Should We Dangle a Carrot? The Effect of Performance-based Incentives in Visualization Experiments

Abhraneel Sarma, Matthew Kay, Sheng Long, Michael Correll, Alexander Lex

Abstract—A perennial research question in visualization involves identifying which visual encodings for a particular dataset are most effective for users in performing a specific task. The relative effectiveness of the different encodings are commonly identified through controlled experiments. However, designing an experiment involves making many, often ad hoc, decisions about the experimental setup such as whether to include a training module, whether to provide performance-based incentives to participants, etc. Yet, there is limited guidance on how these decisions should be made, and we do not fully understand the impact of these subjective decisions on empirical results. In this paper, we investigate the impact of one such key design decision—monetary rewards. Specifically, we ask: does providing or not providing participants with performance-based financial incentives affect the results and the conclusions that we draw from visualization studies? We conducted two crowdsourced studies investigating the impact of incentives on (i) a low-level, perceptual task (perception of correlations in scatterplots or parallel coordinate plots), and (ii) a task involving reasoning (decision-making based on a weather forecast represented as intervals or density plots). In each of these studies, we manipulate both the visual representation and the presence of incentives as between-subject conditions. We expected to find no effect of incentives on the perceptual task, but to see an effect for the decision-making task. However, we found no effect on task performance in either study. While these are results of only two studies and should be replicated, they suggest that performance-based financial incentives may not always have the intended effect on participants that we presumed, and calls for a reflection of how incentivized studies should be designed. A copy of this paper and all supplemental materials are available at `https://osf.io/t8eah`.

Index Terms—Graphical Perception, Decision Making, Quantitative Methods, Experimental Design, Incentives.

1 Introduction

When designing a visualization, one of the goals of a designer is to communicate the data effectively to the viewer. This typically means representing the data using visual encodings which are easy for viewers to decode. To identify the relative effectiveness of visual encodings for a particular task, visualization researchers turn to empirical studies where they compare how well participants perform on a task using two or more visual representations. The encoding which allows the average participant to perform better on a task is then considered more effective for that task. However, in the process of designing an experiment researchers need to make many, often subjective or ad hoc decisions that may have unknown effects. Here, we draw attention to the potential issue of researcher degrees of freedom in the empirical studies that visualization researchers use for determining the effectiveness of various visual encodings, and how this flexibility can impact our theoretical understanding.

This work was partly motivated by our prior experience designing an experiment to study the effect of a new visualization on a decision-making task [56]. Initially, we provided participants with a description of how the data was encoded in the visualization and what the information meant; yet we observed a lot of variance in how participants interpreted the visualization and used the information to make decisions. We speculated that some of this variance could be attributable to people misunderstanding how to read the chart [44]. Subsequently, when we included more explicit information on how to read the chart during onboarding, we observed less heterogeneity in participants' responses. Since our goal was to evaluate the effectiveness of the visual representation, and not whether participants understood the representation correctly, we deployed the modified onboarding procedure.

For a typical experiment, researchers have to make several such experimental design decisions that can potentially have an impact the results. These include: how extensive the instructions should be, or whether participants should be trained on the task, given feedback, provided with performance-based monetary rewards, etc. However, these decisions are often made implicitly or even in an ad-hoc manner, and rarely articulated in a paper. We refer to these degrees of freedom in experimental design decisions as *tacit factors*. These decisions mirror the oft criticized undisclosed flexibility that researchers have in making decisions at various steps in the data analysis process [62, 63], but occur instead during the experimental design stage.

In this work, we take a first step of investigating one such factor: **the impact of performance-based financial incentives**—rewarding correct or optimal decisions by participants through monetary bonuses—in crowdsourced visualization experiments. While some recent visualization studies, primarily on decision-making under uncertainty, have incentivized participants [e.g., 12, 31, 48, 56, 60, 69], most empirical studies in visualization do not. However, **we do not know how or if the decision to incentivize participants' performance with money affects the results or the conclusions that we can draw from the studies.** This can make it tricky to compare and interpret results across studies which differingly incentivize participants but are investigating performance on the same task and using the same set of visualizations.[1] The theoretical argument for employing incentives, put forth primarily by economists, is that they induce more effort, which leads to improved performance on a task. However, we are not aware of any formal experiments on the effects of monetary incentives in crowdsourced visualization studies, a gap we attempt to close with this work.

We chose to study incentives over other *tacit factors* because we expected them to be fairly straight-forward to test in a controlled experiment, and to have a significant impact on participant performance. As visualization studies encompass a wide spectrum in terms of complexity, the types of cognitive processes that are involved, etc., incentives might impact different types of studies differently. We chose two prototypical visualization studies—perception of correlation [e.g., 8, 24, 52], which involves a lower-level perceptual task; and decision-making under uncertainty [e.g., 28, 37, 48, 56], which involves a higher-level reasoning task—to examine the effect of incentives on. We conducted two preregistered crowdsourced experiments where we varied incentives: participants were either paid a fixed amount independent of their performance, or were paid a smaller base amount, with the majority of their compensation tied to their performance.

In study 1 (perception of correlation), we partially replicated Harrison et al's experiment [24]. We visualized pairs of positively correlated datasets using either scatterplots or parallel-coordinates plot and asked participants to choose the plot with the greater correlation of the

- *Abhraneel Sarma and Alexander Lex are with Graz University of Technology.*
- *Matthew Kay and Sheng Long are with Northwestern University*
- *Michael Correll is with Northeastern University*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

[1] This tension is not merely theoretical—a real world example of this is the study by Oral et al. [46], which conducts a variation of the study by Kale et al. [31] but without performance-based incentives. However, the two studies used different dependent variables making the results not directly comparable.

two shown. We expected incentives to have no impact on performance (using just noticeable difference as the metric) in this scenario, as the difference between the plots either is or is not evident, and additional effort would likely not make a difference on task performance.

In study 2 (decision-making), we replicated the decision-making task from Joslyn et al. [28, 29, 42], using either an interval or a density plot to show the probability of temperatures falling below freezing, and tasked participants to make decision on whether to spend a budget to salt roads, or to preserve their budget and forgo salting. We expected incentives to have an effect on performance (using expected utility as the metric) in this scenario, as participants must make judgements about the optimal decision and estimate probabilities based on the visualized uncertainty distribution, where we expected additional effort to improve outcomes.

In both studies, our results showed no significant performance differences between the incentivized and non-incentivized conditions. While these are results of only two studies and therefore the findings need to be replicated before stronger conclusions are drawn, they seem to suggest that, at least in the context of crowdsourced studies on platforms such as Prolific, performance-based financial incentives may have fewer benefits than we presumed. Moreover, we do find some evidence that participants spend more time performing the task in the incentivized conditions, which raises some important ethical considerations for researchers regarding ensuring fair wages for crowdworkers. We discuss the implications of our results, their possible causes, and the scope in which our results may generalize.

2 Background

2.1 Researcher Degrees of Freedom in Scientific Studies

During the course of designing, collecting and analyzing data, and reporting the results of empirical studies, researchers have to choose between several reasonable alternative choices [61, 67]. In the absence of theoretically or empirically grounded guidance, they often make these choices arbitrarily. Also referred to as *researcher degrees of freedom*, this flexibility, especially in the data analysis stage of a study, has received considerable attention in recent years as it can considerably increase the probability of a false-positive finding [e.g., 11, 23, 57, 58, 62, 63]. Our focus in this work is instead on the degrees of freedom that exist in the design stage of an empirical study (*tacit factors*). These tacit factors make it difficult to "establish the generalizability of an effect across contexts" [36], as, at least in the case of visualization and HCI research, different study designs (e.g., whether or not participants are trained on how to read a chart) actually correspond to different contexts of usage [59].

2.2 The Effect of Financial Incentives

While the effect of incentives in experiments have been explored in prior work, primarily in the field of behavioral economics, the findings have been mixed. While some studies have found incentives to improve performance, most studies have found no effect or even a detrimental effect of incentives on performance [1, 2, 3]. Many of the studies which found that incentives did not improve performance also found that incentives reduced variance in responses. Based on their review of prior work investigating the effect of incentives, Camerer and Hogarth [3] propose the *capital-labor-production* theory to describe the potential effect of incentives in tasks. Here, *capital* refers to "cognitive capital" or knowledge that participants possess coming into the task. *labor* refers to effort that is required to perform the tasks. In most visualization and HCI studies, participants primarily exert cognitive effort. *Production* refers to the kinds of capital that is necessary for performing a task.

The *capital-labor-production* theory [3] rests on a few key intuitions: (i) people dislike exerting effort, but will put in more effort if that means increased rewards; (ii) effort generally, monotonically, improves performance; (iii) capital can be a substitute for labor (for example, if one is familiar with the biases in area perception implied by Steven's power law [64], they may be more accurate at comparing areas of two circles, but someone who does not know of these biases can be just as accurate by comparing the radii and computing the areas mathematically, which entails more effort); (iv) if a task's production requirements (i.e., the effort it requires) are too low or too high, there would be little marginal gain for the participant from the extra effort that they would exert due to the presence of incentives.

One important caveat is that these studies were often conducted in physical lab spaces (not on crowdsourcing platforms), and involved very different types of tasks than what we typically employ in visualization studies. In a study conducted with crowdworkers on Mechanical Turk, Mason and Watts [40] found that participants performed a greater number of tasks as pay increased, but the accuracy of the performed tasks did not improve with increases in pay. However, the tasks that participants were asked to perform in these experiments were either quite menial (ordering images temporally) or were word games, both of which are tasks where incentives are unlikely to help [3]. Other studies have investigated factors that can influence performance through an interacting effect with incentives such as intrinsic motivation [4]. As our goal was to study the *total effect of incentives*regardless of participants' intrinsic motivation, we do not control for motivation in our study design or analysis [41].

Based on the literature [3], we may expect the effect of incentives to manifest through improved performance on certain (but not all) tasks. Specifically, the tasks which are most likely to effected by incentives are those where most or all participants possess the required cognitive capital to perform the task, and the task does not require too little or too much effort. To investigate the effect of incentives, we chose two visualization tasks—(1) perception of correlation and (2) decision-making under uncertainty—which require varying degrees of cognitive capital and effort. In addition, we may expect reduced variance in participants' responses, and an increase in the amount of time participants spend on the task (which might be a proxy for increased effort).

2.3 Perception of Correlation

There has been a great deal of work in visualization and graphical perception on ranking different visual representations on their ability to help viewers accurately estimate correlation (and small differences in correlation). How accurately a viewer is able to perceive correlation using a specific representation is often determined by estimating just-noticeable differences (JNDs)—the minimum difference in stimuli at which a viewer can reliably (75% of the time) detect a difference.

Prior work has found that viewers can perceive correlation fairly accurately when the data is represented using scatterplots [24, 52], and the JNDs can be described using Weber's law. Harrison et al. [24] evaluate perceptions of correlation using other visualizations such as parallel coordinate plots, stacked bars, ordered lines etc. The work of Harrison et al. [24], and a re-analysis by Kay and Heer [33] show that participants' perceptions of *negative* correlation are almost as accurate when the data is represented using either scatterplots or parallel coordinates; however, participants' perception is much less precise for *positive* correlation when the data is represented using parallel coordinates (compared to scatterplots).

Perception of correlation is often considered to be a more basic or lower-level cognitive function [25]. Prior work on perceptions of correlation have exclusively paid participants a flat amount for participating in the study. Our hypothesis was that if incentives *were* to have a meaningful effect, we might observe that a visualization such as the parallel coordinates plot, which is less accurate compared to scatterplots for positive correlation in the absence of incentives, is as accurate as scatterplots when incentivized. However, based on the findings of Camerer and Hogarth [3], we speculate that this task is unlikely to be impacted by incentives as it might not benefit from exerting additional effort. Thus, this study represents an important edge case which we seek to evaluate. In our study, we compare scatterplot and parallel coordinates plot for positive correlations.

2.4 Decision-making under Uncertainty

A growing number of studies have investigated the effectiveness of different visual representations in accurately extracting probability information from uncertainty representations [e.g., 6, 27, 34], assessing risk [e.g., 18, 47, 55], or making utility-optimal decisions [e.g.,

12, 29, 31, 60, 69]. In this work, our focus is on the effect of visualizations in helping people make rational (i.e., decisions which are or close to utility-optimal). Prior work has found that uncertainty representations which are more expressive and present viewers with more complete distributional information such as density or violin plots, quantile dotplots, cumulative density plots lead to better decisions compared to interval plots which communicate a 95% (or similar) confidence or credible interval [12, 31, 60].

In these types of studies, researchers have often used financial incentives [e.g., 12, 31, 48, 60] based on the notion that incentivising participants will increase both the internal and ecological validity of a study [12, 30, 68]. When making decisions, people, both in the real world and in experiments, are likely trading off some *subjective costs* against some *subjective benefits*. By explicitly stating what these costs and benefits are to a participant, incentives, in theory, provide benchmarks for defining optimal or good decisions, as well as enable the evaluation of participants' decisions against the benchmark [60, 68]. As incentives reflect the costs and benefits associated with real-world decision-making, it also should make the findings more ecologically valid [12], assuming that the incentives in the experiment align well with the relative costs of real-world tasks.

These types of decision-making tasks require participants to perform relatively more complex (or higher-order) reasoning tasks, which require more effort. In our study, we examine the impact of incentives by comparing how participants perform in a decision-making task where the uncertainty distribution is represented using either an interval representation—specifically 66% and 95% intervals—which are typically considered sub-optimal, or density plots, which are have been found to be better [12, 31]. Based on the *capital-labor-production* theory [3], we expect incentives to impact the results, as participants who exert more effort would be more likely to spend time and more accurately estimate the probability values from the uncertainty representation which is necessary to perform the task. Specifically, in the presence of incentives, our hypothesis was that participants will perform meaningfully better when uncertainty information is presented using density plots compared to interval plots, while in the absence of incentives, we might expect this improvement in performance to be much smaller or even disappear.

3 Experiment Design Preliminaries

We conducted preregistered experiments on the perception of correlation and decision-making tasks to study the effect of incentives. The study procedure for both studies was approved by the ethics board at Graz University of Technology (GZ EK-109/2026).

For both studies, our goal was to ensure that participants' average compensation would be approximately the same ($15/h) across both the baseline and incentivized conditions. We first calculated the average number of correct responses for both experiments based on our pilot studies. On average, participants correctly performed the task in study 1 in 40/65 trials. For study 2, prior work which used the same experimental setup [56] showed that the average participant had 4000 virtual dollars left at the end of the study. We then assigned dollar values for answering a question correctly in study 1 ($0.05) and for the amount of virtual dollars remaining in the bank in study 2 ($0.5 for every $1,000). Thus, we expected the average bonus that participants in the incentivized conditions would receive to be approximately $2 for both studies. The maximum bonus they could theoretically receive was $3.25 for experiments 1 and $9 for experiment 2. We anticipated both tasks to take approximately 9-12 minutes on average, which, at the desired compensation rate of $15/h, translates to a total compensation of $3. Thus, the compensation for the baseline (non-incentivised) conditions was set to $3 for both studies; the compensation for the incentivized conditions were $1.25 and $1.5 for Experiments 1 and 2 respectively. The slightly higher *estimated* average compensation in the incentivized conditions were set in order to comply with Prolific's minimum wage of $8/h, which does not include bonuses.

While these amounts were the compensation advertised to participants *before* taking the study, as participants took significantly longer in the incentivized conditions, we adjusted the guaranteed compensation amount to both comply with Prolific's standards and to match our desired target compensation of $15/h. The exact compensation amounts are reported separately for each study.

4 Experiment 1: Perception of Correlation

We conducted a preregistered study to partially replicate the perception of correlation study by Harrison et al. [24] using either scatterplots or parallel coordinates plot (see `preregistration`). As this task involves more lower-level cognitive functions [25], we test this as a possible edge-case of a visualization task where incentives will likely have little to no effect on performance.

4.1 Experimental Materials

Figure 1 shows a screenshot of the experimental interface. The study is available to browse `here` and the source code for the study is available on `GitHub`.

Experimental Task and Conditions: We adapted the perception of correlation task from Harrison et al. [24] for this experiment. Like Harrison et al. [24], in each trial, participants were shown two charts and were asked to select the one with the higher correlation in a Two-Alternative Forced Choice (2AFC) task. There were four experimental variables in this study: (1) the visualization (scatterplot or parallel coordinates plot); (2) the incentive scheme (advertised[2] as a participation fee of $3 or a participation fee of $1.25 and a bonus of $0.05 for each correct response); (3) the base correlation (r_{BASE}); and (4) the difference in correlation between the two charts (Δr). The visualization and incentive scheme were varied between-subjects whereas r_{BASE} and Δr were varied within-subjects. To keep the experiment simple, we only considered positive correlations, and the *approach from above*—in other words the correlation of the second chart was always greater than the correlation of the first chart and is given by $r_2 = r_{\text{BASE}} + \Delta r$ (the order in which the charts appeared was randomized).

Procedure: We test five levels of $r_{\text{BASE}} = \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and thirteen levels of $\Delta r = \{0.03, 0.04, \ldots, 0.1, 0.12, 0.14, 0.18, \ldots, 0.26\}$ in a fully crossed design resulting in 65 trials. In addition, we included five attention check questions where the correlations for the two charts are 0.01 and 0.99. The order of the trials were completely randomized. Unlike previous studies [8, 24, 52], we did not employ a staircase procedure due to the possibility of a confounding effect with the incentives—we felt that there might be a possibility of two participants ending up with the same payout in the incentivized condition even if they have different JNDs due to the mechanisms of the staircase procedure. Unlike prior studies [8, 24] we provided immediate feedback to participants on whether they answered correctly, as seen in Figure 1. At the end of the experiment, we informed participants

[2]For both this and the subsequent experiment, we had to adjust the guaranteed amounts in the incentivized conditions to ensure an hourly wage of approximately $15/h, as reported in §3. The actual amounts are reported below.

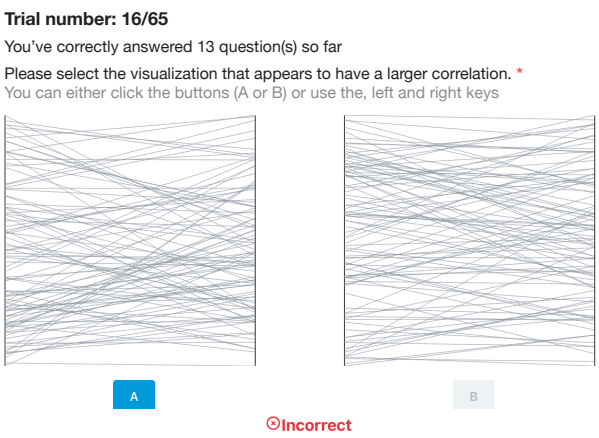

Fig. 1: Example of a stimulus seen by a participant in the non-incentivized parallel coordinates plot condition.

about their total payout. After completing all trials, participants were asked to describe the strategy that they used in performing the task, and about their prior experience with incentivized studies.

Tutorial and Training: Before participants began the study, we provided them with instructions on what correlations are with examples of what different levels of correlations look like, using the visualization that they will encounter subsequently. This is followed by participants completing nine training trials with feedback, using the same set of correlation stimuli that was used by Cutler et al [8].

Participants: We recruited all participants from Prolific. Our experiment was only eligible to participants who were fluent in English, and on desktop devices. We aimed to recruit 200 participants (50 participants in each condition). We excluded five participants who failed to meet our criteria, and recruited additional participants to meet our target. We received explicit consent from all participants to collect and share their responses. The median completion time for participants in the baseline condition was approximately 11.5 mins, and they were compensated $3 (approximately $15.75/h); the median completion time for participants in the incentivized condition was approximately 13.5 mins, and they were received a guaranteed amount of $1.6 (adjusted up from $1.25) and the average bonus was $2.18 (corresponding to an average wage of $16.8/h).

4.2 Model specification

We adapted the approach used by Kale et al. [31], and used a Binomial logistic regression model to estimate the probability of a correct response. The logistic fits (`line 2`) describing each participants' responses are equivalent to psychometric functions [21], which can then be used to derive an estimate of JND (which we discuss further below).

`line 1` $correct \sim \text{Binomial}(p)$

effect of... vis, incentives, r_{BASE}

`line 2` $$\text{logit}(p) = \alpha_i + \beta_{\text{VIS}[i]} \cdot \beta_{\text{INC}[i]} \cdot \gamma_{\text{R}} r_{\text{BASE}[j]} + \beta_{\text{VIS}[i]} \cdot \beta_{\text{INC}[i]} \cdot \gamma_{\text{D}} \Delta r_{[k]}$$

Δr

`line 3` $\alpha_i = \alpha_0 + \delta_{\alpha,i}$

`line 4` $\delta_{\alpha,i} \sim \text{Normal}(0, \sigma)$

`line 5` $i \in \{1...N\}$ (N *participants*)

`line 6` $j \in \{1...J\}$ (J levels of r_{BASE})

`line 7` $k \in \{1...K\}$ (K levels of $\Delta = |r_{\text{BASE}} - r_2|$)

Line 1: Whether a participant correctly chose the graph with the higher correlation in each trial is modeled as a Binomial distribution with probability p, where 1 means the participant made the correct choice.

Line 2: For each stimuli (which consists of two graphs), we assume that participants' decisions would depend on the visualization, the incentive condition, the base correlation (r_{BASE}), and the differential increase in stimuli i.e. correlation (Δr). This model incorporates the (directionless) assumption that the probability of a correct response (p) will change as r_{BASE} changes, and the magnitude of this change in p will also vary with Δr. Based on prior work, both of these effects are likely to be in the positive direction.

Lines 3-4: We expect the intercept (α_i) parameter to vary between participants, as different participants will likely have different capital and perceptive abilities. α is the slope parameters for the *average* participant (when $\delta_{\alpha,i} = 0$); $\delta_{\alpha,i}$ captures differences between participants as random effects.

Priors: We use weakly-regularizing priors centered on zero (no effect) while permitting the possibility of significant distortion: $\alpha \sim \text{Normal}(0, 1)$. We use zero-centered priors with a standard deviation of 2 for the random effects parameters ($\delta_{\alpha,i}$).

Implementation: We implemented these models in `R 4.4.0` [66] and `CmdStanR 0.8.0` [17]. The model ran for four chains with 5,000 warmup samples and 5,000 post-warmup samples each, thinned by 4 for a final total sample size of 5,000. 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 ($\text{ESS}_{min} \approx 2,000$).

JND Calculation: Based on the estimates of this model, we can calculate the JND at a specific level of r_{BASE} for a particular visualization and incentive condition as:

Constant w.r.t. $\Delta r = A$

`line 1` Let, $$F(\Delta r | r_{\text{BASE}}) = \alpha_i + \beta_{\text{VIS}[i]} \cdot \beta_{\text{INC}[i]} \cdot \gamma_{\text{R}} r_{\text{BASE}[j]} + \beta_{\text{VIS}[i]} \cdot \beta_{\text{INC}[i]} \cdot \gamma_{\text{D}} \Delta r_{[k]}$$

changes with $\Delta r = B$ (Slope)

JND is defined as the increase in stimuli needed for the probability of correctly detecting correlation to increase from 50% to 75%:

`line 2` $$\text{JND} = F^{-1}(0.75) - F^{-1}(0.5)$$ ⟶ 0 since this is a 2AFC task

`line 3` $$= F^{-1}(0.75)$$

`line 4` $$= \frac{\text{logit}(0.75) - A}{B}$$

4.3 Results

The results for Experiment 1 are shown in Figure 2 as estimates of JND (just noticeable differences). In the baseline (non-incentivized) condition, we find that the JND for scatterplots is much lower at all levels of the baseline correlation, replicating the finding from prior work [24, 33].

As expected, we also **do not find any evidence to suggest that incentives impact performance in the perception of correlation task**—the JNDs for participants in the incentivized conditions were comparable, if not slightly worse, than the JNDs for participants in the baseline condition (Figure 2A). We also do not find any evidence to suggest that incentivizing participants led to reduced variance in their responses, as measured through the group-level standard deviation parameters in our multi-level logistic regression model (see `supplement` ▶ R ▶ `dm-05-effect_variance.qmd`). However, we did find that participants took slightly longer to complete the task in the incentivized conditions compared to the baseline conditions. Figure 3A, visualizes the bootstrapped median and 95% quantile intervals for the median estimate. The difference in time spent is approximately 30s for parallel coordinates (base parallel coordinates: 288s, [264s, 323s], incentivised parallel coordinates: 318s, [294s, 391s]), and approximately

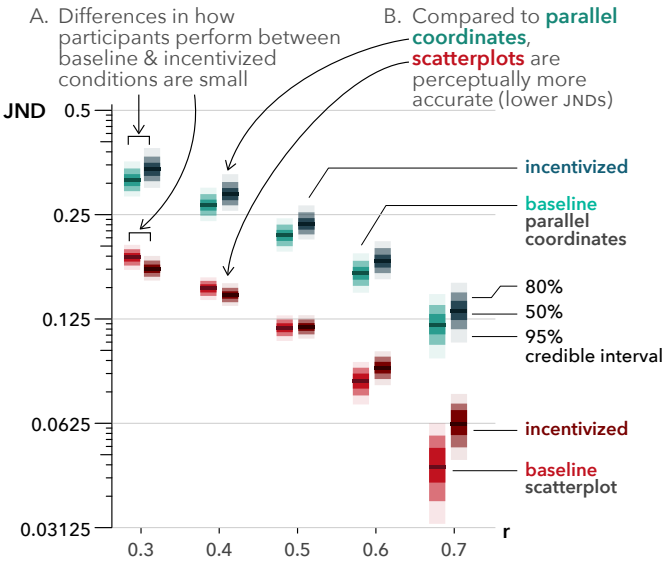

Fig. 2: The main result of Experiment 1. We show the posterior estimates of JND for both visual representations across the incentivized and baseline conditions, at each value of r_{BASE}.

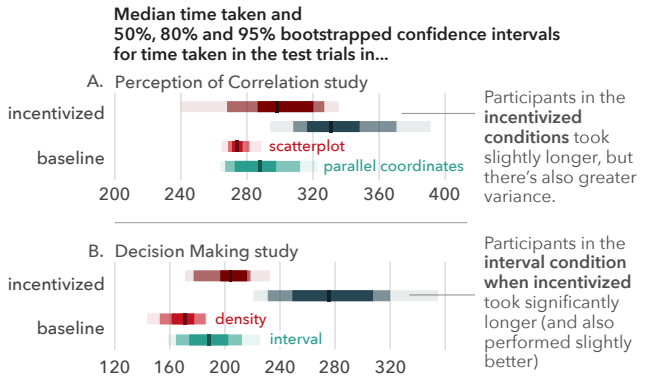

Fig. 3: Median time taken and 50%, 80%, and 95% bootstrapped confidence intervals for time taken in the test trials for the two studies.

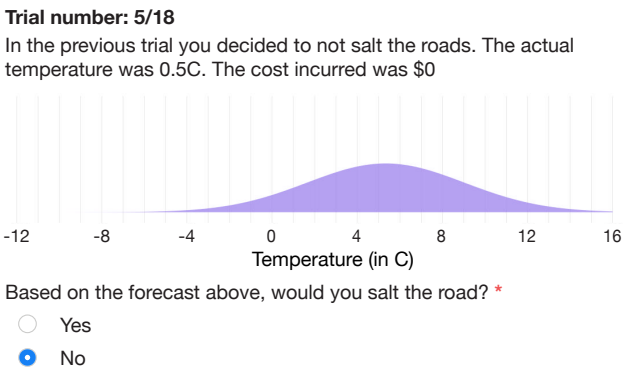

Fig. 4: Example of a stimulus seen by a participant in the incentivized density plot condition.

17s for scatterplot (base scatter: 275s, [265s, 289s], incentivised scatter: 292s, [240s, 336s]). The increase, which is approximately 6-10%, should likely be considered small.

Finally, as an aside, we want to highlight the ability of our experimental setup to replicate JND estimates from prior work *without using a staircase procedure*. An additional advantage of this setup is that it allows us to model participants' decisions directly using the Binomial logistic regression model described above.

5 Experiment 2: Decision-making under Uncertainty

In this preregistered study, we examined the impact of incentives on performance in a decision-making under uncertainty task adapted from prior work [29] using either interval or density plots as uncertainty representations (see `preregistration`). As performing this task involves relatively more complex cognitive functions, our hypothesis was that incentives would have an effect on performance.

5.1 Experimental Materials

Figure 4 shows a screenshot of the experimental interface. The study is available to browse `here` and the source code for the study is available on `GitHub`.

Experimental Task: For this study, we adapted the scenario that was used by Joslyn and LeClerc [29], where participants were presented with the following hypothetical scenario:

> Assume that you are an analyst of a road maintenance company contracted to treat the roads with salt brine to prevent icing in a U.S. town impacted by severe cold. Applying salt brine to the roads is costly for your company and is also detrimental to the environment, as it can pollute groundwater and kill roadside vegetation. However, not salting the roads can cause significant accidents during freezing temperatures, the costs of which are borne by your company. Your job is to salt the roads when temperatures drop below 0°C (32°F). You have a budget of $18,000 for 18 days. Salting all the roads in the town costs $1,000 (per night). If you fail to salt the roads and the temperature drops below 0°C (32°F), it will cost $5,000 from your budget. You will be shown a night-time temperature forecast distribution based on which, you will have to decide whether to salt the roads.

The payoff matrix for the decision problem described above can be represented using the following table:

	$s_1 : \mathrm{T} \leq 0°\mathrm{C}$	$s_2 : \mathrm{T} > 0°\mathrm{C}$
a_1 : salt	-1000	-1000
a_2 : ¬salt	-5000	0

Based on this incentive scheme, a decision-maker who wants to maximize expected utility should prefer the action a_1 (to salt) if and only if the probability of temperature being below freezing, $\Pr(\mathrm{T} \leq 0) = p \geq 0.2$. Thus, $p = 0.2$ is the optimal *crossover point*: the probability at which a decision maker should not have a preference between actions.

Experimental Conditions: There were three experimental variables in this study: (1) the visualization (interval or density plot); (2) the incentive scheme (advertised as either a participation fee of $3 or a participation fee of $1.5 and a bonus of $0.5 for each $1,000 remaining in their account at the end of the experiment); (3) the actual forecasted probability of freezing (9 levels, each repeated twice). The experiment used a mixed-factorial design with visualization and incentive scheme varied between-subjects and the forecast varied within subjects.

Procedure: The experiment consisted of 18 trials and two attention check questions which are interspersed among the 18 trials. In each trial participants are presented with a single forecast in the form of a Normal distribution (see Figure 4). We varied the means and standard deviations of the Normal distribution to generate different forecasts with different probabilities of freezing: $\Pr(T \leq 0°\mathrm{C}) = \{0.595, 0.5, 0.405, 0.315, 0.235, 0.168, 0.115, 0.075, 0.046\}$. The attention check questions showed extreme forecasts ($Normal(-11, 0.4)$ and $Normal(11, 0.4)$). The order of the trials were completely randomized. After completing all the trials, participants were asked open-ended questions similar to experiment 1.

Tutorials and Training: We provided instructions to participants on how to correctly interpret the interval and density plots. This is followed by participants completing four training trials where participants were shown four different temperature forecasts and were asked to report the probability of the temperature falling below a certain marked temperature. Participants were then provided feedback on whether they answered the questions correctly.

Participants: We recruited all participants from Prolific. Our experiment was only eligible to participants who were fluent in English, and on desktop devices. As per our pre-registrations, we aimed to recruit 280 participants (70 participants in each condition). We experienced some data-logging issues which meant the data for 31 participants were not recorded. After excluding participants who failed to meet our pre-registered attention check criteria (six), we had 243 participants (60 in the base interval, 62 in incentivised interval, 59 in base density, and 62 in incentivised density). We received explicit consent from all participants to collect and share their responses. The median completion time for participants in the baseline condition was approximately 16 mins, and they were compensated $3.75 (approximately $14/h); the median completion time for participants in the incentivized condition was approximately 24 mins, and they received a guaranteed amount of $4.2 (adjusted up from $1.5) and the average bonus was $1.6 (corresponding to an average wage of $14.5/h).

5.2 Model Specification

We used a linear-in-log-odds (llo) model [70], which is a good fit to account for the "distortion of judgement or misperception of probabilities" [69] in such decision-making under uncertainty tasks. This model, which has been used in prior work to model the same task [56], is derived directly from the utility-optimal decision criterion by first translating it into log-odds and then applying a linear transformation to derive a model for participants' decisions:

$$\texttt{line 1} \quad salt \sim \text{Binomial}(p_{\text{SALT}})$$

$$\texttt{line 2} \quad \text{logit}(p_{\text{SALT}}) = \alpha_i + \beta_i \cdot [\text{logit}(p) - \text{logit}(0.2)]$$

(intercept, slope, probability of freezing)

$$\texttt{line 3} \quad \alpha_i = \alpha_0 + \alpha_{\text{VIS}[i]} + \alpha_{\text{INC}[i]} + \delta_{\alpha,i}$$

$$\texttt{line 4} \quad \beta_i = \beta_0 + \beta_{\text{VIS}[i]} + \beta_{\text{INC}[i]} + \delta_{\beta,i}$$

$$\texttt{line 5} \quad \begin{bmatrix}\delta_{\alpha,i}\\ \delta_{\beta,i}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma\right)$$

$$\texttt{line 6} \quad i \in \{1 \ldots N\} \qquad (N \textit{ participants})$$

Line 1: The binary decision (where 1 represents the decision to salt and 0 represents the decision to not salt) made by a participant in each trial is modeled as a Binomial distribution with probability p_{SALT}.

Line 2: We assumed that participants' decisions would be a function of the temperature forecast distribution shown, which can be represented using the corresponding probability of freezing value (p). The intercept (α) parameter controls the fixed point of the function—how people map the optimal crossover point ($p = 0.2$) to the probability of salting—which shifts the crossover point. The further α is from 0, the more bias there is, with negative values of α suggesting that crossover point for the average participant is greater than the optimal, and positive values of α suggesting that crossover point for the average participant is less than the optimal. The slope (β) parameter controls the degree of sensitivity. The further it is from 1, the more sensitive a participants is to their subjective crossover points.

Lines 3-5: We expect the intercept (α_i) and the slope (β_i) parameters to vary between participants, as different participants will likely have different decision-making capacities. α and β are the slope parameters for the *average* participant ($\delta_{\alpha,i} = 0$; $\delta_{\beta,i} = 0$), whereas $\delta_{\alpha,i}$ and $\delta_{\beta,i}$ capture differences between each participants' intercepts and slopes compared to the average participant, as random effects.

Priors: Perfectly unbiased responses would yield values of $\alpha_i = 0$. $\beta_i = 1$. We thus use priors centered on these values but which also permit the possibility of significant distortion: $\alpha \sim \text{Normal}(0, 1)$ and $\beta \sim \text{Normal}(1, 1)$. We use zero-centered priors for the random effects parameters ($\delta_{\alpha,i}$ and $\delta_{\beta,i}$).

Implementation: These steps were almost identical to study 1.

5.3 Results

The results for Experiment 2 are shown in Figure 5. We found the value of α to be negative across all conditions (Figure 5A). This suggests that the crossover point for the average participant is greater than 0.2, exhibiting a consistent bias towards risk-seeking behavior. The estimated value of α was smaller for incentivised density compared to the other three conditions. We also found the value of β to be significantly larger than one across all conditions (Figure 5B), indicating that participants are quite sensitive to the stimuli and their subjective crossover point. The similarity in the estimates of the α and β parameters across all conditions suggest that the average participant in every condition likely exhibited very similar decision-making behavior.

We can directly compare decision quality by estimating the expected utility that the average participant in each of these conditions would achieve, which is a measure of how participants performed based on the information that was provided to them regarding how their performance will be evaluated; we report this measure in Figure 5C. Contrary to our apriori expectations, we **find no evidence to suggest that the average participant performs better in the incentivized conditions** in terms of their predicted expected utility (incentivised interval: 2.76, [2.31, 3.12] and incentivised density: 2.33, [1.65, 2.86]) compared to the baseline conditions (base interval: 2.69 [2.19, 3.04] and base density: 2.35 [1.70, 2.86]).[3]

[3] All of these expected utility estimates are in thousands of dollars

We also **found that the average participant in the interval condition performed slightly better than the average participant in the density condition**, regardless of whether they are incentivized or not. This difference is relatively small—the difference in expected utility is 0.66 95% CI: [0.25, 1.1] when incentivised and 0.44 95%CI: [0, 0.87] when not incentivized (see Figure 5C). Figure 6 shows the distribution of utility attained by participants in the various conditions, as well as the posterior predictive intervals (see also Appendix A for more details). The *actual* average utility of participants in the interval condition under incentives was much larger than the other conditions; we suspect that this might have been an artifact of the RNG in JavaScript—we implemented a series of checks but could not determine any systemic issues with the RNG (`supplement ▶ R ▶ dm-05-rng_check.qmd`). Our finding—that intervals perform better than densities—is in contrast to prior work [12, 31], which have typically found that participants make better decisions when information is presented to them using density plots compared to interval plots.We discuss the implications of this finding in §7.2.

As in the first study, we found that participants took slightly longer to complete the task in the incentivized conditions compared to the baseline conditions. Figure 3B, visualizes the bootstrapped median and 95% quantile intervals for the median estimate. The difference in time spent is approximately 80s for intervals (base interval: 187s, [159s, 226s], incentivised interval: 266s, [221s, 355s]), and approximately 32s for density (base density: 173s, [143s, 187s], incentivised density: 205s, [171s, 233s]). This increase of 20-40% is quite large, and might suggest that in tasks such as decision-making under uncertainty the use of incentives may lead to more time spent on the task, and *perhaps* more effort induced (depending on how one views the relationship between time spent on an online experiment and effort).

6 Crowdworkers' Perspectives on Incentives

For both studies, we elicited qualitative responses from participants regarding (i) their prior experiences with performance-based incentives,

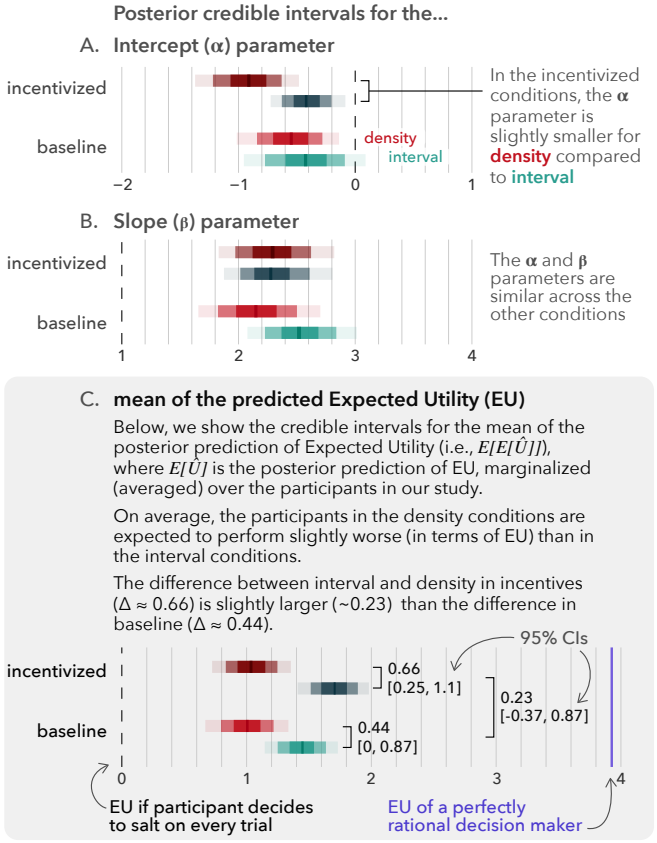

Fig. 5: The main result of Experiment 2. We show the posterior credible intervals of α, β and the mean expected utility for both visual representations across the incentivized and baseline conditions.

and (ii) how they might have behaved if they were in the other condition (i.e., asked people in the incentivized condition how they would behave if they were given a flat amount and vice-versa). We conducted an exploratory qualitative analysis of participants responses using an inductive coding approach to identify a set of codes, which was then used to code all the responses. Below, we report the numbers combined across both studies (N = 444).

Of the responses which we were able to code (224), significantly more participants (115) in our sample claimed to have positive attitudes towards incentives based on their past experience (e.g., "[...] they can make the task more engaging and encourage more careful decision-making" or "Surprised and happy. Its like a little gift you receive because you're doing a good job"), compared to those who explicitly reported mixed or negative feelings (28) towards incentives (e.g., "Very mixed because most times there is little actual chance of receiving any bonus" or "They are often misleading, or sometimes held as a raffle, which is a scammy practice"). Among those who reported positive feelings, some say that it encourages them to do well (34) or makes the task more interesting and engaging (19). On the other hand, those who claim to have negative experience usually believe that it is because they are unfair or conflate bonuses with lotteries or raffles.

For our analysis on how participants self-reported they would behave in the other incentivized condition, we focused only those who were in the incentivized condition as it would be easier for them to imagine the counterfactual. Of the responses we were able to code (192/224), a majority (119) claimed that would have behaved in the same way, there were several participants (35) who believed that the in the absence of incentives they might perform worse because they might be less motivated or focused (e.g., "If I had been guaranteed [...], my performance probably would have been slightly worse because the bonus made me pay closer attention to the tradeoff between salting costs and freezing risk."). In the decision-making study, several (12) reported that they would be more risk seeking (e.g., "[...] would have been slightly less cautious and less incentive-driven."). For these participants, incentives appear to be having the intended effect.

However, there were also a number of participants (13) who believed that they would have performed better in the non-incentivized condition perhaps because it would be less stressful (e.g., "[...] guaranteed payment offers greater financial security and could reduce performance anxiety." and "I think it would be better, as [...] one would treat the study properly because he would get a proper compensation in return"). This might suggest that, because of how incentivized experiments are designed with a lower base pay, for a number of participants, incentives might be having the opposite of the intended effect. In the decision-making study, several participants (16) stated that they might have been less risk averse as they would not be concerned with saving money (e.g., "With a guaranteed reward, the incentive to save money would vanish. My performance would likely have become highly risk-averse"), again suggesting that incentives may not be having the intended effect for many participants.

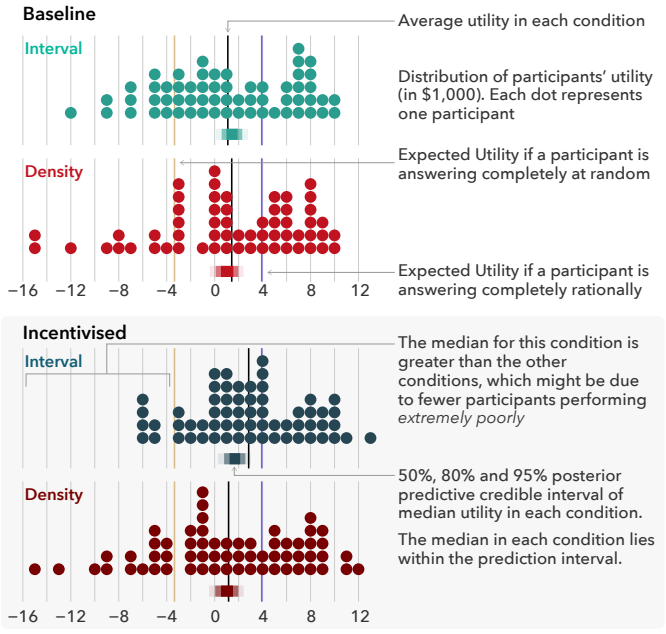

Fig. 6: The distribution of utility obtained by the participants and the *predicted* expected utilities on the exact data shown to participants, but with the simulated temperature calculated using the RNG in R. The observed median utility for all conditions lie within the prediction intervals.

7 Discussion

7.1 What is the Effect of Incentives?

Our work was motivated by several recent studies which have employed incentivized experiments to study decision-making under uncertainty [e.g., 12, 31, 48, 60]. The rationale was that decision-making in the real-world entails consequences and incentives would bring that notion of realism to virtual survey world. Based on the *capital-labor-production* theory, our a priori hypotheses were that: (i) the effect of incentives would likely be small, if any, in the perception of correlation as the effort required (or production requirement) is low; and (ii) performance-based incentives would likely lead to a meaningful improvement in performance (in terms of effect size) in the decision-making task as more effort is required to properly perform the task (i.e., higher production requirements). The results of our two experiments (see Figures 2 and 5) suggests that, compared to a baseline of providing people a flat rate for participation, introducing performance-based financial incentives may not lead to improved performance in either the perception of correlation task or the decision-making task.

While expected for the perception of correlation task, this unexpected violation of our apriori expectations could potentially be explained by the capital-labor-production theory—it is possible that the the decision-making task may have been too difficult. We observed that participants in the incentivized conditions did spend longer in completing the task (see Figure 3B), and this difference in time spent was particularly large for the decision-making task. *If* time spent on a task is considered a reasonable proxy for effort exerted by a participant, this result suggests that the average participant in the incentivized condition is likely putting in more effort then the average participant in the non-incentivized condition, even though they are not necessarily performing better on the task. We discuss potential implications for studying decision quality in §7.3.

Prior work [3] also suggests that incentives often reduce variance in responses, even if they do not have an impact on mean performance. Our model allows us to estimate this variance, and while the variance parameters estimated by our model, for both studies, are smaller for the incentivized condition compared to non-incentivized conditions, these differences appear to be very small (see the "Effect on Variance" subsection in `supplement` ▶ R ▶ `dm-04-analysis.qmd`).

7.2 Why Were Our Results Inconsistent With Prior Uncertainty Visualization Studies?

Perhaps the most surprising aspect of our experimental results was that the average participant making decisions using interval plots was expected to perform marginally better than with density plots. This is contrast to prior work [e.g., 12, 31, 60] which has typically found that, in incentivized experiments, participants make better decisions when the uncertainty distribution is represented using densities rather than 95% interval plots, as densities provide complete distributional information. We expected the same result to hold for our study, even though we varied our design of interval plots.

We initially suspected that it was because, unlike previous experiments, we presented participants with both the 66% and 95% intervals. The two intervals provide additional information relative to prior studies, and supports a visual heuristic to make utility optimal decisions—if the 66% interval is slightly overlapping with the 0°C line, this indicates an approximately 20% probability that the temperature is going to be below freezing Figure 7. In a "just 95% interval" condition, as in prior work, there is no such easy visual heuristic that a participant can rely on. In contrast, a participant in the density condition would have

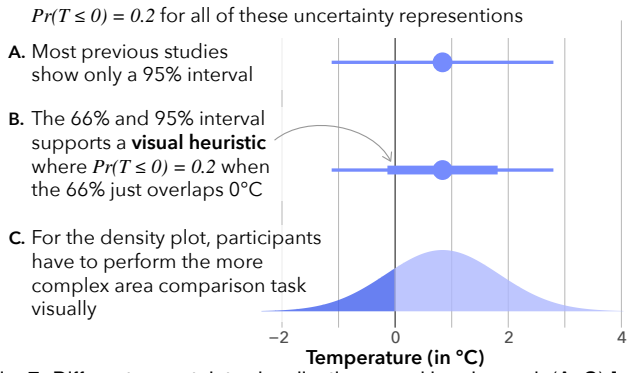

Fig. 7: Different uncertainty visualizations used in prior work (A, C) [e.g., 12, 31, 60], and in our work (B, C).

to compare areas to make the utility optimal decision, which might arguably have been more difficult, given how poor people are at area perception [64]. However, a follow-up study (see Appendix B) comparing 95% intervals with density plots again showed no difference in participants' decision quality between the two conditions. One possible cause could be the probability distributions used as stimuli in our study—the distributions were of varying standard deviations, which can interfere with the visual heuristics that participants use to estimate probabilities from densities [15, 16]. Our results paint a conflicting picture regarding the effectiveness of density plots for uncertainty communication, and warrants further research.

7.3 Ecological Validity of Decision-Making Studies

Given the results of our study, and how they were contrary to our initial hypotheses, we did consider shelving this paper. However, we realized that this would be a terrible idea for two reasons: (i) we would be contributing to the file drawer problem [14, 54] (self-selecting to not publish studies with negative results), and (ii) even though our study does not find evidence for the effect of incentives, it does raise some interesting questions about the use of incentivized experiments for studying decision quality and the ecological validity of studies run on crowd-sourcing platforms such as Prolific.

Why Do We Conduct Incentivized Decision-Making Studies?

Early work in uncertainty visualization compared different representations based on subjective measures such as preference or intuitiveness (e.g., which representation do you prefer?) [10, 22, 39], ease of use [53], and subjective probability (e.g., on a scale of 1-7, how likely do you think the event is going to occur?) or confidence (e.g., on a scale of 1-7, how confident are you about your prediction?) [6, 13, 22, 34]. A critique of these types of studies is the lack of an objective benchmark: a user might perceive visualization A as being more intuitive than visualization B, or they might experience a greater or lesser degree of subjective uncertainty when viewing it. But how do we know that they are perceiving the *correct* level of uncertainty? Is perceiving less uncertainty or more uncertainty better?

Other studies have investigated participants' ability to accurately estimate the probability of an event occuring using uncertainty visualizations (which sometimes reduces to visually estimating tail probabilities) [e.g., 7, 27, 34]. These studies include an objective benchmark—accuracy of probability estimation—which could be used to evaluate participants' performance. However, even if people are able to read and report numerical probabilities accurately, it is possible that their actual behavior diverges from what would be expected based on the reported probabilities, either due to bias or error, or because they associate different, subjective probabilities to uncertainty [38]. As such, measuring uncertainty through accuracy in reading numerical probability values may not necessarily tell us how good people are at using those probabilities to make good decisions.

Since uncertainty visualizations are typically used in the real world to make decisions and a person in the real-world is incentivized to make a correct decision because it has real consequences (e.g., we consult the weather forecast not because we care about knowing the probability of rain, but because we need to decide whether we should carry an umbrella or not), studying decision quality using uncertainty visualizations represented, arguably, a more ecological valid approach of evaluating different uncertainty communication approaches. While other metrics for measuring the effect of uncertainty visualizations could tell us that two visualizations are *different*, they do not necessarily tell us which one is *better*. Asking people to perform incentivized decision tasks, *in theory*, allowed us to measure decision quality against an objective baseline—*the expected utility of a rational decision-maker*—by comparing how far are people from that baseline.

Yet there are a couple of assumptions implicit in this line of reasoning: that all participants in incentivized decision-making tasks are attempting to maximize expected utility; and (assuming they are trying to maximize expected utility) people are able to translate the provided incentive structure or cost function into a decision rule based on probabilities (which is how the uncertainty information is typically presented to them). If these two assumptions hold, incentivized decision-making style experiments can be considered more ecologically valid than other approaches for evaluating uncertainty visualizations. Under these conditions, a decision-making study of uncertainty visualizations measures how people make decisions based on their subjective representation of uncertainty or subjective risk tolerance threshold, and an uncertainty visualization can be considered more effective if they allow a user to make decisions which are close to utility optimal. However, the qualitative analysis of participants responses to the open-ended questions in our study, and some introspective reflection, make us question these assumptions.

Are Participants Trying to Maximize Expected Utility?

In their qualitative responses, a few participants stated to have adopted a risk-seeking strategy, as indicated by statements such as "gambling with [a] small amount is fun and there is no real concern of losing money" or "there is less personal 'skin in the game,' so I might have been more willing to gamble on borderline forecasts." We also observed something similar in prior work [60], where some participants stated that they started taking risks after their initial decision-making strategy did not appear to have been paying off. While a few errant participants are to be expected, and can be easily accounted for in a mixed-effects regression model like the one we used, it is unclear how prevalent these attitudes are. In addition, our qualitative analysis suggests that (§6), due to the competing goals of crowdworkers on Prolific, incentives can have the opposite of the desired effect—as the guaranteed payment was low, instead of trying to maximize expected utility, participants simply behaved as if they were being paid the small base pay. If a significant proportion of participants are not trying to maximize expected utility and instead adopting a risk-seeking attitude when performing decision-making tasks, we might **need to evaluate how incentive structures in experiments are designed so that they elicit the behavior we *normatively* want**.

Can Participants Maximize Expected Utility?

Even if participants were trying to maximize expected utility, it is possible that they were not able to translate this to a decision rule with the optimal crossover point ($p = 0.2$). Of the participants for whom we could deduce a crossover point (53), 25 participants stated that they decided to salt the roads when the chances of freezing were around 30% or higher (the crossover point for the remainder were at or around 20%), suggesting an imperfect translation from the given cost function to a decision rule.[4] This is reflected in the model estimates of α (Figure 5A), and is consistent with findings from prior work on a similar task [56]. There is a qualitative difference between risk-seeking behavior when a participant decides to salt when $p = 0.3$ knowing that the optimal crossover point is $p = 0.2$, and "risk-seeking behavior" when a participant decides to salt when $p = 0.3$ believing that the optimal crossover point is $p = 0.3$. When evaluating decision quality, we want to make sure we are measuring the former and not the latter.

[4] We also observed a very colorful (mis)interpretation of probability, not too dissimilar to the misinterpretations of what *probability of rain* means as described by Gigerenzer et al. [20]: "If more than 50% of the night, or a significant proportion, was below zero, I salted the roads. If there were more hours above 0, I didn't salt because it wasn't necessary"

A different (mis)calculation can be observed among participants who stated that they would be more risk averse in the baseline conditions as they would not be incentivized to "save money." If some participants did not know how to maximize expected utility, it might also mean that they are not sensitive to the incentive structure that is provided to them (i.e., could participants behave similarly if the penalty for not salting was $3,000 instead of $5,000?). From the current study, it is unclear whether some participants' inability to identify a decision rule is simply an artifact of this specific task, the format of presenting incentive structures, difficulties in interpreting single event probabilities properly [20], not being sufficiently sensitive to incentives or something else entirely, and therefore warrants further research.

Other Possibly Important Factors To Consider.
In Experiment 2, we provided feedback to participants after every trial regarding what the simulated temperature was, and the costs incurred. It is possible that this immediate feedback could have led participants to adopt a sub-optimal strategy—according to Achtziger et al. [1], "the human tendency to repeat successful actions and avoid those which led to failure can impair performance by focusing attention on win/lose outcomes and away from the probabilities of the relevant uncertain events." While the setup of the decision-making task used in our study does not allow feedback to be withheld, it is worth exploring whether participants are more rational decision-makers under uncertainty for tasks where immediate feedback is not provided.

The use of incentives assumes that participants who are given a flat payment regardless of how they perform will put in less effort, and overlooks the fact that many participants in the non-incentivised conditions might have strong intrinsic motivation to provide high-quality data and perform well in the study [4]. According to Camerer and Hogarth [3], "some people like mental effort, and those people disproportionately volunteer for experiments!" If this attitude is widespread among crowdworkers on platforms such as Prolific, then the additional effort for implementing incentives might not be worthwhile for most researchers. An additional aspect to consider here are the policies of Prolific as a platform—the quality checks implemented by Prolific could serve as another source of motivation for crowdworkers to provide high quality responses in experiments, reducing the potential effect of performance-based incentives.

Finally, participants in our non-incentivized conditions were still informed on whether they answered a question correctly and how many they answered correctly so far (perception of correlation), or how much budget they have remaining (decision-making). In other words, while participants were not financially rewarded for their performance, they could interpret these as virtual rewards. This was a conscious choice in our experiment in order to minimize the discrepancies between the two experimental conditions. However, it is possible that, in crowdsourcing platforms such as Prolific, if participants are already highly intrinsically motivated to do well, virtual rewards could have a similar effect on effort as real rewards.

7.4 The Possible Role of Incentives in Other Studies

Our work was an initial investigation into the role of incentives in two prototypical types of visualization studies. While our results may not generalize to all visualization tasks, we can speculate on how incentives might impact certain other types of visualization tasks, based on our findings and the *capital-labor-production* theory [3]. Studies on graphical perception [e.g., 5, 9, 26] or color perception [e.g. 49, 50, 51, 65], which require users to perform similarly low-level perceptual tasks as Study 1 are likely not going to benefit from the increased effort than incentives induce.

There are likely a class of visualization tasks such as some matrix-based tasks (e.g., Nobre et al. [43], performed a crowdsourced evaluation of node-link diagrams and adjacency matrices where participants were required to undergo lots of training and perform complex tasks), which either require a lot of a priori knowledge (cognitive capital) or a lot of effort (high production requirements), and thus may not benefit from incentives (and decision-making under uncertainty could possibly fall into this category as well!). In other words, these represent tasks where participants may perform poorly not because they do not want to put in more effort, but rather because they are not sure of how to perform the task well and putting in more effort may not necessarily help them perform the task better.

The type of tasks that we believe may benefit from incentives are ones which do not require a lot of cognitive capital but can feel a bit tedious; incentives might motivate participants to put in the required additional effort to complete the task accurately. An example could be the *finding the shortest path between two nodes* task using node-link or adjacency matrix representations—node-link diagrams have been found to be more accurate [19, 35, 45], but if incentivized, participants may be more willing to put in the effort to perform the tedious steps involved for adjacency matrices. A challenge here is ensuring that the incentive function is well designed and not arbitrary [30, 68]. We hope to explore such tasks in future work.

7.5 Ethical Considerations for Incentivized Experiments

In both experiments, participants in the incentivized conditions took longer to complete the task compared to participants in the baseline condition. Therefore, (i) the initial median hourly wage for participants was lower than the advertised amount, and, after factoring in bonus, payment was below our target fair wage of $15/h; and (ii) the use of performance-based financial incentives increased the variance in wages that participants received—a participant who can perform the task quickly and well can receive a significantly higher wage than a participant who performs it slowly and poorly.

Ensuring a fair median wage for participants is still possible for incentivized experiments, as the researcher can increase the guaranteed amount so that the median wage, taking into account bonuses, reflects what is considered a fair wage. In order to ensure a wage of $15/h, and to comply with Prolific's minimum guaranteed wage of $8/h, we increased the fixed participation fee (by $0.35 in Experiment 1, and by $2.7 in Experiment 2). Addressing the issue of variance in wages, however, is significantly more complicated, and to some degree not compatible with the goal of performance-based incentives. While there is always going to be some degree of variance in compensation in crowdsourced studies, as not every participant takes the same amount of time, using bonuses as a significant portion of the compensation exacerbates this issue. A solution here might be to increase the guaranteed amount to ensure that participants receive a fair hourly wage. However, as research teams typically operate under considerable budget constraints, this would likely mean that they would have to in turn reduce the bonus amounts that participants can earn, making them less "incentivizing."

Running incentivized studies has real costs on the experimenter—increased implementation and design burden while setting up the experiment, the need to manually compensate participants on the crowdsourcing platform, the seemingly low advertised compensation on the crowdsourcing platform (which can have reputational costs for future studies). Taking all of this into account, there is a credible argument to be made against the use of performance-based financial incentives due to the high costs, seemingly low rewards, and the associated ethical questions regarding fair labour practices.

8 Conclusion

We conducted two studies to test the value of performance-based monetary incentives for two different types analysis questions that commonly surface in visualization studies: a lower-level perceptual task, and a higher-level decision-making task. We did not find evidence to suggest that incentives affect performance on either task. To our knowledge, our work was the first formal study of monetary incentives in crowdsourced visualization studies. While we acknowledge that our results may not easily translate to other scenarios (e.g., lab studies, different types of tasks, significantly higher incentive pay, etc.), they seem to suggest that incentives, as currently deployed in many empirical studies, may matter less as a *tacit factor* than we expected. Given recent calls for incorporating decision theory, utility functions and incentives into visualization and HCI experiments [30, 68], our results provide an empirical data point that putting them into practice may not be as straightforward as we might have initially presumed.

Acknowledgments

We would like to thank to Jack Wilburn and Zach Cutler for technical support on this project, and the anonymous reviewers for their feedback on the manuscript. This work was partially funded by the National Science Foundation awards 2213756 and 2403094.

References

[1] A. Achtziger, C. Alós-Ferrer, S. Hügelschäfer, and M. Steinhauser. Higher incentives can impair performance: Neural evidence on reinforcement and rationality. *Social Cognitive and Affective Neuroscience*, 10(11):1477–1483, Nov. 2015. doi: 10.1093/scan/nsv036 2, 9

[2] P. Cala, T. Havranek, Z. Irsova, J. Matousek, Z. Irsova, and J. Novak. Financial Incentives and Performance: A Meta-Analysis of Economics Evidence, Nov. 2022. doi: 10.31222/osf.io/wbe9k 2

[3] C. F. Camerer, R. M. Hogarth, D. V. Budescu, and C. Eckel. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. In B. Fischhoff and C. F. Manski, eds., *Elicitation of Preferences*, pp. 7–48. Springer Netherlands, Dordrecht, 1999. doi: 10.1007/978-94-017-1406-8_2 2, 3, 7, 9

[4] C. P. Cerasoli, J. M. Nicklin, and M. T. Ford. Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, 140(4):980–1008, 2014. doi: 10.1037/a0035661 2, 9

[5] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, Sept. 1984. doi: 10.1080/01621459.1984.10478080 9

[6] 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, Dec. 2014. doi: 10.1109/TVCG.2014.2346298 2, 8

[7] M. Correll, D. Moritz, and J. Heer. Value-Suppressing Uncertainty Palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11. ACM, Montreal QC Canada, Apr. 2018. doi: 10.1145/3173574.3174216 8

[8] Z. Cutler, J. Wilburn, H. Shrestha, Y. Ding, B. Bollen, K. A. Nadib, T. He, A. McNutt, L. Harrison, and A. Lex. ReVISit 2: A Full Experiment Life Cycle User Study Framework, Aug. 2025. doi: 10.48550/arXiv.2508.03876 1, 3, 4

[9] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The Risks of Ranking: Revisiting Graphical Perception to Model Individual Differences in Visualization Performance. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1756–1771, Mar. 2024. doi: 10.1109/TVCG.2022.3226463 9

[10] M. Dong, L. Chen, L. Wang, X. Jiang, and G. Chen. Uncertainty Visualization for Mobile and Wearable Devices Based Activity Recognition Systems. *International Journal of Human–Computer Interaction*, 33(2):151–163, Feb. 2017. doi: 10.1080/10447318.2016.1224527 8

[11] P. Dragicevic, Y. Jansen, A. Sarma, M. Kay, and F. Chevalier. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, Glasgow Scotland Uk, May 2019. doi: 10.1145/3290605.3300295 2

[12] 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*, pp. 1–12. ACM, Montreal QC Canada, Apr. 2018. doi: 10.1145/3173574.3173718 1, 3, 6, 7, 8, 12

[13] N. Ferreira, D. Fisher, and A. C. Konig. Sample-oriented task-driven visualizations: Allowing users to make better, more confident decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 571–580. ACM, Toronto Ontario Canada, Apr. 2014. doi: 10.1145/2556288.2557131 8

[14] A. Franco, N. Malhotra, and G. Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, Sept. 2014. doi: 10.1126/science.1255484 8

[15] R. Fygenson, E. Bertini, and L. M. Padilla. Croissant Charts: Modulating the Performance of Normal Distribution Visualizations with Affordances. *Computer Graphics Forum*, n/a(n/a):e70463, Nov. 2026. doi: 10.1111/cgf.70463 8

[16] R. Fygenson and L. Padilla. Impact of Vertical Scaling on Normal Probability Density Function Plots. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):984–994, Jan. 2025. doi: 10.1109/TVCG.2024.3456396 8

[17] J. Gabry, R. Češnovar, A. Johnson, and S. Bronder. Cmdstanr: R Interface to 'CmdStan', 2025. 4

[18] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28(2):210–216, 2009. doi: 10.1037/a0014474 2

[19] M. Ghoniem, J.-D. Fekete, and P. Castagliola. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In *IEEE Symposium on Information Visualization*, pp. 17–24. IEEE, Austin, TX, USA, 2004. doi: 10.1109/INFVIS.2004.1 9

[20] G. Gigerenzer, R. Hertwig, E. Van Den Broek, B. Fasolo, and K. V. Katsikopoulos. "A 30% Chance of Rain Tomorrow": How Does the Public Understand Probabilistic Weather Forecasts? *Risk Analysis*, 25(3):623–629, 2005. doi: 10.1111/j.1539-6924.2005.00608.x 8, 9

[21] M. Gonzalez-Rubio, P. A. Iturralde, and G. Torres-Oviedo. Weber's Law in walking: Sensory scaling is observed in multi-sensory, dynamic tasks. *Scientific Reports*, June 2026. doi: 10.1038/s41598-026-54948-5 4

[22] 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*, pp. 1–12. ACM, Montreal QC Canada, Apr. 2018. doi: 10.1145/3173574.3174079 8

[23] B. D. Hall, Y. Liu, Y. Jansen, P. Dragicevic, F. Chevalier, and M. Kay. A Survey of Tasks and Visualizations in Multiverse Analysis Reports. *Computer Graphics Forum*, 41(1):402–426, 2022. doi: 10.1111/cgf.14443 2

[24] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, Dec. 2014. doi: 10.1109/TVCG.2014.2346979 1, 2, 3, 4

[25] P. D. Harvey. Domains of cognition and their assessment. *Dialogues in Clinical Neuroscience*, 21(3):227–237, Sept. 2019. doi: 10.31887/DCNS.2019.21.3/pharvey 2, 3

[26] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 203–212. Association for Computing Machinery, New York, NY, USA, Apr. 2010. doi: 10.1145/1753326.1753357 9

[27] J. Hullman, P. Resnick, and E. Adar. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE*, 10(11):e0142444, Nov. 2015. doi: 10.1371/journal.pone.0142444 2, 8

[28] S. Joslyn and J. LeClerc. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science*, 22(4):308–315, Aug. 2013. doi: 10.1177/0963721413481473 1, 2

[29] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126–140, 2012. doi: 10.1037/a0025185 2, 3, 5

[30] A. Kale. Toward a Logic of Generalization about Visualization as a Decision Aid. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pp. 1–5, Nov. 2025. doi: 10.1109/VIS60296.2025.00005 3, 9

[31] 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, Feb. 2021. doi: 10.1109/TVCG.2020.3030335 1, 3, 4, 6, 7, 8, 12

[32] M. Kay. Tidybayes: Tidy Data and Geoms for Bayesian Models, Sept. 2024. doi: 10.5281/zenodo.13770114 12

[33] M. Kay and J. Heer. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, Jan. 2016. doi: 10.1109/TVCG.2015.2467671 2, 4

[34] 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*, pp. 5092–5103. ACM, San Jose California USA, May 2016. doi: 10.1145/2858036.2858558 2, 8

[35] R. Keller, C. M. Eckert, and P. J. Clarkson. Matrices or Node-Link Diagrams: Which Visual Representation is Better for Visualising Connectivity Models? *Information Visualization*, 5(1):62–76, Mar. 2006. doi: 10.1057/palgrave.ivs.9500116 9

[36] J. F. Landy, M. L. Jia, I. L. Ding, D. Viganola, W. Tierney, A. Dreber, M. Johannesson, T. Pfeiffer, C. R. Ebersole, Q. F. Gronau, A. Ly, D. van den Bergh, M. Marsman, K. Derks, E.-J. Wagenmakers, A. Proctor, D. M. Bartels, C. W. Bauman, W. J. Brady, F. Cheung, A. Cimpian, S. Dohle, M. B. Donnellan, A. Hahn, M. P. Hall, W. Jiménez-Leal, D. J. Johnson, R. E. Lucas, B. Monin, A. Montealegre, E. Mullen, J. Pang, J. Ray, D. A. Reinero, J. Reynolds, W. Sowden, D. Storage, R. Su, C. M. Tworek, J. J. Van Bavel, D. Walco, J. Wills, X. Xu, K. C. Yam, X. Yang, W. A. Cunningham, M. Schweinsberg, M. Urwitz, The Crowdsourcing Hypothesis Tests Collaboration, and E. L. Uhlmann. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5):451–479, 2020. doi: 10.1037/bul0000220 2

[37] J. LeClerc and S. Joslyn. The Cry Wolf Effect and Weather-Related Decision Making. *Risk Analysis*, 35(3):385–395, 2015. doi: 10.1111/risa.12336 1

[38] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch. Risk as feelings. *Psychological Bulletin*, 127(2):267–286, 2001. doi: 10.1037/0033-2909.127.2.267 8

[39] 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, Dec. 2012. doi: 10.1109/TVCG.2012.279 8

[40] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pp. 77–85. Association for Computing Machinery, New York, NY, USA, June 2009. doi: 10.1145/1600150.1600175 2

[41] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman and Hall/CRC, New York, 2 ed., Mar. 2020. doi: 10.1201/9780429029608 2

[42] L. Nadav-Greenberg and S. L. Joslyn. Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, Sept. 2009. doi: 10.1518/155534309X474460 2

[43] C. Nobre, D. Wootton, L. Harrison, and A. Lex. Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, Honolulu HI USA, Apr. 2020. doi: 10.1145/3313831.3376381 9

[44] C. Nobre, K. Zhu, E. Mörth, H. Pfister, and J. Beyer. Reading Between the Pixels: Investigating the Barriers to Visualization Literacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17. ACM, Honolulu HI USA, May 2024. doi: 10.1145/3613904.3642760 1

[45] M. Okoe, R. Jianu, and S. Kobourov. Node-Link or Adjacency Matrices: Old Question, New Insights. *IEEE Transactions on Visualization and Computer Graphics*, 25(10):2940–2952, Oct. 2019. doi: 10.1109/TVCG.2018.2865940 9

[46] B. Oral, P. Dragicevic, A. Telea, and E. Dimara. Decoupling Judgment and Decision Making: A Tale of Two Tails. *IEEE Transactions on Visualization and Computer Graphics*, 30(10):6928–6940, Oct. 2024. doi: 10.1109/TVCG.2023.3346640 1

[47] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1):40, Dec. 2017. doi: 10.1186/s41235-017-0076-1 2

[48] L. M. K. Padilla, M. Powell, M. Kay, and J. Hullman. Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. *Frontiers in Psychology*, 11, Jan. 2021. doi: 10.3389/fpsyg.2020.579267 1, 3, 7, 12

[49] K. Reda. Rainbow Colormaps: What are They Good and Bad for? *IEEE Transactions on Visualization and Computer Graphics*, 29(12):5496–5510, Dec. 2023. doi: 10.1109/TVCG.2022.3214771 9

[50] K. Reda, P. Nalawade, and K. Ansah-Koi. Graphical Perception of Continuous Quantitative Maps: The Effects of Spatial Frequency and Colormap Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–12. Association for Computing Machinery, New York, NY, USA, Apr. 2018. doi: 10.1145/3173574.3173846 9

[51] K. Reda and D. A. Szafir. Rainbows Revisited: Modeling Effective Colormap Design for Graphical Inference. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1032–1042, Feb. 2021. doi: 10.1109/TVCG.2020.3030439 9

[52] R. A. Rensink and G. Baldridge. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x 1, 2, 3

[53] D. P. Retchless and C. A. Brewer. Guidance for representing uncertainty on global temperature change maps. *International Journal of Climatology*, 36(3):1143–1159, 2016. doi: 10.1002/joc.4408 8

[54] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979. doi: 10.1037/0033-2909.86.3.638 8

[55] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, Apr. 2016. doi: 10.1080/13875868.2015.1137577 2

[56] A. Sarma, M. Hedayati, and M. Kay. More Forecasts, More (Decision) Problems: How Uncertainty Representations for Multiple Forecasts Impact Decision-Making, Feb. 2025. doi: 10.31219/osf.io/t4e9u_v1 1, 3, 5, 8

[57] A. Sarma, K. Hwang, J. Hullman, and M. Kay. Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, Honolulu HI USA, May 2024. doi: 10.1145/3613904.3642375 2

[58] A. Sarma, A. Kale, M. J. Moon, N. Taback, F. Chevalier, J. Hullman, and M. Kay. Multiverse: Multiplexing Alternative Data Analyses in R Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, Hamburg Germany, Apr. 2023. doi: 10.1145/3544548.3580726 2

[59] A. Sarma, S. Long, M. Correll, and M. Kay. Tasks and Telephones: Threats to Experimental Validity due to Misunderstandings of Visualisation Tasks and Strategies Position Paper. In *2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 33–40, Oct. 2024. doi: 10.1109/BELIV64461.2024.00009 2

[60] A. Sarma, X. Pu, Y. Cui, M. Correll, E. T. Brown, and M. Kay. Odds and Insights: Decision Quality in Exploratory Data Analysis Under Uncertainty. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–14. Association for Computing Machinery, New York, NY, USA, May 2024. doi: 10.1145/3613904.3641995 1, 3, 7, 8

[61] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366, Nov. 2011. doi: 10.1177/0956797611417632 2

[62] U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, Nov. 2020. doi: 10.1038/s41562-020-0912-z 1, 2

[63] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5):702–712, Sept. 2016. doi: 10.1177/1745691616658637 1, 2

[64] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957. doi: 10.1037/h0046162 2, 8

[65] D. A. Szafir. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, Jan. 2018. doi: 10.1109/TVCG.2017.2744359 9

[66] R. C. Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2024. 4

[67] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, and M. A. L. M. van Assen. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7, Nov. 2016. doi: 10.3389/fpsyg.2016.01832 2

[68] Y. Wu, Z. Guo, M. Mamakos, J. Hartline, and J. Hullman. The Rational Agent Benchmark for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):338–347, Jan. 2024. doi: 10.1109/TVCG.2023.3326513 3, 9

[69] F. Yang, M. Hedayati, and M. Kay. Subjective Probability Correction for Uncertainty Representations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–17. Association for Computing Machinery, New York, NY, USA, Apr. 2023. doi: 10.1145/3544548.3580998 1, 3, 5

[70] H. Zhang and L. T. Maloney. Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. *Frontiers in Neuroscience*, 6, Jan. 2012. doi: 10.3389/fnins.2012.00001 5

A Interpreting the Results of Experiment 2

Calculation of Benchmarks

As described in section 5, there are two payoff relevant states (of nature), $S = \{s_1 = Pr(T \leq 0), s_2 = Pr(T > 0)\}$ and two possible actions $A = \{a_1, a_2\}$. Following prior studies in decision-making under uncertainty in visualization [e.g., 12, 31, 48], to evaluate participant decisions for each trial, we simulate a draw from the visualized temperature distribution which determines the state of the world S (i.e., whether the temperature is below or above freezing). We then calculate the cost incurred based on the payoff matrix described in section 5. To assess participants performance, we use three benchmarks—the expected utility of decision-maker responding at random, the expected utility of decision-maker with extreme risk-aversion (deciding to salt on every trial except the attention check ones), and the expected utility of completely rationally decision-maker.

$$\begin{aligned}
\mathbb{E}(U|\text{random}) &= -1000 \cdot \left(\sum_{k}^{n} 0.5 \cdot 1 + 0.5 \cdot 5 \cdot p_k \right) \\
&= -3272 \\
\mathbb{E}(U|\text{risk-averse}) &= \sum_{i}^{n} -1000 \\
&= 0 \\
\mathbb{E}(U|\text{rational}) &= -1000 \cdot \left(\sum_{k}^{n} 1 \cdot \mathbb{I}(p_i \geq 0.2) + 5 \cdot \mathbb{I}(p_i < 0.2) \cdot p_k \right) \\
&= 3960
\end{aligned}$$

where p_k represents the probability of freezing in a particular trial k and lies in $\{0.595, 0.5, 0.405, 0.315, 0.235, 0.168, 0.115, 0.075, 0.046\}$ each repeated twice, and $n = 18$.

Calculation of the Posterior Predictive Credible Interval

In Figure 6, we show the posterior predictive credible interval for expected utility in each condition based on the model estimated probability of salting for the average participant on a specific trial k: $p_{\text{SALT}} = \text{logit}^{-1}(\alpha_i + \beta_i \cdot [\text{logit}(p_k) - \text{logit}(0.2)])$. The expected utility is given by:

$$\mathbb{E}(U) = -1000 \cdot \left(\sum_{k}^{n} \mathbb{E}(p_{\text{SALT}}) + 5 \cdot (1 - \mathbb{E}(p_{\text{SALT}})) \cdot p_k \right)$$

We sample draws from the expectation of the posterior predictive distribution using the `add_epred_draws` function from `tidybayes 3.0.7` [32].

The Probability of Performing Better than the Rational Benchmark In Figure 6, we find that approximately 30% of the participants performed better than the rational benchmark. While on the face of it, this might seem surprising, it is actually not that unlikely—the rational benchmark is an asymptotic guarantee given an infinite number of trials; for obvious purposes, we limit participants to 18 trials. To perform better than the rational benchmark ($\mathbb{E}(U) > \mathbb{E}(U|\text{rational})$), a participant simply has to get lucky and not encounter a freezing state of nature (i.e., s_1) on any of the trials where they decide to not salt. This is given by:

$$\prod_{k}^{n} (1 - p_i) \cdot \mathbb{I}(p_i \leq d)$$

where d is the decision boundary and $d = 0.2$ for the rational decision-maker. The probability of outperforming the rational benchmark for a participant making decisions using $d = 0.2$ is 0.42. If $d = 0.3$, the probability of outperforming the rational benchmark is 0.25.[5]

[5]This is actually the lower bound probability. In reality, if $d \geq 0.3$, the participant can actually encounter a freezing state of nature for one of the trials and still end up with utility greater than the rational benchmark. Such a participant will salt on eight of the trials where $p_k \geq 0.3$ leaving them with a budget of $10,000; if they encounter a freezing state of nature in one of the remaining trials, their remaining budget will be $5,000 which is still greater than the rational benchmark

B Experiment 3: Decision-Making under Uncertainty

We conducted a replication of Experiment 2, where we examined the impact of incentives on performance in the same decision-making task using either 95% interval (as opposed to 66% and 95% interval) or density plots as uncertainty representations. The rest of the experimental materials (task, procedure, tutorials and training) and data analysis model were the same as Experiment 2. As this experiment was conducted after the initial review cycle, we only include it as an appendix.

Participants: We recruited all participants from Prolific. Our experiment was only eligible to participants who were fluent in English, and on desktop devices. As per our pre-registrations, we aimed to recruit 240 participants (60 participants in each condition). After excluding participants who failed to meet our pre-registered attention check criteria (eleven), we had 229 participants (56 in the base interval, 60 in incentivised interval, 55 in base density, and 58 in incentivised density). We received explicit consent from all participants to collect and share their responses. The median completion time for participants in the baseline condition was approximately 13 mins, and they were compensated $3.5 (approximately $16/h); the median completion time for participants in the incentivized condition was approximately 13.5 mins, and they received a guaranteed amount of $1.9 (adjusted up from $1.5) and the average bonus was $1.5 (corresponding to an average wage of $15/h).

The results from this experiment are shown in Figure 8. Similar to Experiment 2, we found the value of α to be negative across all conditions (Figure 8A)—this means that the crossover point for the average participant is greater than 0.2, suggesting a consistent bias towards risk-seeking behavior. We also found the value of β to be significantly larger than one across all conditions (Figure 8B), and highest for the incentivised interval condition, indicating that participants are quite sensitive to the stimuli and their subjective crossover point. Figure 8C allows to compare decision quality. We again do not find any meaningful effect of incentives. In addition, we found the expected utility to be greater for both interval conditions (base interval: 1.73, [1.42, 2.00] and incentivised interval: 1.53, [1.22, 1.80]) compared to

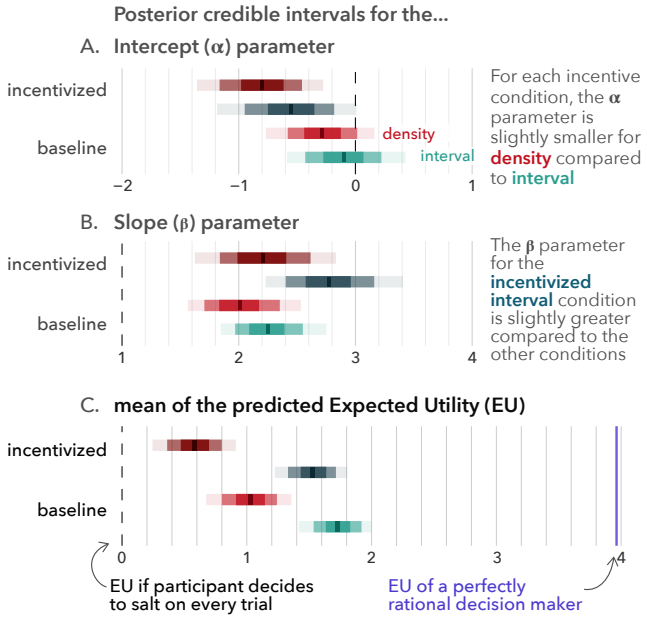

Fig. 8: The main result of Experiment 3. We show the posterior credible intervals of α, β and the mean expected utility for both visual representations across the incentivized and baseline conditions.

the density conditions (base density: 1.03, [0.67, 1.36] and incentivised density: 0.58, [0.24, 0.91]). While this result is consistent with our previous experiment, it is contrary to the findings of prior work which have generally found that visualizing uncertainty using density plots lead to better decision quality.

As in the previous two studies, we again find that participants took slightly longer to complete the task in the incentivized conditions compared to the baseline conditions. Figure 3B, visualizes the bootstrapped median and 95% quantile intervals for the median estimate. The difference in time spent is approximately 22s for intervals (base interval: 179s, [160s, 201s], incentivised interval: 201s, [170s, 217s]), and approximately 30s for density (base density: 194s, [161s, 239s], incentivised density: 224s, [186s, 292s]). This increase of 10-15% is smaller than what was observed for Experiment 2, and similar to the difference observed in Experiment 1.