhat unsure about their responses in every *representation* conditions as indicated by the high estimated probability of the average participant responding 2 or 3 on the Likert-style question compared to the other

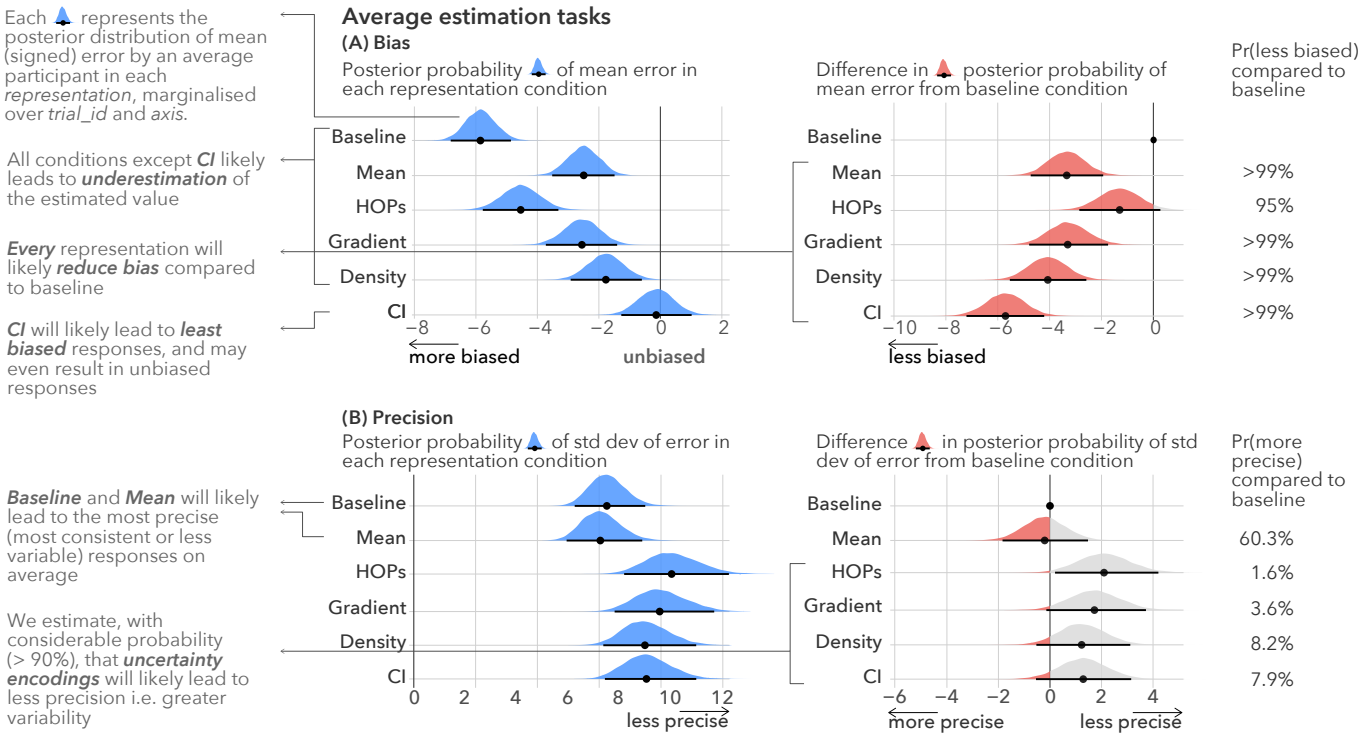


Fig. 5. Experiment results for the mean estimation tasks: posterior probability densities, mean and 95% credible intervals for bias and precision for an average participant in each representation condition (left). We compare the difference in bias and precision compared to the baseline (right) and estimate the probability of improvement for each representation condition.

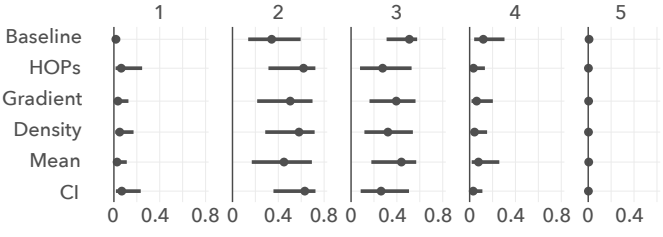
values (Fig. 6). As such we compare the cumulative probability for an average participant to respond to an item 3 or greater on the Likert-style question, $Pr(\geq 3)$ for each *representation* condition. The average participant was likely to be most confident in the *no imputation* condition, with $Pr(\geq 3) = 64\%$. Presenting imputations without uncertainty (*mean* condition) results in slightly less confidence, compared to the baseline, with $Pr(\geq 3) = 52\%$. The other *representation* conditions which explicitly encode uncertainty information result in even lesser confidence in their response, with $Pr(\geq 3) = 31\%, 33\%, 38\%$ and 46% for *CI*, *HOPs*, *density* and *gradient* conditions respectively.

5.2 Trend estimation task

Unlike in the *average estimation* task, for *trend estimation*, we observe a larger and sometimes inconsistent effect of proportion on the quality of participants' responses. Hence, in the following sections, we look at the results separately for each level of *proportion*.

A. Bias in participants' responses (trend estimation)

The average participant, in all *representation* conditions, at both levels of proportion appears to systematically overestimate the trend in the data (Fig. 7A). At lower levels of missing data *proportion*, only the *mean* condition may reduce bias in responses, with an estimated probability of 88%, compared to the baseline. When compared to the baseline condition, the average participant's responses in *gradient* and *HOPs* were likely to be just as biased when shown; however, the responses in *density* and *CI* conditions may likely become more biased. At higher levels of missing data *proportion*, only the *HOPs* condition may reduce the bias in responses, with an estimated probability of 91%, compared to the baseline. The average participant in all the other *representation* conditions may exhibit increased bias in their responses.



Posterior mean and 95% confidence interval ⊸ for the estimated probability that an average participant would rate their confidence as on the five-item Likert-style question, in each condition

Fig. 6. Experiment results for participants' confidence in their responses for the *trend estimation* task

B. Precision of participants' responses (trend estimation)

For the average participant, when the *proportion* of missing data increases, precision will likely decrease in the *baseline*, *mean*, and *CI* conditions, while it may remain similar in the *HOPs*, *gradient*, and *density* conditions (Fig. 7B). The average participant was most precise when presented with *HOPs*, at both levels of *proportion*. Compared to the baseline, *HOPs* is more likely to result in more precise responses with an estimated probability of 79% and 99% when the *proportion* of missing values is 30% and 50% respectively. All the other *representation* conditions will either not improve precision (*mean*), or may lead to less precision (*gradient*, *density*, *CI*).

C. Self reported confidence (trend estimation)

Compared to the *average estimation* task, the average participant appears to be more confident in their responses when performing the *trend estimation* task, as indicated by the greater estimated probability of responding 3 or higher on the Likert-style question across all *representation* conditions (Fig. 8). The average participant is likely to have similar confidence in the *no imputation* condition ($Pr(\geq 3) = 66\%$), as the *mean* ($Pr(\geq 3) = 67\%$) condition, while they are likely to be only marginally less confident in *gradient* ($Pr(\geq 3) = 64\%$), and *HOPs* ($Pr(\geq 3) = 58\%$) conditions.

6 Discussion

The effect of imputating and encoding imputations with uncertainty in a chart can vary based on the task, and in some cases, the amount of missing data. In the following, we discuss how our findings may impact the design of charts for data with missing values.

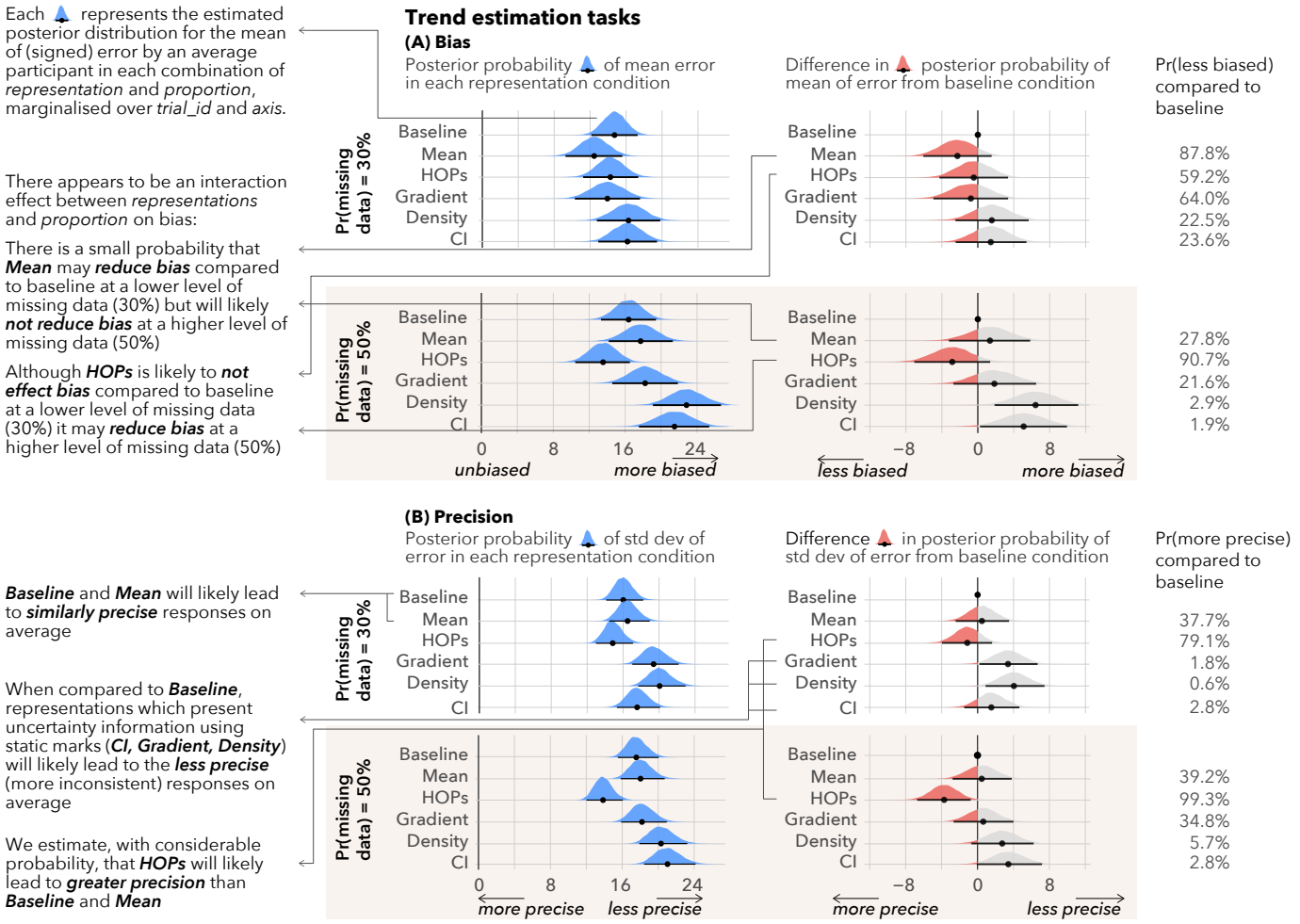
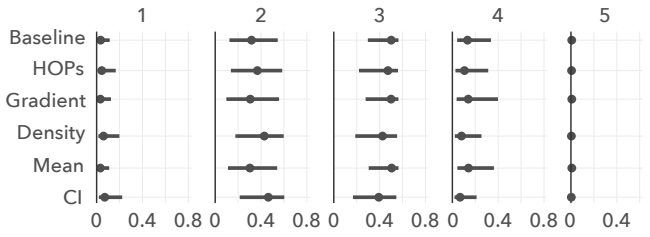


Fig. 7. Results for trend estimation tasks: posterior probability densities, mean and 95% credible intervals for bias and precision for an average participant in each representation, conditional on proportion (left). We compare the difference in bias and precision compared to the baseline (right) and estimate the probability of improvement for each representation condition, conditional on proportion.

6.1 To Add or Not To Add (Uncertainty Information)?

In the *average estimation* task, there is an almost deterministic benefit of showing imputed estimates ($Pr(\textit{less biased}) > 99\%$ for all representations but HOPs; Fig. 5), as all our treatment visualisation conditions are likely to decrease error and reduce bias when compared to the baseline. However, the benefit of adding uncertainty information may be more inconsistent. Only the *mean* condition, which does not present any uncertainty information, led to no decrease in precision compared to the baseline, while all conditions which encode uncertainty information (*CI*, *density*, *gradient* and *HOPs*) decreased precision, suggesting that presenting uncertainty information may lead to participants being less consistent. A possible explanation for this result may be that the presence of additional marks, which are used to encode uncertainty information, may interfere with participants ability to read values precisely from the charts or perform ensemble processing [45], or require greater cognitive effort [9] on the part of the viewer.



Posterior mean and 95% confidence interval ⊸ for the estimated probability that an average participant would rate their confidence as on the five-item Likert-style question, in each condition

Fig. 8. Experiment results for participants' confidence in their responses for the *average estimation* task

Due to this tradeoff, our results suggest that the *mean* representation may be superior to *HOPs*, *gradient*, and *density*, as the average participant will be more precise but will not be more biased. However, the comparison between *mean* and *CI* is more difficult, as when compared to *mean*, *CI* is likely to be less biased with an estimated probability of 98%, but also more imprecise with an estimated probability of 87%.

In the *trend estimation* task, none of the *representation*s appeared to consistently improve participants' performance. The use of *CI*, *density*, and *gradient* appear to neither consistently reduce bias nor improve precision, even when compared to the baseline condition of not showing imputation information. Presenting participants with only *mean* imputation information results in similarly precise responses when compared to the baseline, but leads to an inconsistent effect on bias—while there may be a reduction in bias at lower *proportion*s of missing data it is unlikely to do so, or even lead to more bias, at higher *proportion*s. On the other hand, *HOPs* appears to improve precision consistently as well as reducing bias at higher *proportion*s, but is likely going to result in responses which are just as biased at lower *proportion*s, when compared to the baseline.

Our results suggest that the number of discrete sources of uncertainty that a user needs to process when performing a task may potentially impact their performance. While prior work has suggested that users may be able to aggregate information from two uncertainty distributions [18], in our study, each visualised dataset consisted of 50 observations,

resulting in 15 or 25 distinct uncertainty distributions being encoded in the chart. The average participant was less biased compared to the *baseline* at *average estimation*, but not so for *trend estimation*, across all uncertainty representation conditions. As *average estimation* is performed along a single axis, they had to consider only half the uncertainty distributions represented in the chart. Thus, depending on the task, the additional uncertainty information, represented in this way, may not always be beneficial to participants. However, if the task that the user has to perform with a chart is clear, there may still be considerable benefit from using uncertainty representations such as *CI* for *average estimation* and *HOPs* for *trend estimation*.

Unlike previous studies of uncertainty visualisation, since we present uncertainty information regarding points in a scatterplot, there may exist certain limitations to using representations such as the *density* plot. *Density* plots use length along the orthogonal dimension to encode probability density; but, in scatterplots, both x and y dimensions are already used as visual channels to represent points. Even though we provide participants with a tutorial on how to interpret each uncertainty representation, *density* plots may still be potentially more challenging for viewers to interpret. Our results do not seem to suggest that participants in the *density* condition performed worse than the other uncertainty conditions; rather in the *average estimation* task, the *density* condition is likely to result in less biased responses (see Fig. 5). However, we hope that future work can investigate further how the uncertainty information of points in scatterplots represented by density plots are perceived by users.

6.2 Is Greater Confidence Desirable?

Our initial hypothesis regarding participants' self-reported confidence was that presenting only *mean* imputation estimates (no uncertainty information) would lead to participants being the most confident. We expected that not showing imputations would lead to participants being less confident, as they would be operating under insufficient information, and would be aware of this fact. We expected that the uncertainty conditions would lead to participants being less confident than the *mean* condition due to the explicit presentation of uncertainty information.

While our results suggest that the latter hypothesis is supported, the former is not. The presence of imputed values (which used hollow circles as marks as opposed to filled circles for actual values) may have served as an indication to participants that the data may not be as reliable, leading them to often be less confident in their responses compared to the baseline. In an exploratory analysis, we compare the correlation between participants error and their self-reported confidence. We find a small negative correlation (-0.16) for the *average estimation* task and no correlation (-0.01) for the *trend estimation* task. While some prior studies have used self-reported confidence as a metric to compare uncertainty representations [19, 41], our exploratory analysis suggests that, in some situations, participants' self-reported confidence could potentially be an unreliable indicator of task performance. Since for the *average estimation* task, the average participant was likely to be most confident in the condition they were likely to be most biased (*baseline*), we posit that being more uncertain may even be desirable in certain cases, as that may lead to participants refraining from making strong, but incorrect, conclusions about the data. In future work, we would like to further investigate the effect of representing imputed values in charts through incentivised decision making tasks.

6.3 Does Missing Data Proportion Impact Performance?

As the amount of missing data increases, the amount of available information in the chart decreases. Thus, we would expect participants' performance to worsen. We would also expect this degradation to be greatest when missing observations are simply omitted (i.e., the *no imputation* condition), and relatively smaller in all the remaining *representation* conditions, when compared to *no imputation*.

For *average estimation*, contrary to our expectations, the average participant in every *representation* condition was likely to be just as biased or even slightly less biased (smaller errors), but not more biased, at higher *proportion*s (see Fig. 5.2 in the supplementary materials). This result is quite surprising, and may suggest that the presence of additional marks may not be as distracting for *average estimation* tasks when compared to *trend estimation*.

For *trend estimation*, as expected, the average participant is more biased at higher *proportion*s in every *representation* condition except *HOPs*. However, since participants are more likely to perform worse in the *CI*, *density*, and *gradient* conditions when compared to *baseline*, the degradation in performance between lower and higher *proportion* is, contrary to expectation, greater than *baseline* (§7.1 in the supplementary materials). If the presence of additional marks is indeed making *trend estimation* more challenging to participants, this result may be expected, as at higher *proportion*, participants would have to consolidate a greater number of marks encoding uncertainty information.

6.4 Limitations and Future Work

Even though we consider bivariate missingness, we only test data where one of the quantitative variables was missing at a time. In real-world datasets, multiple variables may be missing in conjunction, which may require uncertainty representations which encode two or more joint probability distributions together. However, two-dimensional representations of uncertainty are fairly complex to visually encode and communicate to a viewer, and perhaps even more challenging to interpret. We hope this can be further explored in future work.

When comparing performance between the *average* and *trend* tasks, our results appeared to suggest that, as the number of discrete sources of uncertainty increases, users' task performance may decline. However, the effect of proportion—a factor which impacts the number of discrete uncertainty points in the chart—on performance was unclear, and we cannot make claims regarding the exact relationship between these two variables. We initially included a 20% missingness proportion condition in our pilot study, but, similar to prior work [41], we found that participants were not likely to perform better in the 20% condition compared to the 30% condition. Thus, we decided to drop the 20% condition and test only two levels of proportion, allowing us to also obtain greater statistical power. Despite this, as the proportion of missing data increases, one should still expect performance to worsen because of the increase in the number of discrete sources of uncertainty represented in the chart.

The number of data points is another factor that may influence the number of discrete uncertainty points in the chart. However, due to resource constraints, we only focused on "medium" number of 50 data points [39]. We speculate that a study design with more levels of proportion and varying number of data points in the chart will be able to better identify the effect of the number of "discrete points with uncertainty", which we leave for future work.

More work is also needed to determine the role of factors such as cognitive load on task performance. Castro et al. [9] highlight differences in effort, measured using NASA-TLX, when users make decisions with uncertainty representations. As users in our study had to aggregate information from multiple probability distributions, we may expect uncertainty conditions to impact the effort required in performing the tasks. We hope to explore this in future studies.

7 Conclusion

We contribute the results of a crowdsourced study investigating visualisation users' performance on *average* and *trend estimation* tasks with an incomplete dataset and the role of imputing missing values. We vary how imputations are encoded in the chart using different uncertainty representations. For *average estimation*, when compared to not presenting uncertainty information, we find that showing only *mean* imputations will likely reduce bias (desirable), while showing uncertainty (*CI*, *density* and *gradient*) may likely reduce bias but may also reduce precision (undesirable). For *trend estimation*, we find that showing only *mean* of imputations may have a small effect on reducing bias at a lower proportion of missing data, but may likely not reduce bias at a higher proportion of missing data; on the other hand, while *hypothetical outcome plots* may not reduce bias at a lower proportion of missing data, it is likely to reduce bias at a higher proportion of missing data.

REFERENCES

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117, 2005. doi: 10.1109/INFVIS.2005.1532136

[2] R. Andreasson and M. Riveiro. Effects of visualizing missing data: An empirical evaluation. In *2014 18th International Conference on Information Visualisation*, pp. 132–138, 2014. doi: 10.1109/IV.2014.77

[3] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006. doi: 10.1007/3-540-31190-4

[4] N. J. Barrowman and R. A. Myers. Raindrop plots. *The American Statistician*, 57(4):268–274, 2003. doi: 10.1198/0003130032369

[5] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005. doi: 10.1037/1082-989X.10.4.389

[6] K. Bhaskaran and L. Smeeth. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339, 04 2014. doi: 10.1093/ije/dyu080

[7] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124

[8] S. Casner and J. H. Larkin. Cognitive efficiency considerations for good graphic design. 11:275–282, 1989.

[9] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla. Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):411–421, 2022. doi: 10.1109/TVCG.2021.3114803

[10] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014. doi: 10.1109/TVCG.2014.2346298

[11] M. Correll and J. Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 1387–1396. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025922

[12] M. Correll, D. Moritz, and J. Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–11. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174216

[13] E. Dimara and J. Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2022. doi: 10.1109/TVCG.2021.3114813

[14] S. Dray and J. Josse. Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, 216(5):657–667, 2015. doi: 10.1007/s11258-014-0406-z

[15] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. In M. F. Costabile and F. Paternò, eds., *Human-Computer Interaction - INTERACT 2005*, pp. 861–872. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[16] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1117–1127, 2021. doi: 10.1109/TVCG.2020.3029413

[17] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718

[18] M. Greis, A. Joshi, K. Singer, A. Schmidt, and T. Machulla. Uncertainty visualization influences how humans aggregate discrepant information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174079

[19] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao. Visualizing uncertainty and alternatives in event sequence predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300803

[20] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. doi: 10.1080/00031305.1998.10480559

[21] 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. doi: 10.3758/s13423-013-0572-3

[22] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222, 2011. doi: 10.1109/TVCG.2011.175

[23] 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*, p. 25, 2015. doi: 10.1371/journal.pone.0142444

[24] H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis*, 7(4):519–529, 1987. doi: 10.1111/j.1539-6924.1987.tb00488.x

[25] C. H. Jackson. Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347, 2008. doi: 10.1198/000313008X370843

[26] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021. doi: 10.1109/TVCG.2020.3030335

[27] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2019. doi: 10.1109/TVCG.2018.2864909

[28] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, p. 3363–3372. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1978942.1979444

[29] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 5092–5103. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858558

[30] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003. doi: 10.1023/A:1021564703268

[31] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. *Computer Graphics Forum*, 37(3):157–167, June 2018. doi: 10.1111/cgf.13409

[32] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987. doi: 10.1016/S0364-0213(87)80026-5

[33] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019. doi: 10.1002/9781119013563

[34] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012. doi: 10.1109/TVCG.2012.279

[35] C. M. McColeman, F. Yang, T. F. Brady, and S. Franconeri. Rethinking the ranks of visual channels. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):707–717, 2022. doi: 10.1109/TVCG.2021.3114684

[36] R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018. doi: 10.1201/9780429029608

[37] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willigen, and R. Maintainer. nlme : Linear and nonlinear mixed effects models. r package version 3.1-103. *http://cran.r-project.org/web/packages/nlme/index.html*, 2017.

[38] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani. A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 1–6, 2016. doi: 10.1109/ICITSI.2016.7858189

[39] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018. doi: 10.1109/TVCG.2017.2744184

[40] H. Song, Y. Fu, B. Saket, and J. Stasko. Understanding the effects of visualizing missing values on visual data exploration. In *2021 IEEE Visualization Conference (VIS)*, pp. 161–165, 2021. doi: 10.1109/VIS49827.2021.9623328

[41] H. Song and D. A. Szafir. Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer*

Graphics, 25(1):914–924, 2019. doi: 10.1109/TVCG.2018.2864914

[42] D. J. Spiegelhalter. Surgical audit: statistical lessons from nightingale and codman. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):45–58, 1999. doi: 10.1111/1467-985X.00120

[43] D. F. Swayne and A. Buja. Missing data in interactive high-dimensional data visualization. *Computational Statistics*, 13(1):15–26, 1998.

[44] D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998. doi: 10.1080/10618600.1998.10474764

[45] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11–11, 03 2016. doi: 10.1167/16.5.11

[46] M. Templ, A. Alfons, and P. Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1):29–47, 2012. doi: 10.1007/s11634-011-0102-y

[47] E. R. Tufte. *The Visual Display of Quantitative Information*, vol. 2. Graphics Press Cheshire, CT, 2001.

[48] A. Unwin, G. Hawkins, H. Hofmann, and B. Siegl. Interactive graphics for data sets with missing values—manet. *Journal of Computational and Graphical Statistics*, 5(2):113–122, 1996. doi: 10.1080/10618600.1996.10474700

[49] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03

[50] G. Wilkinson and C. Rogers. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(3):392–399, 1973. doi: 10.2307/2346786

[51] B. W. Wong and M. Varga. Black holes, keyholes and brown worms: Challenges in sense making. vol. 56, pp. 287–291, 2012. doi: 10.1177/1071181312561067