Odds and Insights: Decision Quality in Exploratory Data Analysis Under Uncertainty

Abhraneel Sarma
abhraneel@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Xiaoying Pu*
xpu@umich.edu
Independent
Seattle, WA, USA

Yuan Cui
yuancui2025@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Eli T. Brown
eli.t.brown@depaul.edu
DePaul University
Chicago, Illinois, USA

Michael Correll
m.correll@northeastern.edu
Northeastern University
Portland, Maine, USA

Matthew Kay
mjskay@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

ABSTRACT

Recent studies have shown that users of visual analytics tools can have difficulty distinguishing robust findings in the data from statistical noise, but the true extent of this problem is likely dependent on both the incentive structure motivating their decisions, and the ways that uncertainty and variability are (or are not) represented in visualisations. In this work, we perform a crowd-sourced study measuring decision-making quality in visual analytics, testing both an explicit structure of incentives designed to reward cautious decision-making as well as a variety of designs for communicating uncertainty. We find that, while participants are unable to perfectly control for false discoveries as well as idealised statistical models such as the Benjamini-Hochberg, certain forms of uncertainty visualisations can improve the quality of participants' decisions and lead to fewer false discoveries than not correcting for multiple comparisons. We conclude with a call for researchers to further explore visual analytics decision quality under different decision-making contexts, and for designers to directly present uncertainty and reliability information to users of visual analytics tools. This paper and the associated analysis materials are available at: https://osf.io/xtsfz/

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization**; *Visualization design and evaluation methods*; **Visual analytics**;

KEYWORDS

multiple comparisons problem, uncertainty visualization, decision-making

*Currently at Apple.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3641995

ACM Reference Format:
Abhraneel Sarma, Xiaoying Pu, Yuan Cui, Eli T. Brown, Michael Correll, and Matthew Kay. 2024. Odds and Insights: Decision Quality in Exploratory Data Analysis Under Uncertainty. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3613904.3641995

1 INTRODUCTION

Imagine that you are an analyst charged with the responsibility of identifying the profits of stores of a company. The company operates in 10 different regions, each of which has 200 stores. The company's manager have recently implemented an ambitious new store policy, and wants you to assess the impact of the new policy on the stores' profitability. However, only 20 of the stores in each region have provided you their financial reports. To assess profitability, you use an *exploratory data analysis* (EDA) system to create visualisations of the profits of the 20 stores in each of the 10 regions that you have data for; you then calculate the average profit in each region and quickly identify the regions which are likely to have been profitable on average.

Modern visual analytics systems such as Tableau and PowerBI allow analysts of varying levels of expertise to quickly create and modify such visualisations. With simple UI components, analysts can easily sort, filter, and slice data. This encourages analysts to rapidly explore the data, generate numerous charts, and, potentially, a great number of insightful findings. In other words, these tools enable EDA: an unconstrained and often visual search for interesting and meaningful trends or patterns in a data set [68, 69].

Returning to our example, how can you, as the analyst in this scenario, be sure if the insights you have generated are reliable and accurate? By lowering the barrier to engage in EDA, and allowing the analyst to rapidly and iteratively test multiple hypotheses, visual analytic systems may be susceptible to the *multiple comparisons problem* [4, 49, 55, 75, 80]. Well-known in statistics, the multiple comparisons problem can occur when a user tests multiple hypotheses, sometimes even implicitly, increasing the likelihood of finding a false positive. For instance, if an analyst applies a statistical test at the 5% significance level to a dataset where the null hypothesis is true, the probability of a false positive is, by definition, 5%. However, applying the same test on 10 null datasets will increase the chance of finding at least one false positive to $1-(1-0.05)^{10} \approx 40\%$.

Similarly, the more charts an analyst generates through an unconstrained search of possible combinations of data fields or subsets, the higher the odds of encountering, just by chance, a finding that appears insightful but which is merely statistical noise.

Recent literature suggests that multiple comparisons and false discoveries could be a significant concern for EDA systems. In one study, over 60% of "insights" found were false when participants were asked to report "any reliable observations" in a synthetic dataset [78]. In another study where participants were asked to perform visual analysis on specific tasks with an EDA tool, approximately 20% of the answers provided by participants were incorrect [4]. If these error rates generalise to EDA in practice, the ease of discovery in current tools could lead to poor decision-making. However, prior work falls short of a realistic evaluation of decision-making quality in EDA in two crucial ways: (1) the lack of uncertainty representations, and (2) the lack of incentives.

In prior work [78], participants were asked to generate *insights*—an essentially inferential task, requiring participants to make generalisations based on a small data sample. A data analyst would typically generate *hypotheses* and then use inferential statistical methods to make generalised claims about the data. The visualisation analog of conducting such inferential statistical tests would be to use uncertainty representations which visualise a summary statistic, such as the mean or median, and the uncertainty associated in estimating this statistic, such as the standard error [19]. Different types of uncertainty visualisations have been found to improve decision quality when users make decisions under risk [21, 41, 42]. An EDA system that fails to visualise uncertainty does not explicitly provide the information analysts need to identify reliable insights, control error rates, or potentially mitigate the multiple comparisons problem.

In most decision-making scenarios, the consequences of such "mistakes" can be determined. For instance, consider a scenario where there exists a high cost for making an intervention, such as flagging a potential finding for a full-scale, double-blind clinical trial. An analyst cognisant of this cost might be unwilling to act on an apparent finding unless they are very certain. Conversely, as in many monitoring scenarios, the cost of investigating a few false positives might be considerably lower than the cost of missing an important event (a false negative). The commonly used value for $\alpha = 0.05$ in null hypothesis statistical testing (NHST) is one such acknowledgement of the cost of a false positive relative to the reward of a true positive, albeit a completely arbitrary one. As such, a realistic evaluation of participants' decision making requires explicitly defining the rewards and penalties for correct and incorrect decisions. In experiments, this can be achieved through the use of incentives [74]. Failing to consider incentives in experimental design can reduce validity—can we claim that a 60% false positive rate is bad if participants may have falsely believed that the reward for true positives is arbitrarily high? Without an explicit incentive structure, participants in an EDA task might simply attempt to maximise the quantity of insights identified and ignore the quality, leading to a high false discovery rate.

To understand the impact of uncertainty representations and incentives, we conduct a pre-registered,[1] crowd-sourced, incentivised experiment where participants make decisions from multiple datasets visualised simultaneously, mimicking an EDA-like multiple comparisons setting. We investigate (1) **whether participants in a multiple comparisons scenario adjust for multiple comparisons** and (2) **whether uncertainty representations affect participants ability to correct for multiple comparisons**. Specifically, we present participants with 12, 16 or 20 graphs at the same time, and ask them to make a decision for each graph. We visualise the data in the graphs using a scatterplot (**baseline**), 50% confidence intervals (**ci**) or probability density functions (**pdf**).

We measure participants performance using three metrics: (1) the probability of a *false positive,* (2) the *false discovery rate,* and (3) *points* accumulated based on our incentive scheme. Compared against two normative strategies as benchmarks—uncorrected (not correcting for multiple comparisons) and Benjamini-Hochberg (a multiple comparisons correction procedure which controls for the false discovery rate)—we find that participants, in the **ci** and **pdf** conditions, on average performed better than the uncorrected benchmark, but worse than the Benjamini-Hochberg benchmark. However, participants in the **baseline** condition, who were shown the data samples directly, perform worse than the uncorrected benchmark. These results suggest that appropriate uncertainty representations can improve participants' decision quality, and when provided with such information, participants may be able to control for False Discoveries to a certain extent. Further, participants report using heterogeneous strategies to complete the task, many employing the visual affordances of the displays they saw, suggesting that different ways of conveying the same uncertainty information can influence decision-making.

2 RELATED WORK

Visualisation researchers have long argued that the primary objective of visualisation is to help users gain *insight* and make data-driven decisions [9, 12, 14, 16, 51, 76]. Several visual analytics systems are designed to achieve that objective, including Tableau, Microsoft PowerBI, TIBCO Spotfire, and Voyager [72, 73]. Although the target is amorphous, some definitions of *insight* in the visualisation literature include, "an individual observation about the data by the participant, a unit of discovery" [58] or "a non-trivial discovery about the data" or "a complex, deep, qualitative, unexpected, and relevant assertion" [51].

The goal of insight discovery in EDA can conflict with the goal of validating and verifying patterns in confirmatory data analysis. Designing for the *detection* or serendipitous *discovery* of insights can require virtues like open-mindedness, perseverance, and a system that supports fluid and extemporaneous exploration [66] that matches the idiosyncratic ways that people can move through the various states of EDA [57]. Such an EDA system might value the speed or ease of constructing new views or queries (as with TimeSearcher, which was designed to "provide analysts with the power to construct queries quickly, [...] and examine results" [26]) but lack a similar focus on tools for verifying or validating the patterns seen. For instance, EDA system designers may refrain from

[1] https://aspredicted.org/2fw4r.pdf

including information about uncertainty (that can be crucial for determining the robustness of a visual pattern) to avoid confusing or overwhelming users, among other reasons [31]. On the flip side, the focus on verification can cost the analyst opportunities for making discoveries: e.g., an analyst approaching a dataset with a specific set of hypotheses to validate can be biased or incurious concerning otherwise obvious data quality concerns [75].

2.1 Pitfalls in Visual EDA

The freedom to create novel views in EDA without considerations for robustness of insight can create two types of errors: errors due to the multiple comparisons problem and errors due to overlooking uncertainty in the visualised data.

Multiple Comparisons in the Garden of Forking Paths: During an unconstrained exploration, an analyst makes many, often implicit, decisions. These decisions create branches in the analysis path, possibly affecting the subsequent exploratory steps. These decisions occur at all stages of the sense-making process, and can result in compounding levels of uncertainty and variability [40]. Gelman and Loken [22] describe this phenomenon as wandering in the *garden of forking paths*. The large degrees of freedom in exploration can result in problematic conclusions that fail to generalise to the entire dataset or the population. In statistical testing, the issue of finding non-generalisable insight can be framed as the multiple comparisons problem: the chance of making a false discovery increases as the analyst tests more hypotheses on the same data or tests the same hypothesis on multiple datasets [55].

There are many approaches in EDA for addressing the garden of forking paths problem in general and the multiple comparisons problem in particular. *Multiverse* visualisations can show that a particular conclusion is robust (or not) across a set of reasonable analyses applied to a dataset [20, 60, 61]. Visual analytics systems may also calculate and display metrics of "insight quality" [8, 15, 62, 80]. In statistical testing, multiple comparisons correction methods adjust p-values to control different error rates in multiple testing scenarios. The Bonferroni correction is a common method that controls the family-wise error rate [64], and the Benjamini-Hochberg procedure controls the false discovery rate [6].

For our study, we want to realistically evaluate whether EDA users implicitly correct for the multiple comparisons problem by assessing their false discovery rates in data decision-making. We use statistical testing (t-tests) and the Benjamini-Hochberg procedure as baselines to interpret participant performance.

Missing Uncertainty Information: The absence of uncertainty information in a chart prevents users from easily judging the reliability of an effect, either through heuristics linked to statistical tests ("inference by eye" [19]) or through more holistic estimations involving both mean and error [17, 41]. At best, this uncertainty information can be recovered *implicitly* [18], either through estimation of spread based on underlying values, or through "graphical inference" [10, 29, 70], where a particular visual pattern's robustness is evaluated by contrasting it with visualisations of data generated under some null hypothesis. Despite the value of uncertainty information, per a survey by Hullman [31], visualisation designers often intentionally omit this information. Common reasons for this omission include the cost in additional visual and cognitive complexity incurred by including uncertainty, the perceived inability of audiences to correctly interpret the uncertainty information, and the difficulty in quantifying this uncertainty in a useful way.

Even if uncertainty is directly visualised, different methods of conveying uncertainty can impact decision quality. Traditional forms of communicating uncertainty, such as error bars or box plots, express distributional information by encoding summary statistics as marks. These contain some amount of uncertainty information and are consistent with the design goal of cognitive efficiency [13, 32, 46, 67]. However, they may not be ideal for many decision-making tasks, as they are subject to biases or non-ideal heuristics [17, 41], may require the reader to have a baseline understanding of statistics [5], or may simply be subject to inconsistencies in what is being encoded (e.g., confidence intervals v.s. standard errors, etc.) [17, 27].

While there are many more uncertainty visualisation techniques for various data types [36, 38, 47, 54, 63], the ones most relevant for our study visualise univariate distributions. They include probability density function (PDF) plots, with variants such as violin plots [25] and gradient plots [17, 37]. In theory these convey complete information about the underlying probability distribution. Uncertainty visualisations have been shown to improve accuracy in statistical reasoning in certain tasks [23, 34, 41–43, 59]. Therefore, we postulate that uncertainty visualisations can improve the quality of decisions and help users reduce excessive error rates from the multiple comparisons problem. In our study, we compare two types of uncertainty visualisations and a baseline to cover a range of previously-evaluated techniques.

2.2 Assessing Insight Reliability and Robustness in EDA

Faced with the issue of analysts discovering insights during EDA that may fail to generalise, a small but growing body of visual analytics research attempts to *quantify* the reliability of insights. Zgraggen et al. [78] ran an experiment where participants were free to explore datasets in an exploratory visual analysis tool, asking participants to report "any reliable observations." These observations were then manually coded into testable hypotheses. For example, one such insight could be "the average age is 50." Zgraggen et al. found that the false discovery rate among those insights was over 60%. On the other hand, Battle and Heer [4] had participants complete "goal-oriented" tasks in Tableau. These participants answered questions that can be judged correct or incorrect, like "which [one of the four] parts of the aircraft appear to get damaged the most." Correspondingly, Battle and Heer report the error rate (not false discovery rate) to be at most 25% in participants' responses to these focused-task questions, and conclude that "participants were cautious analysts" [4]. Comparing their results against the False Discovery Rate reported in Zgraggen et al. [78], Battle and Heer speculate that their lower error rate may be due to more data-proficient participants and more "focused" tasks [4]. That is, if participants have to choose from a small set of answers, they can be more deliberate or cautious, therefore getting more answers correct (true positives); by contrast, if they write down however many insights they may find, participants might make more false positives.

These recent studies of insight quality in EDA may be limited by the lack of *incentives* for decision quality. Without appropriate incentives—such as penalising false positives and rewarding true positives in some proportion to each other—it is difficult to say what the optimal False Discovery Rate should be. Further, the decisions that participants make are less likely to reflect the way that analysts would trade off risks and benefits in a real-world, data-driven decision-making task; instead, participants might be motivated to maximise the total number of insights generated regardless of reliability. Our experimental task is closer to the goal-oriented approach of Battle and Heer (we ask participants to make decisions about a given hypotheses), with the addition of incentives for decision quality.

2.3 Incentivising Decision-Making in Human Subjects Studies

Previous studies in Psychology and HCI have adopted financial incentives to motivate participants in decision-making tasks under uncertainty [21, 39, 50], with incentives determined by maximising a utility function. Many regard incentivised experiments as a more realistic way to study decision quality [30, 33, 44, 45, 56]. Though these studies usually have repeated decision trials, they do not study multiple-comparison scenarios: each trial consists of only *one* decision, and the participant does not make decisions at the same time or make them on the same dataset. Since one decision does not impact the quality of the next, participants do not have to control for error rates across multiple comparisons. However, during EDA, an analyst may test *multiple* hypotheses at once or make multiple decisions from the same dataset, increasing their error rates overall due to the multiple comparisons problem [1, 78]. Our incentivised study evaluates decision quality in a multiple-testing scenario, requiring the participant to test multiple hypotheses in each trial.

3 EXPERIMENT DESIGN

The primary goals of this study were to investigate (1) whether users' decisions reflect implicit multiple comparisons correction when users perform EDA under a specific incentive structure, and (2) whether the type of uncertainty visualisations affects users' decision quality. To that end, we designed an online experiment where participants made decisions in an EDA-like setting with multiple comparisons and were compensated based on their decision quality. Study materials, data, and analyses are in the Supplementary Material and available on OSF[2].

3.1 Task Description and Experimental Apparatus

In our study, we asked participants to play the role of a business analyst who decides which sales regions are likely to have made a profit on average. This task did not assume extensive domain knowledge. In each trial, participants selected the profitable regions from a panel of sales data visualisations. As shown in Figure 1, the task contains a multiple comparisons problem: the participant tests one hypothesis ("Is the average profit in this region greater than zero?") on multiple datasets from different sales regions. We use the task of identifying profitable (significantly different from zero in the positive direction) regions as a reasonable proxy for real-world analytical tasks: in EDA, data analysts often explore many facets of their data and report on only the subsets that appear important. Participants saw the following prompt:

> You are a manager supervising the sales of stores. In each region there are 200 stores. Your task will be to guess whether the average profit of the stores in a region is greater than zero. However, you only receive the sales data for 20 stores, and you have to make the decision based on this limited information.

There are two experimental variables in this study: (1) the number of sales regions, *i.e.,* the number of possible comparisons in each trial (m), a within-subjects variable with three levels: 12, 16 and 20; and (2) the type of uncertainty visualisation representation (*vis*), a between-subjects variable with three levels: **baseline** scatterplot, **ci** and **pdf**. Each experiment consisted of 30 trials broken into three blocks. Within each block, there were ten trials of the same m (number of graphs shown). The order of the blocks, and the order in which the trials were presented within each block were randomised.

In the beginning, participants went through a training which consisted of three parts. First, participants were presented with an onboarding page introducing them to the background story for the

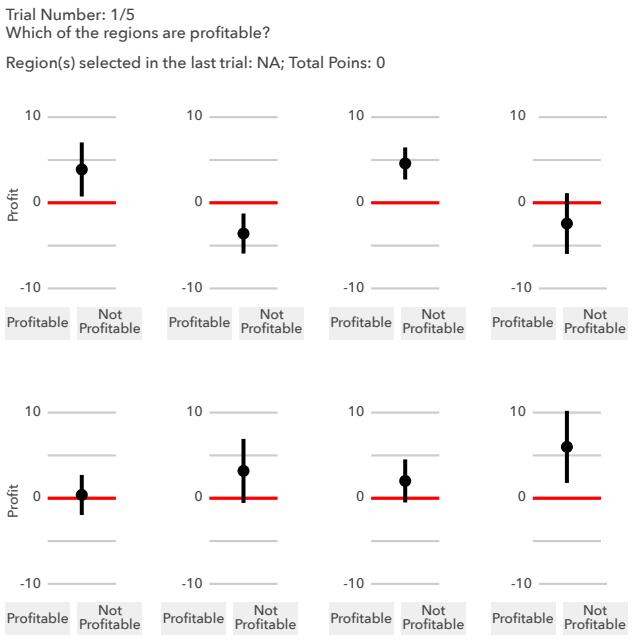


Figure 1: Experimental interface: participants are asked to indicate each region (individual graphs) as *profitable* or *not profitable*, based on data shown. In the training phase, participants are presented with 8 graphs (as shown here). In the test phase, participants are either presented with 12, 16, or 20 graphs.

[2]https://osf.io/xtsfz/

task and an explanation of how to interpret the visual representations used (see `supplement ▶ survey ▶ survey-example.pdf`). This was followed by an introduction to the incentives and an explanations of how participants' job performance would be evaluated (Figure 2). Finally, participants were presented with five training trials with eight graphs (Figure 1). After each training trial, participants were informed of their cumulative *points,* and given detailed feedback regarding which graphs were correctly selected, which were actually *positive* and which were actually *negative*. The *points* were reset to zero after the training phase ended. In the test phase, participants were only informed of their cumulative *points* after each trial; feeedback regarding the quality of their decisions were not provided. After completing all trials, participants were asked to report their strategy for completing the task in a free text field.

3.2 Simulating the Stimuli for a Multiple Comparisons Task

An important consideration in designing an experiment for investigating the multiple comparisons problem is the value of adjusting for multiple comparisons. Consider two statistical golems—one which makes statistical decisions without adjusting for multiple comparisons (**uncorrected**) and the other which makes statistical decisions while controlling for the *false discovery rate* at a particular α level (**Benjamini-Hochberg**). In our desired experiment, the performance of these two golems represent two benchmarks, against which we can compare participants' performance. This requires the difference between these two benchmarks to be larger than the measurement and estimation error in our experiment.

Ensuring this difference can be challenging. For example, in a previous iteration of this experiment, we generated stimuli such that the number of possible comparisons (m) varied between 8 and 12, the probability the null hypothesis is true (p_0) *i.e.* that the region shown in the graph was not profitable was set to 0.5, and participants were incentivised to control for *false discovery rate* at $\alpha = 0.05$. In this scenario, the maximum possible number of *false positive* is given by $p_0 m = 6$, when $m = 12$. The p-value is known to be uniformly distributed when the null hypothesis is true [35]. Thus, the expected number of *false positives* was $E(FP) = \alpha p_0 m = 0.3$. Thus the **uncorrected** statistical golem would be expected to make 0.3 *false positives* on average, while the **Benjamini-Hochberg** golem would be expected to make somewhere between 0 and 0.3 *false positives*[3] on average. The magnitude of the difference between the two benchmarks was too small to determine whether participants were performing any form of multiple comparisons correction.

However, by manipulating the values of p_0, m and α we can design an experiment with a greater difference between the **uncorrected** and **Benjamini-Hochberg** benchmarks. We can then use these benchmarks to investigate people's performance on a multiple comparisons problem. Another variable which impacts the *false discovery rate* is the (standardised) effect size (δ)—larger effect sizes are more easily "discoverable" (more *true positives* and less *false negatives*), whereas smaller effect sizes have the opposite effect. We conducted a grid search, varying $p_0 \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, $m \in \{10, 15, 20, 30, 50, 100\}$, $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\alpha \in \{0.2, 0.25, 0.3\}$, to determine a combination of p_0, m, δ and α that ensures a measurable difference between the **uncorrected** and **Benjamini-Hochberg** benchmarks. We also preferred smaller values of m (to better fit on participants' screens), smaller α (to more closely align with the typical $\alpha = 0.05$ used most commonly in NHST), and p_0 close to 0.5.

For each combination of variables, we ran 1,000 simulations, estimated the *false discovery rate* using **uncorrected** and **Benjamini-Hochberg** strategies, and calculated the mean and standard deviation of the *false discovery rate* from each strategy. Based on the results of our simulations, we selected values for the variables such that the estimated difference in the mean of the *false discovery rate* between the **uncorrected** and **Benjamini-Hochberg** (Δ) was at least twice as large as the relative standard deviation. More precisely, we determined the following values for the parameters: $m \in \{12, 16, 20\}, p_0 = 0.7, \delta = 0.4$ and $\alpha = 0.25$ to be reasonable for our experiment. We use these values to simulate the datasets used as stimuli for the participants which consisted of 10 trials for each value of m, resulting in 30 total trials. Due to variance, it is still possible that the difference in *false discovery rate* between the two normative strategies in our sample of 30 trials is smaller than the average difference across 1,000 simulations. As such, we use rejection sampling to ensure a minimum average difference, for each value of m, of at least the estimated average difference (Δ) from our simulations. Further details, including the code used for our simulation and the generated stimuli, can be found in `supplement ▶ R ▶ 01-experiment-design.Rmd`.

3.3 Incentives

We want an incentive structure which encourages participants to control the *false discovery rate* at the determined α level of 0.25 across multiple comparisons. This means that for every 100 discoveries, at least 75 of them should be true discoveries and less than 25 should be false discoveries on average [35]. Therefore we want 25 false discoveries to be as expensive as 75 true discoveries: the ratio between the *false positive* penalty and *true positive* reward is $75 : 25 = 3 : 1$. Since statistical tests are typically conducted with a power of 0.8, we similarly adjust the ratio between *false negative* and *true negative* rewards to be $4 : 1$. Moreover, we want to ensure our incentive structure does not encourage participants from adopting a trivial strategy where marking all of the regions are as not profitable (a *reject none* strategy). To make such a trivial strategy

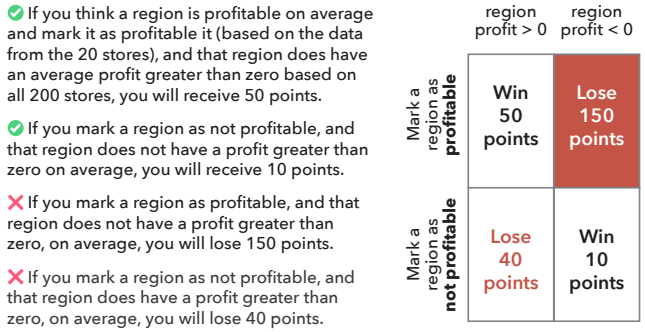


Figure 2: Incentive structure as shown to participants

[3]Because the BH procedure is dependent on the actual distribution of the p-value, we cannot provide a theoretical estimate of $E(FP|\text{strategy} = BH)$

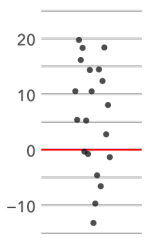
non-viable, we again rely on simulations to test various incentive structures and determine an appropriate reward ratio between a *true positive* and *true negative*. Based on our simulation results, we find a ratio of 5 : 1 between the rewards for *true positives* and *true negatives* to result in measurable differences between the payouts from using **Benjamini-Hochberg** and the trivial *reject none* strategies. The complete incentive matrix presented to participants is shown in Figure 2.

3.4 Uncertainty Displays

We varied the type of uncertainty visualisations (*vis*) between subjects to investigate whether visualisation type can improve the quality of participant decisions. In addition to a baseline which represents the data directly, we included two types of uncertainty visualisations, varying in amount of information (interval *vs.* density). While there are many other strategies for visualising uncertainty [34, 36, 38, 43, 47, 54, 63], we chose these types because they are either common in visual analytics tools or have been shown to improve decision-making under risk.

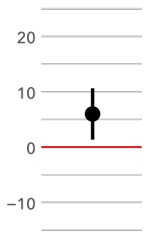
To decide whether a region is profitable, the participant needs to estimate arbitrary one-sided confidence intervals using each type of visualisation. For example, if the participant wants to control the false discovery rate to be under 0.25 for a single decision, they should decide that a region is profitable on average when the confidence mass below the zero line is less than 0.25, *i.e.*, $P(X < 0) < 0.25$.

Baseline: scatterplot



Displays of raw data without explicitly encoded uncertainty (such as scatterplots and strip plots) are easy to understand and commonly used in exploratory data analysis systems, making them a good baseline for comparison. These plots present the intrinsic uncertainty contained in raw data. An alternative baseline is to show only the mean, which hides all uncertainty information. However, showing no uncertainty is known to lead to bad performance [39] and can be unfair to the participant, whose pay depends on their decision-making quality.

CI: Mean Point Estimate and 50% Confidence Intervals



Point estimates of the mean with 95% confidence intervals are perhaps the most commonly used graphical plot for communicating uncertainty. Our task is designed such that, in the absence of multiple comparisons, participants should perform a one-tailed t-test and reject the null hypothesis at $\alpha = 0.25$. Hence, we visualise the 50% confidence interval, which is the equivalent of showing a 95% interval when rejecting the null hypothesis at $\alpha = 0.05$. We included these plots due to their familiarity and ubiquity. While the interval does not directly encode the confidence mass that a given region is profitable, it does provide some information on how reliable the estimate of the mean is for the broader population.

PDF: Probability Density Plot

Probability density plots use height to represent the probability density function (PDF) of the confidence distribution of the mean. This approach is similar to eyeball plots proposed by Spiegelhalter [65], which use width instead of height to encode density (similar variations are also called raindrop plots [2] or violin plots [17]). Density plots are a common uncertainty representation that shows information about the shape of the entire distribution. To make judgements about the confidence an estimate is greater than a particular value, the viewer must compare ratios of areas, which may be a difficult task and lead to lower accuracy [21, 36, 43].

3.5 Participant Information

For our pre-registered (https://aspredicted.org/2fw4r.pdf) study, we deployed the experiment on the Prolific crowd-sourcing research platform [52]. We recruited participants who were on desktop devices and fluent in English. We collected responses from 182 participants in total. Per our preregistration, participants who failed any of the three attention checks were not allowed to finish the study and therefore not included in the $N = 182$ sample. Two participants appeared to have retaken the survey after they were disqualified for failing the attention check, and were excluded from the analysis. This resulted in 180 participants for our final analysis, with 60 participants in each *vis* (between-subjects) condition. The median completion time was approximately 26 minutes, and the average wage was $11.44/hr ($14.40/hr including bonuses). All participants who performed better than the uncorrected strategy *i.e.* had accured greater than -7650 cumulative *points* (60%; 108 / 180) received a bonus which was awarded in stepwise increments of $0.5 up to a maximum of $4.5

4 STATISTICAL MODELING AND ANALYSIS

We describe the methods involved in our pre-registered quantitative and qualitative analyses.

4.1 Quantitative Analysis

Our Bayesian hierarchical model is specified in the Wilkinson-Rogers-Pinheiro-Bates syntax [3, 53, 71] as:

$$
\begin{aligned}
&1: \quad outcome \mid trials(m) \;\sim\; multinomial(p) \\
&2: \quad softmax(p_k) \;=\; vis \times trial \times block \times m + \\
&3: \qquad\qquad\qquad\quad (trial \times block \times m \mid participant)
\end{aligned}
$$

Line 1: decision outcomes modeled as a multinomial distribution. There are four possible outcomes for each decision: *true positive* (TP), *true negative* (TN), *false positive* (FP), and *false negative* (FN), and we use a multinomial distribution for the likelihood to estimate the mean probability for each *outcome*. The multinomial distribution estimates the probability of each outcome as a vector, $p = \{p_1, p_2, p_3, p_4\}$. $trials()$ is a `brms`[4] keyword that specifies how

[4]More explanation on `trials()`: https://cran.r-project.org/web/packages/brms/vignettes/brms_customfamilies.html

many decisions ("*trials*") are in each observation, and the *m* integer variable indicates that the participant performed 12, 16 or 20 comparisons in a particular trial.

Line 2: population-level effects. In our experiment, *vis* is a between-subjects variable for different uncertainty displays; *trial*, which indicates the trial number within each block (1-10), captures any potential learning or fatigue effects over the course of the experiment; *m* encodes the number of graphs presented in the trial; and *block* captures potential order and learning effects. We encode *vis*, *m* and *block* as discrete variables, and *trial* as a continuous variable (*i.e.*, with a linear effect). Since we want to compare decision outcomes across these variables, they are specified as population-level effects (predictors) with interactions.

Line 3: group-level effects to account for individual differences and repeated trials. Different participants can have different decision capacities, and the effects of *trial* and *block* (learning and order effects as a participant goes through the trials) and *m* variables can be different for each participant. To account for these individual differences, we use a multilevel model, including varying slopes and intercepts for the effect of *trial* and *m* within each *participant* (the grouping variable). In addition, participants completed repeated trials within each condition. Using multilevel models and grouping by participants or other clusters often gives improved estimates for repeated trials [48].

Model Run and Validation

We fit the model with the `brms` R package [11] and used weakly-informed priors. The model ran four chains with 5,000 warmup samples and 5,000 post-warmup samples each, thinned by 5 for a final total sample size of 4,000. We assessed convergence using the Gelman-Rubin diagnostic ($\hat{R} = 1.00$ for all population-level parameters, correlations and standard deviations) and the (bulk and tail) effective sample sizes ($Tail\_ESS_{min} \approx 3,000$). One method for model validation is with posterior retrodictive checks. Instead of predicting responses for an average participant as results, here our model *retrodicts* existing participant responses [7]. Figure 3 shows the ●– posterior retrodictives (mean and 95% credible interval) alongside with the participant response averages. In this visual comparison, the means of the posterior retrodictives are close to the average responses in most cases. The retrodictive checks do not show any signs of consistent model bias, and indicates a good model fit.

4.2 Qualitative Analysis

We preregistered an exploratory qualitative analysis of the participants' self-reported free text responses for the strategies they used to complete the main experimental task. We employed a hybrid coding approach on these responses. In line with our research aims, we coded whether participants reported a *sensitivity to the incentive structure* we presented (*i.e.*, expressing caution over the cost of false

Figure 3: Average participant response and validation of the model by recovering the participant response means from the model fit. The columns show average predicted and observed probabilities of true positive, true negative, false positive and false negative for a given trial. The false discovery rate is computed as $FDR = FP/(TP + FP)$. The tapered ends of each gray bar represents the complementary cumulative distribution function (CCDF) of the posteriors of the average probability. As indication for a good model fit, the means of posteriors are close to the means of participant response in most cases, without consistent bias.

positives). We also coded whether participants self-reported employing a *correction strategy*: that is, incorporating the number of comparisons into their decision making, or changing their decision-making strategy in response to past performance on previous trials. We were also interested in whether different uncertainty visualisations would promote different strategies for task completion, but given the expected diversity of strategies we relied on emergent rather than pre-defined codes. A paper author acted as initial coder employing our two closed codes and then performed open coding to generate initial codes representing categories of strategies. A second paper author then independently coded the responses; the two coders then met to discuss mismatches and ambiguities to generate a final consensus codebook and consensus labels for use in our thematic analysis. The full codebook, per-rate responses, and analysis of inter-rater reliability are included in `supplement ▶ qualitative-analysis ▶ qual-responses.xlsx`.

5 RESULTS

Our model estimates the number of *true positive*, *true negative*, *false positive* and *false negative* for a particular trial, for each *vis* and *m* condition. We can divide these estimates by *m* to obtain the probability of making a *true positive*, *true negative*, *false positive* and *false negative*. Figure 3 reports participants' • **average** and ●– posterior retrodictive estimates of the probability of *true positives*, *true negatives*, *false positives* and *false negatives*, reflecting participants' overall decision quality. Since our research questions are not concerned with other potential factors such as learning or order effects, we marginalise over the *trial* and *block* variables (see Appendix A).

5.1 Do Participants' Decisions Reflect Implicit Multiple Comparisons Correction?

Our first research question concerns whether participants' decisions reflect multiple comparisons correction. We compare the estimated probability of a *false positive* (Figure 4) by an average participant, for each *vis* and *m* conditions, averaged over *trial* and *block*, against the normative **uncorrected** benchmark—the expected number of *false positive* and *false discovery rate* from using an **uncorrected** strategy. If we find participants having lower probability of a *false positive* when compared to the **uncorrected** benchmark it would suggest that participants may be performing some form of multiple comparisons corrections.

Overall, we find that the average participant in the **ci** condition is expected to make 0.148 (95% credible interval (CI): [0.124, 0.175]), 0.142 (95% CI: [0.118, 0.167]) and 0.142 (95% CI: [0.120, 0.167]) *false positives* on average, when the number of possible comparisons (*m*) is 12, 16 and 20 respectively; the average participant in the **pdf** condition is estimated to make 0.104 (95% CI: [0.086, 0.123]), 0.124 (95% CI: [0.103, 0.147]) and 0.132 (95% CI: [0.111, 0.158]) *false positives*. As shown in Figure 4, this is lower than the normative benchmark of using an **uncorrected** strategy but greater than the **Benjamini-Hochberg** benchmark. On the other hand, the average participant in the **baseline** scatterplots condition is expected to make more *false positives* than the **uncorrected** benchmark. This suggest that participants, on average, are likely performing some form of multiple comparisons to be making fewer *false positives* than the **uncorrected** strategy. However, a typical participant is likely not able to exactly control for *false discoveries* at the desired α-level, as incentivised, as the *false discovery rate* across all *vis* conditions exceeds $\alpha = 0.25$ (Figure 5). Additionally, this falls short of the performance achieved by procedures which can guarantee *false discovery rate* control at any pre-determined α-level such as **Benjamini-Hochberg**.

As the number of comparisons, *m*, changes, the criterion for rejecting the null hypotheses in normative procedures such as Benjamini-Hochberg or Bonferroni becomes stricter. In our results, we observe that the probability of a *false positive* remains comparable. This likely suggests that, while some participants may be performing some form of multiple comparisons correction, they may not be adjusting their strategy as the number of possible comparisons changes. Additionally, in an (not pre-registered) exploratory

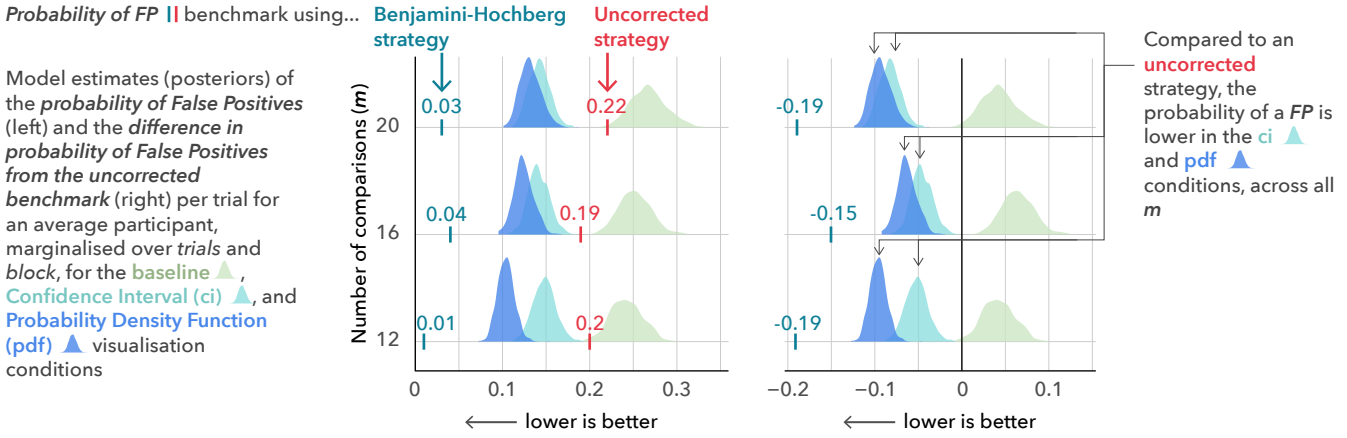


Figure 4: Estimated posterior probability of a false positive when the number of possible comparisons, m, is 12, 16 and 20 graphs, represented as densities (left). The corresponding differences in the estimated posterior probability of a false positive from the uncorrected benchmark, represented as densities (right).

analysis we examine the probability of rejecting the null hypothesis, and find that the average participant in the CI and PDF conditions are likely to reject the null hypothesis (that the region is not profitable) less frequently than the **uncorrected** strategy, lending further evidence to suggest that participants are likely performing some form of multiple comparisons correction (see `supplement ▶ R ▶ 04-modeling_and_analysis.Rmd`).

From the self-reported strategy data from our 180 participants, while 37/180 reported making adjustments to their strategy, only two reported doing so in an a priori-way, based on the number of comparisons. One participant reported that they "[h]ad a look at the graphs overall and see how all stores looked before making a decision.", indicating at least an awareness of potential issues in multiple comparisons, while another indicated that they "[a]dded 50% to length of bars to see if it still indicated a profit (mad I know!)" which is somewhat analogous to an increased level of significance produced by something like a Bonferroni correction. More common was reporting an adaptive or reactive strategy based on feedback from the trials, in which participants became more conservative in reaction to a low score. E.g., "[i]nitially I was very risky and went with all stores with over 50% chance of being profitable. I changed to very conservatively picking only sure bets as my score was very low." Or, from another participant, "[i]nitially I largely trusted the distribution and if less than about 25% of the bell curve was below the profit line then I would mark it as profitable. But this didn't work very well and gradually I began only marking as profitable if about 90% was above the line and the middle of the bell curve was at around about 4 or higher."

An interesting reaction was a (mal-)adaptive strategy in response to negative feedback, which was reporting adding randomness or otherwise giving up. E.g., "Tried to play it safe and go for 'profitable/not profitable' when it looked like a sure thing, but that didn't go too well for me! Started taking more risks/gambles as my points spiralled and by the end, there wasn't really too much of a strategy" from one participant, or "If the [business analyst] put the dot above the profit line, generally I said it would be profitable. However when I kept accruing negative points, I did try and throw some random ones in there to see if it helped but unfortunately it did not and I kept getting further into the negative numbers." In all, nine participants indicated employing some degree of randomness in their guesses, from giving up or "spiralling" as mentioned above, to those who reported trying to "...throw some random ones in there to see if it helped."

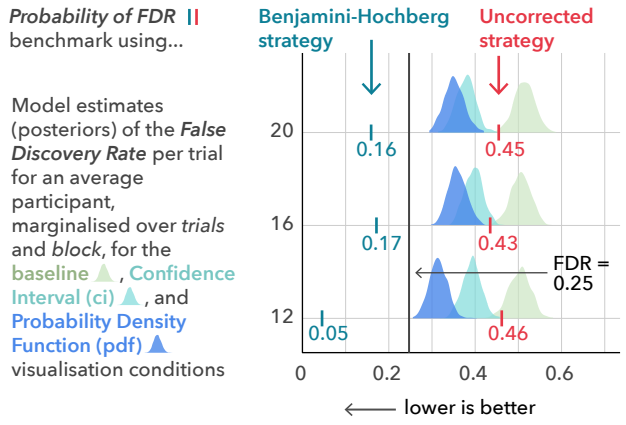


Figure 5: Posterior estimates of the false discovery rate points per trial, based on our incentive structure, for each visualisation (vis) and number of possible comparisons (m)

5.2 Do Uncertainty Visualisations Affect Decision Quality?

Our second research question is concerned with the impact of different uncertainty representations on participants' decision quality, and more specifically, in their ability to correct for multiple comparisons. As a between-subjects condition, we tested three uncertainty visualisations—a (**baseline**) scatterplot of the data with no summary or uncertainty information, a discrete interval representation of a summary statistic (mean) and the associated uncertainty using a point estimate and 50% confidence intervals (**ci**), and a continuous uncertainty representation of the mean using a probability density function (**pdf**).

Figure 6A shows the estimated probability of *false positive,* for the average participant, marginalised across all *trial, block* and *m* variables. Compared to the baseline, the average participant in both uncertainty representations is likely to have a lower probability of a *false positive* with **pdf** showing marginally greater reduction (mean: -0.13; 95% CI: [-0.18, -0.10]) compared to the **ci** (mean: -0.11; 95% CI: [-0.15, -0.07]). As seen in Figure 4, these differences are consistent across all levels of *m*.

In addition to comparing *false positive,* we can also compare participants' performance by calculating *false discovery rate* and *points,* using the estimated quantities of *true positive, true negative, false positive* and *false negative.* Figure 6B shows the estimated *false discovery rate* marginalised across all *trial, block* and *m* variables. Here, we observe that the average participant in the **baseline** condition has a higher *false discovery rate* than the **uncorrected** benchmark. However, while the typical participant in the **pdf** and **ci** conditions are not able to control for False Discoveries at the desired α-level of 0.25, these uncertainty representations nevertheless may lead to a lower *false discovery rate* when compared to both an uncorrected strategy, in addition to the **baseline**. Figure 6C shows the *points,* estimated to be accumulated by a typical participant, marginalised across all *trial, block* and *m* variables. Like the previous metrics, we again observe that the average participant in the **baseline** condition performs worse than the **uncorrected** benchmark, while the typical participant in the **pdf** and **ci** conditions outperform both the uncorrected strategy and the **baseline**.

Our qualitative responses indicate that one potential factor in the differences in performance observed in the different visualisation types may be a result of the strategies that a reader can adopt based on the affordances of those charts. 28/60 participants in the **pdf** condition report using the proportion of the visualised density that overlaps 0 (or some other value of profit) to help them make a decision. For instance, a participant reported that "anything with 25% or more of the distribution below the red line went immediately to 'unprofitable"' and another that "I eventually realised that only tasks where <5% of the bell curve fell below the profit line were worth recommending as profitable." This density information

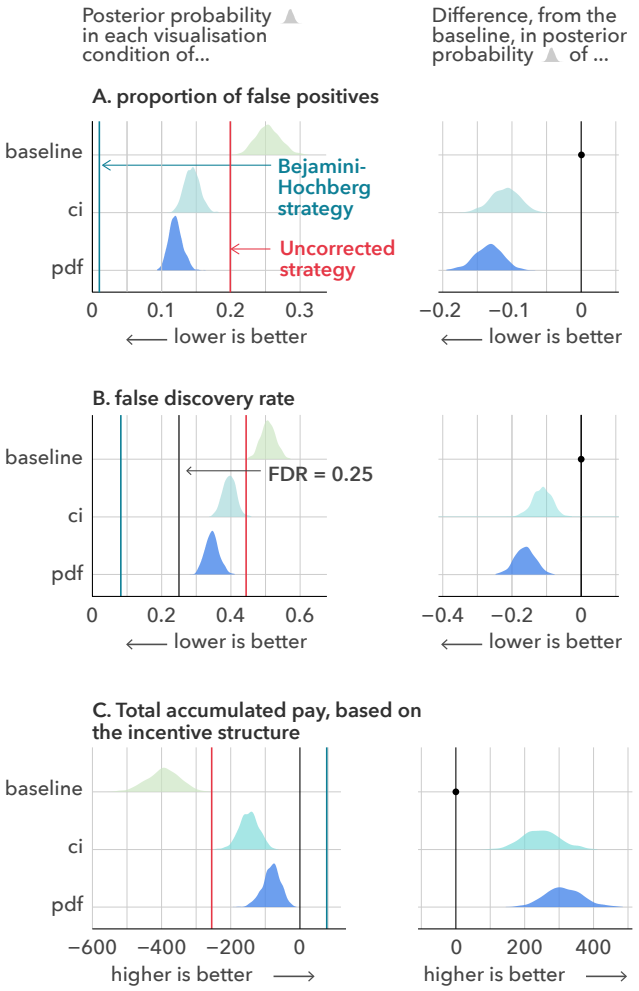


Figure 6: Posterior estimates of: (A) the probability of making a false positive, (B) the false discovery rate, and (C) the average accumulated points per trial, based on our incentive structure (left) as well as the difference in the posterior estimates, for each of these metrics, from the baseline (right)

was not directly available in the other charts, and so participants reported other strategies.

In keeping with prior work that reports that confidence intervals encourage dichotomous thinking [17, 24], 8/60 participants in the **ci** condition reported using whether or not the confidence interval overlapped 0 (or some other threshold) to make their decision. E.g., "I simply selected profitable for the the ones where the whole range was above 0" or selecting unprofitable "[i]f the bar dropped below the red line at all". As with the *density overlap* strategy above, this information was only directly available in the **ci** condition.

For the **baseline** condition, where no data about mean or variability was directly presented, participants often used other strategies. The majority of participants (39/60) would count the number of points above and/or below zero, and use the resulting total to either make a decision directly, or as part of a process of estimating the mean in order to make a final decision. E.g., "I figured the more dots above the red line there were, the more profitable the store was". We note that this sort of counting is a not a true calculation of mean value and, in keeping with the other perceptual proxies that have been shown to factor into the extraction of mean values from graphs [77], can produce incorrect or biased results. 12/39 participants who reported employing dot counting strategies counteracted for this fact by giving outliers special treatment, e.g. "I looked at how dense the clusters were on either side of the 0 mark to decide whether or not they were profitable or not then looked for any outliers like a lone sample or two that were above or below the average and decided if [I] thought they were high or low enough to counteract my initial assumption" from one participant, and "first [I'd] see which side had most points, but then if it was close or if sides had a noticeable quantity of points further from 0 [I'd] weigh them higher, as 1 point at lets say 25 is worth 10x points around the 2.5 mark."

6 DISCUSSION

6.1 The Potential Promise of Uncertainty Representations

The visual representations used in prior work did not directly visualise the mean and the associated uncertainty in the mean (standard error) that is necessary for the inferential tasks that participants in prior work [78] were asked to perform. Instead, this information was left implicit—participants could get a sense of the mean and the standard error based on the visualised data sample, and the spread of the data sample. The average participant in our **baseline** condition performed poorly across all metrics, suggesting that participants may be struggling to recover such inferential statistical estimates from a plot which does not visualise it directly.

On the other hand, we find that the average participant, in the **ci** and **pdf** conditions, makes fewer false discoveries than the golem using an **uncorrected** strategy. Along with the qualitative descriptions of the strategies used in performing this task, this result indicates that participants in these two conditions are potentially making some adjustments for multiple comparisons. As participants' *false discovery rates* are still greater than the incentivised $\alpha = 0.25$ level (Figure 6), and falls far short of the *false discovery rates* achieved by the golem using the **Benjamini-Hochberg** strategy, it appears that the result of this adjustment is perhaps an imperfect multiple comparisons correction.

However, the estimated averages hide a great degree of variability [28, 79], both in participants' performance and their reported strategies. From Figure 7, we observe that approximately 23% and 32% of the participants in the **ci** and **pdf** conditions respectively (0% in the **baseline** condition) have a positive *points* on average across the trials, which is close to the **Benjamini-Hochberg** benchmark (96 points on average). This suggests that a subset of participants are in fact able to optimise for the incentives and perform almost as well as the best statistical golem. Conversely, 80%, 48% and 35% of the participants in the **baseline**, **ci** and **pdf** conditions respectively perform worse than **uncorrected** golem, with some performing considerably worse. We conduct an exploratory analysis to help us understand what strategies might participants be using. We believe that a small subset of the participants may be employing a mix of a

mean strategy (considering only the mean of the visualised data, and indicating that a "region is profitable" if this mean is greater than zero) and answering at random (not responding to the stimulus). This is supported by some the qualitative responses as well (e.g., a participant in the **pdf** condition who claimed "if the middle of the shape was above zero [I] said profitable, if not then I said not profitable"), and suggests that there exists a subset of people who, even when presented with uncertainty information, are likely to use sub-optimal strategies to perform this task.

We recommend that designers of EDA systems should explicitly visualise uncertainty, if they expect the user to be performing inferential tasks. However, this may not be sufficient. Certain forms of uncertainty visualisations have different affordances, and may promote certain decision-making strategies that rely on these affordances. For instance, a proportion of participants reported employing binary decision criteria in the **ci** condition (which has been criticised in the past for presenting uncertainty information in a dichotomous way), and a proportion of participants counted the number of dots above and below zero as a perceptual proxy for estimating mean and/or variability. However, the wide variety of reported strategies both within and across conditions points to differing levels of expertise and experience with uncertainty visualisation: designs could therefore draw from the strategies used by participants who excel at this task in order to help those who may struggle with it.

6.2 The Precarious Entanglement of Incentives and Evaluation

The expected proportion of false discoveries, for the average participant in the **ci** and **pdf** conditions, was 39% and 34% respectively. In contrast, prior work found *false discovery rates* of 60% [78]. This raises the question: is a *false discovery rate* of 34% good? Or conversely, is a *false discovery rate* of 60% bad? The answer surely varies depending on the data analysis and decision-making context. This decision-making context can and should be translated into incentive structures. When analysts decide which data patterns may be real, or perhaps "statistically significant", they usually consider the context. For example, there might be a budget constraint against taking action on too many discoveries, or false discoveries might lead to adding an ineffective new product feature that loses users and revenue. As we demonstrate in this paper, we can encode a false discovery rate threshold, such as $\alpha = 0.25$, through experimental incentives. Only when this additional context is provided can we evaluate participants' decision-making quality and performance. In the absence of such explicit incentives, it is impossible to determine whether the 60% false discoveries reported in prior work is excessive. Rather, a number of alternative, plausible explanation may explain participants behavior in the study. For instance, participants in the study may have deemed the value of *true positive* to be arbitrarily high, and have attempted to maximise this implicit incentive structure even though they were actually evaluated on minimising *false positive*.

The participants in our study were unable to control their *false discovery rate* at the incentivised $\alpha = 0.25$ level. Even in the **ci** and **pdf** conditions, participants' *false discovery rates* were 10–15 percentage points greater than the desired *false discovery rate,* based on our incentives. However, they did, on average, adjust their behaviour in light of the incentives, as evidenced by the lower *false discovery rate* when compared to the **uncorrected** benchmark as well as from their qualitative descriptions. This suggests that well-designed and explicit incentive structures *may* encourage better decisions and more realistically reflect the quality of the EDA system in terms of the multiple-comparison problem.

6.3 Should We Care About False Discoveries Only?

We note that the potential improvements in *false positive* and *false discovery rate* come at the cost of lower *true positive* and higher *false negative*. This is the case for both the statistical golems (**Benjamini-Hochberg** *v.s.* **uncorrected**) and the uncertainty displays (**baseline** *v.s.* **ci** or **pdf**). This is to be expected, as our incentives penalised *false positives* most strongly. However, depending on the decision context other incentive structures besides the one we tested may be valid. For instance, it is possible that the cost of *false negatives* is

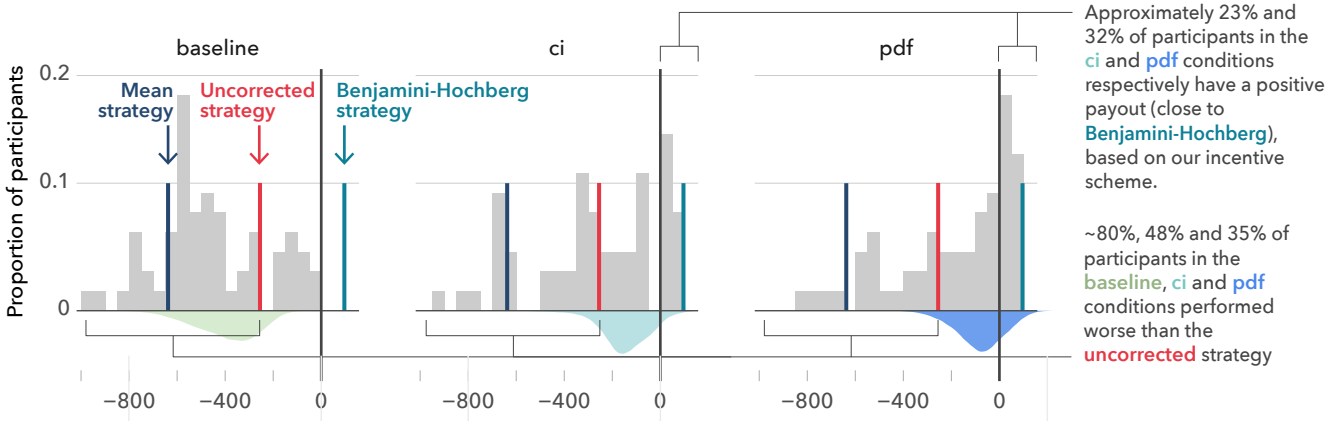


Figure 7: Distribution of participants' average accumulated points per trial (*points*), and posterior estimates of average *points*, marginalised over the participants in our study, across each condition.

much greater than the rewards from *true positives* or *true negatives* in certain scenarios.

Due to our experimental design, we are unable to disentangle the effects of the incentive structure presented to participants and the uncertainty representations used on participants' performance. While it may be possible that such uncertainty representations provide participants with a holistic representation of the information required to perform the task, such that they still may perform better even under different evaluation metrics, further research is necessary. Designers of EDA systems should carefully choose uncertainty visualisations that best suit their evaluation metrics.

6.4 Limitations and Future Work

Impact of immediate feedback. In our experiment, we decided to provide participants with immediate feedback regarding their performance on both the previous trial, and their performance overall, in the form of "points." This design choice had some impact on participants' behavior in the experiment, and even across different trials. A handful of participants described using a reactive strategy—they became more conservative about rejecting the null if they scored a lot of negative points in a previous trial. In real world decision-making contexts, however, the feedback may not always be immediate, but may be delayed, or even ambiguous. In the absence of immediate feedback, we speculate that participants may be less reactive; it is also possible this may increase the likelihood of users failing to account for multiple comparisons when performing exploratory data analysis. In scenarios where immediate feedback may not be feasible, a possible solution could be to proactively train participants on the need to account for multiple comparisons. We hope to explore the impact of feedback presentation in future work.

Impact of the magnitude of incentives. One design choice we made in our experiment was to use comparatively large values for the incentives. While this was intended to make sure that participants did not perceive the difference between a correct and incorrect decision in the task as trivial, it also meant that participants could potentially end up with a large, negative number of points. Figure 7 shows that there were indeed some participants who accrued very large, negative points, raising the question of how our specific incentive structure impacted performance. A small number of participants (9/180) reported adopting risky strategies or randomly guessing due to poor performance in a previous trial and/or an accumulation of negative points (see §5.1). We speculate that a different incentive structure (*e.g.*, if the rewards are scaled down by a factor of 10) may have reduced the number of participants adopting such mal-adaptive strategies. In general, we believe that the psychological and statistical impact of differing incentive structures in visual analytics (both in experimental settings and in practice) is understudied.

The broader space of uncertainty representations. Recent work on uncertainty visualisations have recommended many other forms of representations such as dotplots [21, 41, 43], hypothetical outcome plots [34, 42], gradient plots [17] etc. As a preliminary exploration of the impact of uncertainty representations, we decided to focus on two uncertainty representations which are commonly used and which provide successively greater degree of uncertainty information. We hope to explore the impact of these alternative uncertainty representations, all of which, like the **pdf**, convey complete distributional information, in future work.

7 CONCLUSION

We set out to improve the evaluation of multiple comparison problems in EDA systems in two ways: using realistic decision incentives and uncertainty visualisations. We conduct an experiment to investigate the impact of providing explicit incentives and using uncertainty representations on participants decision-making quality. We found that, for an average participant, uncertainty representations such as confidence intervals or probability density functions may lead to *false discovery rates* which are lower than the uncorrected (no corrections for multiple comparisons) benchmark, but higher than the Benjamini-Hochberg (a multiple comparisons correction procedure) benchmark. However, in the absence of uncertainty information, participants perform worse than the uncorrected benchmark. In a qualitative analysis of users' strategies, we find that some participants may be adapting to the information presented to them and employing strategies which produce a similar effect to an imperfect multiple comparisons correction procedure.

ACKNOWLEDGMENTS

We would like to thank Fumeng Yang and Ziyang Guo for their thoughtful feedback on this research. We also thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. 2019. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

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

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67 (Oct. 2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[4] Leilani Battle and Jeffrey Heer. 2019. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum* 38, 3 (June 2019), 145–159. https://doi.org/10.1111/cgf.13678

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

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[7] Michael Betancourt. 2020. *Towards A Principled Bayesian Workflow.* https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

[8] Carsten Binnig, Lorenzo De Stefani, Tim Kraska, Eli Upfal, Emanuel Zgraggen, and Zheguang Zhao. 2017. Toward Sustainable Insights, or Why Polygamy is Bad for You. In *Conference on Innovative Data Systems Research.* https://api.semanticscholar.org/CorpusID:15852012

[9] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. 2015. Storytelling in Information Visualizations: Does It Engage Users to Explore Data?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15).* Association for Computing Machinery, New York, NY, USA, 1449–1458. https://doi.org/10.1145/2702123.2702452

[10] Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, 1906 (2009), 4361–4383.

[11] Paul-Christian Bürkner. 2017. Advanced Bayesian Multilevel Modeling with the R Package brms. *R J.* 10 (2017), 395. https://api.semanticscholar.org/CorpusID:54534499

[12] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. 1999. Using vision to think. In *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 579–581.

[13] Stephen Casner and Jill H Larkin. 1989. *Cognitive efficiency considerations for good graphic design.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY ....

[14] Remco Chang, Caroline Ziemkiewicz, Tera Marie Green, and William Ribarsky. 2009. Defining insight for visual analytics. *IEEE Computer Graphics and Applications* 29, 2 (2009), 14–17.

[15] Yeounoh Chung, Sacha Servan-Schreiber, Emanuel Zgraggen, and Tim Kraska. 2018. Towards Quantifying Uncertainty in Data Analysis & Exploration. *IEEE Data Eng. Bull.* 41, 3 (2018), 15–27.

[16] Kristin A Cook and James J Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics.* Technical Report. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

[17] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2142–2151.

[18] Michael Correll and Michael Gleicher. 2015. Implicit Uncertainty Visualization: Aligning Perception and Statistics. In *Workshop on Visualization for Decision Making under Uncertainty.* https://api.semanticscholar.org/CorpusID:16691049

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

[20] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300295

[21] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173718

[22] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 348 (2013), 1–17.

[23] Miriam Greis, Aditi Joshi, Ken Singer, Albrecht Schmidt, and Tonja Machulla. 2018. Uncertainty Visualization Influences How Humans Aggregate Discrepant Information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174079

[24] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besancon. 2021. Can visualization alleviate dichotomous thinking? Effects of visual representations on the cliff effect. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2021), 3397–3409.

[25] Jerry L Hintze and Ray D Nelson. 1998. Violin plots: a box plot-density trace synergism. *The American Statistician* 52, 2 (1998), 181–184.

[26] Harry Hochheiser and Ben Shneiderman. 2004. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization* 3, 1 (2004), 1–18.

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

[28] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. 2020. How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376454

[29] Heike Hofmann, Lendie Follett, Mahbubul Majumder, and Dianne Cook. 2012. Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2441–2448.

[30] Jessica Hullman. 2016. Why Evaluating Uncertainty Visualization is Error Prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (Baltimore, MD, USA) *(BELIV '16)*. Association for Computing Machinery, New York, NY, USA, 143–151. https://doi.org/10.1145/2993901.2993919

[31] Jessica Hullman. 2019. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 130–139.

[32] Jessica Hullman, Eytan Adar, and Priti Shah. 2011. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2213–2222.

[33] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 903–913.

[34] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE* 10, 11 (11 2015), 1–25. https://doi.org/10.1371/journal.pone.0142444

[35] H. M. James Hung, Robert T. O'Neill, Peter Bauer, and Karl Kohne. 1997. The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics* 53, 1 (1997), 11–22. http://www.jstor.org/stable/2533093

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

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

[38] Amit Jena, Ulrich Engelke, Tim Dwyer, Venkatesh Raiamanickam, and Cecile Paris. 2020. Uncertainty visualisation: An interactive visual survey. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 201–205.

[39] Susan L Joslyn and Jared E LeClerc. 2012. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied* 18, 1 (2012), 126.

[40] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 202, 14 pages. https://doi.org/10.1145/3290605.3300432

[41] Alex Kale, Matthew Kay, and Jessica Hullman. 2021. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

[42] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 892–902.

[43] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[44] Christoph Kinkeldey, Alan M MacEachren, Maria Riveiro, and Jochen Schiewe. 2017. Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science* 44, 1 (2017), 1–21.

[45] Christoph Kinkeldey, Alan M MacEachren, and Jochen Schiewe. 2014. How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal* 51, 4 (2014), 372–386.

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

[47] Alan M. MacEachren. 1992. Visualizing Uncertain Information. *Cartographic Perspectives* 13 (Jun. 1992), 10–19. https://doi.org/10.14714/CP13.1000

[48] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC.

[49] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing Visualization Mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376420

[50] Limor Nadav-Greenberg and Susan L Joslyn. 2009. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making* 3, 3 (2009), 209–227.

[51] Chris North. 2006. Toward measuring visualization insight. *IEEE computer graphics and applications* 26, 3 (2006), 6–9.

[52] Stefan Palan and Christian Schitter. 2018. Prolific.ac — A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[53] José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, and R Maintainer. 2017. nlme : Linear and nonlinear mixed effects models. R package version 3.1-103. *http://cran.r-project.org/web/packages/nlme/index.html* (2017). https://cir.nii.ac.jp/crid/1570854174288831360

[54] Kristin Potter, Paul Rosen, and Chris R. Johnson. 2012. From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches. In *Uncertainty Quantification in Scientific Computing*, Andrew M. Dienstfrey and Ronald F. Boisvert (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 226–249.

[55] Xiaoying Pu and Matthew Kay. 2018. The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 37–45. https://doi.org/10.1109/BELIV.2018.8634103

[56] P Samuel Quinan, Lace M Padilla, Sarah H Creem-Regehr, and Miriah Meyer. 2015. Towards ecological validity in evaluating uncertainty. In *Proceedings of Workshop on Visualization for Decision Making Under Uncertainty (VIS'15). http://vdl. sci. utah. edu/publications/2015_vdmu_ecologicalvalidity*.

[57] Khairi Reda, Andrew E Johnson, Michael E Papka, and Jason Leigh. 2016. Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization* 15, 4 (2016), 325–339.

[58] Purvi Saraiya, Chris North, Vy Lam, and Karen A Duca. 2006. An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 12, 6 (2006), 1511–1522.
[59] Abhraneel Sarma, Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du, Eunyee Koh, and Matthew Kay. 2023. Evaluating the Use of Uncertainty Visualisations for Imputations of Data Missing At Random in Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 602–612. https://doi.org/10.1109/TVCG.2022.3209348
[60] Abhraneel Sarma, Kyle Hwang, Jessica Hullman, and Matthew Kay. 2024. Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642375
[61] Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. 2023. Multiverse: Multiplexing Alternative Data Analyses in R Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 148, 15 pages. https://doi.org/10.1145/3544548.3580726
[62] Sacha Servan-Schreiber, Olga Ohrimenko, Tim Kraska, and Emanuel Zgraggen. 2019. STAR: Statistical Tests with Auditable Results. *arXiv:1901.10875 [cs, stat]* (2019). http://arxiv.org/abs/1901.10875 arXiv: 1901.10875.
[63] Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. 2008. Revealing Uncertainty for Information Visualization. (2008), 376–379. https://doi.org/10.1145/1385569.1385637
[64] Rachel A Smith, Timothy R Levine, Kenneth A Lachlan, and Thomas A Fediuk. 2002. The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway ANOVA. *Human Communication Research* 28, 4 (2002), 515–530.
[65] David J Spiegelhalter. 1999. Surgical audit: statistical lessons from Nightingale and Codman. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162, 1 (1999), 45–58.
[66] Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale. 2012. The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1461–1470. https://doi.org/10.1145/2207676.2208607
[67] Edward R Tufte. 2001. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT.
[68] John W Tukey. 1965. The technical tools of statistics. *The American Statistician* 19, 2 (1965), 23–28.
[69] John W Tukey. 1977. *Exploratory data analysis*. Vol. 2. Addison-Wesley Pub. Co.
[70] H. Wickham, D. Cook, H. Hofmann, and A. Buja. 2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov 2010), 973–979. https://doi.org/10.1109/TVCG.2010.161
[71] GN Wilkinson and CE Rogers. 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 22, 3 (1973), 392–399. https://doi.org/10.2307/2346786
[72] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.
[73] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2017. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2648–2659. https://doi.org/10.1145/3025453.3025768
[74] Yifan Wu, Ziyang Guo, Michails Mamakos, Jason Hartline, and Jessica Hullman. 2023. The Rational Agent Benchmark for Data Visualization. *IEEE transactions on visualization and computer graphics* 22, 1 (2023), 649–658.
[75] Itai Yanai and Martin Lercher. 2020. A hypothesis is a liability. *Genome Biology* 21, 1 (Dec. 2020), 231, s13059–020–02133–w. https://doi.org/10.1186/s13059-020-02133-w
[76] Ji Soo Yi, Youn-ah Kang, John T. Stasko, and Julie A. Jacko. 2008. Understanding and Characterizing Insights: How Do People Gain Insights Using Information Visualization?. In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization* (Florence, Italy) *(BELIV '08)*. Association for Computing Machinery, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/1377966.1377971
[77] Lei Yuan, Steve Haroz, and Steven Franconeri. 2019. Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review* 26 (2019), 669–676.
[78] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 479, 12 pages. https://doi.org/10.1145/3173574.3174053
[79] Sam Zhang, Patrick R. Heck, Michelle N. Meyer, Christopher F. Chabris, Daniel G. Goldstein, and Jake M. Hofman. 2023. An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences* 120, 33 (2023), e2302491120. https://doi.org/10.1073/pnas.2302491120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2302491120
[80] Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. 2017. Controlling False Discoveries During Interactive Data Exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 527–540. https://doi.org/10.1145/3035918.3064019

A MARGINALISATION OF POSTERIOR ESTIMATES

To answer our research questions, we need to estimate the marginal effect of *vis* and *m* on the posterior probability of the four decision outcomes for an average participant. However, because *trial* and *block* were included population-level effects in our regression model, we need to average out the effects of these two variables on our parameters of interest. For instance, if we want to show the effect of *vis* on participants' decisions, we would need to marginalise over the predictors *trial, block* and *m*. Figure 8 describes the process of computing the average marginal effect[5] by marginalising over the predictor *trial*. This can then be repeated for other the other predictors.

Marginalising over *trial* is justified because the learning effects captured by *trial* do not alter our main results (section 5). With exploratory comparisons, we find that as participants progress and potentially improve through the trials, *vis* affects FDRs in similar patterns when we look at $m = 12$, $m = 16$ and $m = 20$ separately. If we compare the FDRs in the last trial of the first experiment block and that of the second block, and do the same comparison for the last five trials of each block, we see similar differences in FDR among visualization types, even though participants might have gotten better at the task by the end of the second block of the experiment. Details of this exploratory comparison are in `supplement► R ► 04-modeling_and_analysis.Rmd`.

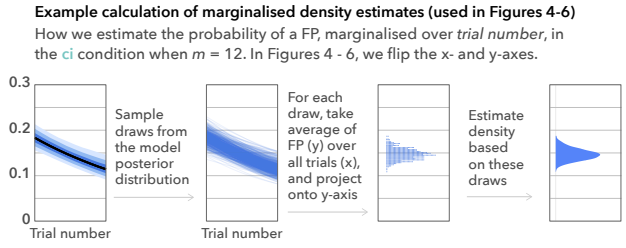


Figure 8: How we estimate the probability of a decision outcome (FP) for an average participant, with a given *vis* and *m* condition, marginalising over *trial*.

[5] An article on model interpretation using average marginal effects: https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf