The Rational Agent Benchmark for Data Visualization

Yifan Wu, Ziyang Guo, Michalis Mamakos, Jason Hartline, and Jessica Hullman

Abstract—Understanding how helpful a visualization is from experimental results is difficult because the observed performance is confounded with aspects of the study design, such as how useful the information that is visualized is for the task. We develop a rational agent framework for designing and interpreting visualization experiments. Our framework conceives two experiments with the same setup: one with behavioral agents (human subjects), and the other one with a hypothetical rational agent. A visualization is evaluated by comparing the expected performance of behavioral agents to that of a rational agent under different assumptions. Using recent visualization decision studies from the literature, we demonstrate how the framework can be used to pre-experimentally evaluate the experiment design by bounding the expected improvement in performance from having access to visualizations, and post-experimentally to deconfound errors of information extraction from errors of optimization, among other analyses.

Index Terms—Evaluation, decision-making, rational agent, scoring rule

1 Introduction

Intuition-driven design guidelines for designing data visualizations are increasingly being replaced with data-driven recommendations based on visualization studies. To assess the extent to which modern empirical study of visualizations does in fact capture the value of visualization, however, requires accounting for the design of the experimental task and conditions for study. To understand how well people performed with a visualization in a controlled study, or how important an observed difference in performance between two visualization is, we must understand what sorts of performance differences an experimental scenario admits.

However, it can be difficult in designing a study to predict how the choices one makes impact the experiment's capability for capturing meaningful performance differences. We can liken the experiment design process to setting various "knobs" that will impact the difficulty of the task, the extent to which participants are motivated to study the visualization to complete the task, and the best achievable performance on the task. These knobs include the input distributions used to generate stimuli, the allocation of these inputs across participants, and the payoff function that will reward participants for making good decisions. More broadly applicable experiment design decisions include how many participants to target and how to compare key interventions (e.g., between-subjects, pre-post design, etc.).

While it is difficult to define "optimal" choices for these myriad decisions, the results of a study can still provide useful knowledge about visualization performance when properly conditioned on the *potential* the study had, whether to show differences between visualization strategies or to evaluate a specific strategy. For example, a canonical form of conditioning used to assess a study's potential to detect an effect such as a difference between treatments ensures that the study design provides sufficient statistical power to detect an effect of the hypothesized size.

More generally, we would like an approach to interpreting the results of a study comparing visualization strategies that helps a reader answer questions like the following:

- How hard is the task? For example, how well could we expect someone do without consulting the visualized data at all?
- Considering the study design alone, how incentivized would we expect participants to be to use the visualized information?
- To what extent are observed differences in performance likely to stem from informational asymmetries in the visualizations (e.g., visualizing only a mean versus a more expressive depiction of a distribution)?
- To what extent is sub-optimal performance with a visualization due to participants not differentiating the task-relevant information it provides, versus not being able to properly use the information they gained to choose a response?

Our inability to answer the above questions from many empirical research papers highlight how visualization research lacks clear comparison points, or performance *benchmarks* that can aid the design and interpretation of controlled visualization experiments. Answering such questions contextualizes what was learned from observing the performance of any single visualization in absolute terms defined on the experiment design. Without clear benchmarks, readers and authors alike tend to draw conclusions from coarse, *relative* information like visualization performance rankings. A good set of benchmarks are necessary to assess the fitness of the experiment design itself for studying a given visualization research question.

We contribute a rational agent framework based on quantifying the value of information to a judgment or decision problem. Our framework defines benchmark measures representing attainable performance given a visualization experiment design. Benchmarks defined in the rational agent framework can be applied before an experiment is run to vet how capable the experiment design is of showing important differences between visualizations and of resolving good performance with any single visualization. Applying the framework after an experiment provides further insight into behavioral agent performance, by enabling the researcher to deconfound sources of erroneous answers. For example, agents might be unable to extract the information from the visualization, or unable to optimally translate the information to a decision.

We apply the framework to two well-regarded visualization experiments from the literature: one on the impact of visualization design on effect size judgments and decisions [19] and one on the impact of visualization design on transit decisions [6]. In both cases, we identify 1) ways in which the experiment design could have been improved (through different measures or payoff functions) and 2) sources of loss that help explain behavioral results but were not fully addressed in the original presentations of results.

- *Yifan Wu is with Northwestern University. E-mail: yifan.wu@u.northwestern.edu*
- *Ziyang Guo is with Northwestern University. E-mail: ziyangguo2027@u.northwestern.edu*
- *Michalis Mamakos is with Northwestern University. E-mail: michailmamakos2022@u.northwestern.edu*
- *Jason Hartline is with Northwestern University. E-mail: hartline@northwestern.edu*
- *Jessica Hullman is with Northwestern University. E-mail: jhullman@northwestern.edu*

Manuscript received 31 March 2023; revised 1 July 2023; accepted 8 August 2023. Date of publication 23 October 2023; date of current version 21 December 2023. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2023.3326513, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2023.3326513

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.
1077-2626 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

2 Related Work

2.1 Visualization Evaluation

Our work aims to improve evaluation methods in visualization. Previously, researchers have contributed overviews of qualitative and quantitative approaches [16, 25, 33] and conceptual models and approaches for ensuring that one selects an evaluation that is appropriate for a given task, context, or contribution type [15, 28, 30].

Whenever visualizations are meant to support inference in addition to merely describing an observed dataset [12], the evaluation approach should define a standard for assessing the quality of the inference. However, several recent surveys of evaluative studies for visualizations [5] and uncertainty visualizations specifically [13, 23] suggest that the use of well-defined judgment and decision tasks is rare. Instead, a majority of uncertainty visualization studies rely on measures of perceptual accuracy and/or self-reports of satisfaction, confidence, or other properties that may have an unclear or even opposite relationship with rational use of the information for the problem at hand [13, 23]. This has led some researchers to advocate for adopting Bayesian inference as a benchmark against which to compare reactions to visualizations [12, 20, 22]. These models use the deviation of human performance from the Bayesian ideal as a means of better understanding patterns in human judgments, and for inspiring new design approaches [12, 18]. While human judgments need not be perfectly Bayesian for such approaches to lead to a better understanding of how people use visualizations, if there is no correspondence between human behavior and the Bayesian agent's behavior, design suggestions aimed at aligning the human behavior with the Bayesian's them may not be effective. In contrast to prior applications of Bayesian theory to visualization, the value of the rational agent framework does not depend on actual humans acting like rational agents. Our work is related to ideal observer analysis, used in psychophysics, which theoretically upperbounds behavioral performance by a Bayesian agent in the same situation in order to reason about factors influencing human perception [24]. However, our framework defines the baseline performance in addition to the upperbound, and hence provides a "scale" for interpreting behavioral performance and a means to separate sources of loss in decision-making.

2.2 Interpreting experiment results

Our work is related to recent integrative modeling [10] approaches to benchmarking the irreducible variance in data used for modeling [1, 7]. For example, the explanatory power of theories embedded in behavioral models can be assessed by quantifying irreducible error inherent in an experimental task [7], grounding a perspective for how well a model performs. We take a similar approach, but with the goal of benchmarking how well humans can be expected to do under different assumptions when faced with an experimental task.

3 The Rational Agent Framework

The value of the information presented in a visualization can be quantified by how much it improves the expected payoff in a decision problem. The visualized information reduces uncertainty about a payoff-relevant state, thus helping the agent make better decisions. The value of the visualization can be understood as the expected improvement in payoff when an agent has access to the visualization.

Our framework conceives two studies, an experimental study and a theoretical one. The first occurs in the real world with behavioral participants, and the other is based on an analysis of a hypothetical rational world with a rational agent participant. We assume an experiment design as input, including information on how stimuli will be generated, what decisions or beliefs participants will report, and how their responses will be incentivized and scored. If the experiment has already been conducted, the raw or modeled behavioral results are also part of the input. The two studies assume exactly the same decision problem and data-generating process, enabling analysis of an experiment both before and after it is run.

Below we establish preliminaries, including what constitutes a visualization experiment in our framework, the conceptual devices of the rational and behavioral agent, and how they are used in pre- and post-experimental analyses. We apply these definitions to an example forecast visualization experiment.

Payoff-relevant state	$\theta \in \Theta$
Signal (visualization)	$v \in V$
Data generating process	$\pi \in \Delta(V \times \Theta)$
Agent's action	$a \in A$
Scoring rule (payoff)	$S : A \times \Theta \to \mathbb{R}$

Table 1: Notation for defining a visualization experiment (assuming a single visualization strategy).

3.1 Decision Problems

Decision theory provides a natural framework for understanding an agent's task in a visualization study. A decision problem starts by assuming a state space Θ that describes the set of finite values (scenarios) that an uncertain state can take. Each possible state $\theta \in \Theta$ is a description of reality, and only one may hold at a time. A *data generating model* defines a distribution over scenarios $p \in \Delta(\Theta)$. In many experiments the distribution over states is uniform.

A decision problem is defined by a distribution over states $p \in \Delta(\Theta)$ an action space A and a *scoring rule* $S : A \times \Theta \to \mathbb{R}$ that maps the action and state to a quality or payoff. Given a distribution p and scoring rule S denote the expected score of an action by $S(a, p) = \mathbf{E}_{\theta \sim p}[S(a, \theta)]$. The optimal decision for a distribution p is the one with the highest expected quality, i.e. $a^* = \arg\max_{a \in A} S(a, p)$.

In decision problems corresponding to prediction tasks, the action space is a probabilistic belief over the state space, i.e., $A = \Delta(\Theta)$. For such problems, a scoring rule is said to be *proper* if the optimal action is to predict the true distribution, i.e., $p = \arg\max_{a \in A} S(a, p)$. Squared loss, a.k.a., the quadratic scoring rule, is an example of a proper scoring rule that measures the accuracy of beliefs. For any scoring rule $S : A \times \Theta \to \mathbb{R}$ there is an equivalent *proper scoring rule* $\hat{S} : \Delta(\Theta) \times \Theta \to \mathbb{R}$ defined by playing the optimal action under the reported belief. Formally,

$$\hat{S}(p, \theta) = S(\arg\max_{a \in A} S(a, p), \theta). \tag{1}$$

Example We illustrate the framework with a hypothetical weather forecast experiment, loosely inspired by [29]. Imagine a researcher who wants to compare people's performance in making a decision using several visualization strategies for presenting a predicted daily low temperature with uncertainty (i.e., a temperature distribution). They define a task in which the participant must decide whether to salt the parking lot or not, i.e., by selecting action a from action space $A = \{0 = \text{no salt}; 1 = \text{salt}\}$. They plan to score the participants for each decision task by simulating a temperature according to the predicted distribution. The payoff relevant state θ is from state space $\Theta = \{0 = \text{not freezing}, 1 = \text{freezing}\}$, corresponding to whether the simulated temperature was above or below the freezing point. Given the state space $\Theta = \{0 = \text{not freezing}; 1 = \text{freezing}\}$ the experimenter endows the following payoff function as a scoring rule:

$$S(a, \theta) = \begin{cases} 0 & \text{if } a = 0, \theta = 0 \quad \textit{no salt, not freezing} \\ -100 & \text{if } a = 0, \theta = 1 \quad \textit{no salt, freezing} \\ -10 & \text{if } a = 1, \theta = 0 \quad \textit{salt, not freezing} \\ 0 & \text{if } a = 1, \theta = 1 \quad \textit{salt, freezing} \end{cases} \tag{2}$$

3.2 Information Structures and Visualizations

In a visualization experiment, the subject is given a stimulus in the form of a visualization that is associated with the state. Since the visualization is associated with the state, if the subject understands the visualization well, he can improve his performance at the decision task.

To gauge the performance of a behavioral subject in such a task we introduce the rational agent who faces the same task with the same stimulus. Formally, a visualization strategy induces an information structure that is given by a joint distribution $\pi \in \Delta(V \times \Theta)$ over signals $v \in V$ (corresponding to the visualization) and states $\theta \in \Theta$. This joint distribution assigns to each realization $(v, \theta) \in V \times \Theta$ a probability denoted $\pi(v, \theta)$. The joint distribution allows us to calculate expected performance in the experiment. In the data generating process, there

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

may be a fine-grained state $x \in X$ which determines the payoff-relevant state θ, i.e. there exists a function $\hat{\theta}$ that $\theta = \hat{\theta}(x)$.

Our framework allows us to study the performance of a single visualization strategy, or to compare a set of k visualization strategies, inducing information structures $\pi_1, \pi_2, \ldots, \pi_k$, respectively.

Example The experimenter decides to evaluate a few different visualization strategies that can be used to present a weather forecast (Figure 1) for the decision problem they designed (Section 3.1). One shows only the expected daily low temperature. Another shows the expected low plus an interval expressing a 95% confidence interval on the point estimate. Two others depict the probability distribution over possible low temperatures as a gradient plot (plotting probability as opacity) and animated hypothetical outcome plot (HOPs) [14] (plotting probability as frequency).

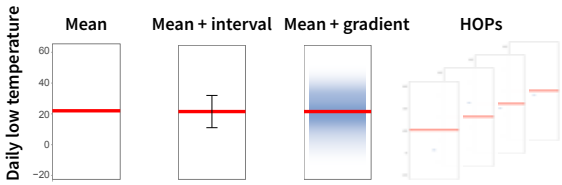

Fig. 1: Example visualizations for a hypothetical weather forecast task.

They define a data-generating process as follows: the daily low temperature is generated from a Gaussian distribution $N(\mu, \sigma^2)$ with a deterministic mean $\mu = 5°C$ and standard deviation σ. The standard deviation σ is uniformly drawn from $\{2,3,4,5\}$.

For visualization strategies that depict uncertainty (CI, gradient, HOPs), the signal v is (μ, σ); for the visualization of the mean, the signal v is deterministically μ.

The data-generating process results in a joint distribution $\pi \in \Delta(V \times \Theta)$ on signal and state for the three non-trivial visualization strategies. The joint distribution allocates probability to getting a decision task for different combinations of θ and σ in Section 3.2.

visualization v for σ	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
$\theta = 0$	0.24845	0.23805	0.2236	0.2103
$\theta = 1$	0.00155	0.01195	0.0264	0.0397

Table 2: The joint distribution $\pi \in \Delta(V \times \Theta)$ on signal and state for the three non-trivial visualization strategies in the weather forecasting experiment.

The notation for the weather forecasting experiment is summarized in Table 3.

Payoff-relevant state	$\theta \in \{0,1\}$ $=$ {not freezing, freezing}
Data generating model	• fine-grained state: daily low temperature $x \sim N(\mu, \sigma^2)$; $\theta = \hat{\theta}(x) = \mathbf{1}[x \leq 0]$ $\Pr[\theta = 1] = \Pr[t \leq 0]$; $\mu = 5$ fixed; σ uniformly from $\{2,3,4,5\}$. • equivalently, $\Pr[\theta = 1]$ uniformly from 0.62%, 4.78%, 10.56%, 15.87%.
Agent's action	$a \in \{0 =$ no salt, $1 =$ salt$\}$
Signal (visualization)	$v^{\text{vis}} \in V^{\text{vis}}$, vis $=$ visualization strategies vis $\in$ {mean, CI, gradient, HOPs} of temperature
Scoring rule (payoff)	$S(a, \theta)$ (see Eq. (2))

Table 3: Notation for the freezing-salting example.

The agent's belief about the freezing state θ can be represented by the probability $p = \Pr[\theta = 1]$ of freezing. The corresponding proper scoring rule is

$$\hat{S}(p,\theta) = \begin{cases} 0 & \text{if } p \leq 0.1, \theta = 0 \quad \textit{no salt, not freezing} \\ -100 & \text{if } p \leq 0.1, \theta = 1 \quad \textit{no salt, freezing} \\ -10 & \text{if } p > 0.1, \theta = 0 \quad \textit{salt, not freezing} \\ 0 & \text{if } p > 0.1, \theta = 1 \quad \textit{salt, freezing} \end{cases} \tag{3}$$

3.3 The Rational Agent: Baseline, Benchmark, and Information Value

Two key constructs in our analysis of a behavioral agent are the decisions of a rational agent without the visualization and with the visualization. In each case, the rational agent makes perfect use of the information available to them. In the case where they have access to a visualization, they do so by Bayesian updating from the joint distribution π to a posterior belief. Here we define the rational agent for a single visualization strategy.

The rational agent's belief prior to the stimulus is their *prior distribution*:

$$p(\theta) = \sum_{v \in V} \pi(v, \theta). \tag{4}$$

The rational agent's belief after the stimulus is their *posterior distribution*. The posterior belief is defined by following Bayes rule:

$$q(\theta) = \pi(\theta|v) = \frac{\pi(v,\theta)}{\sum_{\theta \in \Theta} \pi(v,\theta)}. \tag{5}$$

These two constructs induce a performance of the rational agent which can be compared to the performance of the behavioral agent. For a scoring rule S and information structure π, denote the corresponding proper scoring rule by $\hat{S}$, prior distribution by p, and posterior distribution by $\pi(\theta|v)$. Consider:

rational baseline: The rational baseline is the performance of the rational agent without access to the signal, i.e., with only the prior belief.

$$\mathrm{R}_{\varnothing} = \mathbf{E}_{\theta \sim p}[\hat{S}(p, \theta)]. \tag{6}$$

rational benchmark (visualization optimal) The rational benchmark is the performance of the rational agent with access to the signal, i.e., with the posterior belief.

$$\mathrm{R}_V = \mathbf{E}_{(v,\theta) \sim \pi}[\hat{S}(\pi(\theta|v), \theta)]. \tag{7}$$

The expected payoff of any behavioral agent with the same visualization is below the rational benchmark.

value of information: The difference between the rational benchmark and the rational baseline quantifies the value of the information being visualized in the context of the scoring rule:

$$\Delta = \mathrm{R}_V - \mathrm{R}_{\varnothing}.$$

The value of information provides a unit of difference in expected score for comparing behavioral performance.

3.3.1 Multiple Visualization Strategies

When the framework is applied to multiple visualization strategies, the visualization optimal may vary. To compare multiple visualization strategies, the rational benchmark is defined with regards to the most helpful visualization. Suppose the experimenter is comparing a set of k different visualization strategies, with information structures $\pi^1, \ldots, \pi^k$.

visualization optimal: The visualization optimal is the performance of the rational agent with access to the signal, i.e., with the posterior belief.

$$\mathrm{R}^k_V = \mathbf{E}_{(v,\theta) \sim \pi^k}[\hat{S}(\pi^k(\theta|v), \theta)]. \tag{8}$$

The expected payoff of any behavioral agent with the same visualization is below the visualization optimal.

rational benchmark: Given multiple visualization strategies, the rational benchmark is instead defined as the best performance of the rational agent across different visualization strategies. Suppose the experimenter aims to compare visualization formats $1 \ldots k$, inducing information structures $\pi^1, \ldots, \pi^k$. The rational benchmark is defined as

$$\mathrm{R}^R_V = \max_i \mathbf{E}_{(v,\theta) \sim \pi^i}[\hat{S}(\pi^i(\theta|v), \theta)]. \tag{9}$$

In addition to behavioral losses due to not properly receiving information or not optimizing one's decision (discussed below), we define an information loss induced by information asymmetry across visualizations, quantifying the extent to which visualization strategies provide varying amounts of information about the uncertain state.

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

information loss The information loss captures the loss of information when data is summarized into a less informative visualization. We measure the information loss for a given visualization strategy by the difference $(\mathrm{R}_V^R - \mathrm{R}_V)/\Delta$ between the rational agent benchmark (the rational best performance across visualizations) and the visualization optimal for a particular visualization strategy.

Example We pre-experimentally analyze the hypothetical weather forecast experiment.

We first calculate the prior and posterior distributions of the rational agent. Note that a distribution p on a binary state space $\Theta = \{0,1\}$ can be fully described by the probability that the binary state is $\theta = 1$ (freezing). From Eq. (4) we have the prior probability of freezing $p = 0.0796$. and the posterior probabilities are $\Pr[\theta = 1|\sigma] = 0.62\%, 4.78\%, 10.56\%, 15.87\%$, relatively for $\sigma = 2,3,4,5$, as given in Table 3.

Figure 2 depicts the expected score of the agent for both no-salt and salt actions as a function of her belief p, as specified in Equation (2). Notice that if the belief is certainty either 0 or 1, then the payoff is given explicitly by the scoring rule. For an uncertain belief $p \in (0,1)$ between 0 and 1 the payoff is given by linearly interpolating between certain beliefs, i.e., the payoff is the expected value of the action over the belief. Lines correspond to the no-salt and salt action. The optimal action for each posterior belief – i.e., the action taken by the rational agent – can be read off as well. For each signal, we find its posterior on the horizontal axis, and evaluate which of the two actions give a higher payoff and take that one. From this analysis it is clear that the no-salt action $a = 0$ is taken on the lower two signals $\{2,3\}$ and the salt action $a = 1$ is taken on the higher two signals $\{4,5\}$. The payoff lines cross at $p = 0.1$ where the decision-maker is indifferent between no-salt and salt actions, so the proper scoring rule in Equation (3) sets belief threshold at $p = 0.1$.

The rational agent framework gives the following quantities:

rational baseline: $\mathrm{R}_\varnothing = -7.96$.

The prior $p = 0.08$ is optimized at no-salt and gives an expected payoff of -7.96.

visualization optimal: $\mathrm{R}_V^{\mathrm{CI}} = \mathrm{R}_V^{\mathrm{gradient}} = \mathrm{R}_V^{\mathrm{HOPs}} = -5.69$; $\mathrm{R}_V^{\mathrm{mean}} = -7.96$.

In CI, gradient, and HOPs, each signal arises with probability $1/4$ and the average of the optimal actions under the induced posteriors (read off Figure 2) gives $\mathrm{R}_V = -5.69$. For the visualization of the mean, the rational agent has only the prior information and obtains $\mathrm{R}_V^{\mathrm{mean}} = \mathrm{R}_\varnothing = -7.59$.

rational benchmark: $\mathrm{R}_V^R = \max_{\mathrm{vis}} \mathrm{R}_V^{\mathrm{vis}} = -5.69$, the best achievable across visualizations.

value of information: $\Delta = \mathrm{R}_V^R - \mathrm{R}_\varnothing = 2.27$.

Suppose the experimenter sets the conversion rule $f(r) = \$1 + \$0.01r$ from score r to real dollars as follows: an agent gains a fixed \$1 for completing each trial, plus a \$0.01 in real dollars for each point earned in scoring rule space. The conversion rule is set such that an agent is guaranteed to obtain a positive payment. We calculate the expected real payments to a rational agent in Table 4. If the goal is to incentivize an agent to consult the visualization, we would conclude that the incentive is badly designed because it is a very small fraction of the amount expected without looking at the visualizations (<3%).

$f(\mathrm{R}_\varnothing)$	$f(\mathrm{R}_V)$	Δ_f	$\Delta_f / f(\mathrm{R}_\varnothing)$
\$0.920	\$0.943	\$0.023	2.5%

Table 4: $f(\mathrm{R}_\varnothing)$ shows the expected payment to a rational agent without the visualization, $f(\mathrm{R}_V)$ shows the expected payment to a rational agent who reads the visualization, while $\Delta_f = f(\mathrm{R}_V) - f(\mathrm{R}_\varnothing)$ is the incentive to consult the visualization.

The information loss can also be calculated pre-experimentally.

information loss CI, gradient, and HOPs: $(\mathrm{R}_V^R - \mathrm{R}_V)/\Delta = 0$. Mean: $(\mathrm{R}_V^R - \mathrm{R}_V^{\mathrm{mean}})/\Delta = 100\%$.

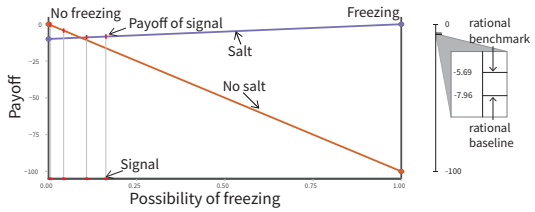

Fig. 2: Score $S(a,p)$ as a function of belief $p \in [0,1]$ as probability of freezing.

From this pre-experiment analysis, the experimenter should also expect the mean visualization to behave badly in payoff compared to the interval, because the mean has a information loss of 100%, i.e. it is not informative for the decision task.

3.4 The Behavioral Agent and Performance Analysis

The behavioral agent faces the same task as the rational agent upon seeing a visualization and choosing an action a from an action space A. Once the experiment has been conducted the collected data implies an empirical joint distribution $\pi^B \in \Delta(A \times \Theta)$ over the behavioral actions and the states.

Experimenters can estimate the following measures to quantify behavioral performance:

behavioral score: The behavioral score is the expected score of the behavioral agent.

$$\mathrm{B} = \mathbf{E}_{(a,\theta)\sim\pi^B}[S(a,\theta)]. \tag{10}$$

behavioral value of information: The behavioral value of information is the difference between the behavioral score and the rational baseline (if non-negative).

$$\Delta^B = \max(\mathrm{B} - \mathrm{R}_\varnothing, 0).$$

The behavioral score B is always below the rational benchmark R_V and can be either above or below the rational baseline $\mathrm{R}_\varnothing$. Importantly, if the behavioral score is below the rational baseline, then from the scores alone we cannot reject the hypothesis that the behavioral agent got no useful information from the visualization. Even with no information, the rational agent performs better. On the other hand, if the behavioral score exceeds the rational baseline, then the behavioral agent systematically performs better than the rational agent with no information and, therefore, must be getting some useful information from the visualization.

To understand how much useful information the behavioral agent is able to get from the visualization, we consider the ratio of the value of information to the behavioral value of information, i.e., $\Delta^B/\Delta \in [0,1]$. If this ratio is large, i.e., close to one, then there is little room to improve the amount of effective communication of the visualization for the decision problem. If this ratio is small, then there is theoretically an opportunity to improve communication.

3.5 Calibrated Behavior and Fine-grained Analysis

The source of behavioral errors can be identified by observing that the joint distribution of behavior and state may contain information that the agent was not able to appropriately act on. In other words, the correlation between behavior and state captures information that is not necessarily reflected by the payoff. The agent's behavior may not be calibrated. The agent's behavior is calibrated if action $a \in A$ is the optimal action on the conditional distribution over states when that action a was taken. The following calibrated behavioral score is always between the rational baseline and the rational benchmark:

calibrated behavioral score The calibrated behavioral score is the score of a rational agent on information structure π^B.

$$\mathrm{R}_B = \mathbf{E}_{(a,\theta)\sim\pi^B}[\hat{S}(\pi^B(\theta|a),\theta)]. \tag{11}$$

The calibrated behavioral agent performance allows for different behavioral errors to be distinguished, and the information conveyed by the visualization to be assessed even when the behavioral score is

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

below the rational baseline. We identify two sources of loss for the behavioral agent:

belief loss The belief loss captures the loss in score as a result of the agent not responding with different beliefs after looking at visualizations of informationally distinct stimuli (e.g., different proportions, probabilities, etc.). We measure the belief loss by calibrating the behavioral decisions and responses. The difference $(\mathrm{R}_V - \mathrm{R}_B)/\Delta$ quantifies the magnitude to which the agent is not able to differentiate between stimuli.

optimization loss Upon viewing a visualization the rational agent would update their beliefs and then choose the optimal action under those beliefs. The optimization loss captures the loss from the agent not properly updating their beliefs about the uncertain state and making the optimal decision given their beliefs. The difference $(\mathrm{R}_B - \mathrm{B})/\Delta$ quantifies the magnitude to which the agent is unable to use the information they have obtained.

3.6 Applying the Framework to Visualization Studies

3.6.1 Scope: What is a decision experiment?

The rational agent framework can be applied widely across empirical visualization studies. To apply the framework the experiment task needs to involve the visualization of states that can take on multiple values and under which the rational agent's optimal decision – for payoff or accuracy – is non-identical. In such experiments, the rational benchmark and the rational baseline are distinct and there is a non-trivial value of information.

It is worth noting that our use of the term "decision" aligns with statistical decision theory, and may conflict with colloquial interpretations promoted elsewhere in visualization research. For example, we could apply the framework to perception studies (like Cleveland and McGill's well-known position-length experiment [3]) and refer to the task participants face as a decision task. The uncertainty in the state comes from the fact that there is a distribution over ground truth proportions that are used to generate stimuli.

There are just two conditions that prevent applying the rational agent framework. The first is in studies where there is no differing state. For example, if the exact same data are presented to all participants in a single-trial between-subjects manipulation of visualization design then there is no uncertainty about the state and the rational benchmark and baseline would coincide. The second is in studies for which the experimenter considers it impossible to define a ground-truth response against which to evaluate participants' reports, such as studies that query agents' emotional states (e.g., angry, excited, sad) after showing a visualization. For such studies, optimal reports by a rational agent are not well defined.

In decision experiments, scoring rules are typically used to incentivize the behavioral agent to make good decisions and to evaluate the quality of the decision made, such as the accuracy of a prediction. The experimenter may use the same scoring rule for both incentives and accuracy; or the experimenter may not incentivize the behavioral agent at all. For example, it is not clear if participants in the position-length experiment [3] were compensated more for doing the tasks well, but mid-mean absolute error is used to evaluate their responses. The rational agent framework applied to either scoring rules for incentives or accuracy can help understand how effectively information is conveyed by a visualization; the framework's application to scoring rules for incentives can additionally help understand the potential effectiveness of the incentives.

For any decision task, we can distinguish between the decision–the reported "action"–and the beliefs that led to that decision. However, when a decision is defined on a coarse action space, such as binary, calibration will be of limited use, because multiple different beliefs will lead to the same decision so the decision is not informative about the agent's belief. Recall that the optimization loss is the difference $\mathrm{R}_B - \mathrm{B}$ between the calibrated score and the raw score. When the calibrated score is not informative about the agent's optimal payoff as dictated by belief, the experimenter does not estimate the optimization loss precisely. Hence, an experimenter could potentially better quantify the usefulness of the visualization by refining the action space or asking for beliefs directly, i.e., with the action space $A = \Delta(\Theta)$, the set of distributions over states.

3.6.2 $\mathrm{R}_\varnothing$ as a simple baseline

The rational baseline $\mathrm{R}_\varnothing$ captures what a rational agent would do in the experiment if they didn't look at the visualizations. This concept is novel in visualization research, where attempts to detect reliance on visualizations remain relatively rare. Instead, observed performance is usually compared only to the best possible performance for the task, as in computing perceptual or decision accuracy.

We can compare $\mathrm{R}_\varnothing$ to different notions of a simple baseline that an experimenter might use to simulate a behavioral agent not paying attention. For example, a researcher might consider random response over the allowable values for the measure (e.g., randomly choosing a value between 0 and 100 for a task that elicits an integer-valued probability) as a useful simple baseline, or designing a study specifically to compare observed behavior to expectations under a heuristic (e.g. Kale et al. [19]). There is nothing wrong with using other simple baselines to estimate bad performance. However, the unique value of $\mathrm{R}_\varnothing$ as a definitive benchmark is for separating cases where participants got information from the visualization from cases where they did not. If we use other forms of "random guessing" as the baseline, agents could still not look at the visualization at all and do better than the random baseline, so long as random guessing performs worse in expectation than using the prior. Only observing that agents did better than the prior lets us evaluate a "null hypothesis" that they did not consult the visualization.

The fact that the prior is not provided to participants in many visualization experiments does not affect its value for evaluating the state of evidence on whether agents consulted the visualization. In some cases, even when a prior is not provided, $\mathrm{R}_\varnothing$ may still be a realistic expectation of how participants who are not carefully consulting the visualization would respond. For example, when the experiment involves repeated measures (trials) and agents receive feedback, with enough trials we might expect behavioral agents to achieve the expected payoff $\mathrm{R}_\varnothing$ by learning that some fixed action guarantees an okay payoff without looking at the visualization. Research into learning from samples (e.g. Gonzalez and Dutt [8]) can inform speculation about particular repeated feedback experiment designs.

3.6.3 Calculating behavioral scores

R_V, the rational agent's payoff under the action dictated by their posterior beliefs, represents the best attainable performance by a behavioral agent who does the experiment. Whenever the goal of the experiment is to compare the performance of visualization strategies that differ in the information they provide for the task, R_V and Δ can be calculated for each visualization condition tested. Different visualization optimal R_V for informationally-inequivalent visualizations give us a sense of how much the results of the experiment can be driven purely by information differences. In general, researchers who are interested in understanding differences that result from visual design choices, rather than informational differences, should aim for equivalent visualization optimal R_V. Exceptions include cases where the goal is to investigate how visualization approaches compare for a real-world inspired task where a conventional representation may not be richly informative, such as situations where point estimates are preferred by convention [11]. Whenever informationally-inequivalent visualizations are compared, the experimenter can use the information loss $(\mathrm{R}_V^R - \mathrm{R}_V)/\Delta$ to study the maximum differences we expect under optimal use of the two visualizations. [1]

[1] Additionally, we can use comparisons between information loss for informationally-different visualizations to weed out claims a researcher makes about one visualization being informationally superior than another: A larger effect than the difference in the two R_V that is claimed to result from informationally-inequality must be an overestimate. More generally, any experiment that presents estimates corresponding to a higher expected score under the scoring rule for a given visualization must be presenting an overestimate confounded, for example, by sampling error [2].

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

Payoff-relevant state	• $\theta_0 \in \{0,1\}$ = lose/win w/o. a new player • $\theta_1 \in \{0,1\}$ = lose/win w. a new player
Data generating model	• fine-grained state (x_0, x_1), where $x_0 \sim N(100, \sigma^2)$ = score w/o. a new player $x_1 \sim N(\mu, \sigma^2)$ = score w. a new player • win: score higher than 100, $\theta_i = \mathbf{1}[x_i \geq 100]$ $\Pr[\theta_i = 1] = \Pr[x_i \geq 100]$ • $\Pr[\theta_0 = 1] = 50\%$ • $\Pr[\theta_1 = 1]$ uniformly drawn from $\{p_1, \ldots, p_8\}$
Signal (visualization)	$v \in V$ visualizing x_0, x_1 e.g. CI, HOPs, densities, QDPs
Agent's action	$a \in \{0 = \text{not hiring}, 1 = \text{hiring}\}$
Scoring rule (payoff)	$S(a, \theta)$

Table 5: Kale et al. [19] decision problem under our framework.

Generally, we employ estimates of joint behavior of the agent with the state, $\pi^B \in \Delta(A \times \Theta)$, from a statistical model that accounts for the design of the experiment. This is because rarely can the results of an experiment be interpreted without accounting for confounding induced by the design in the form of order effects, random effects of participants or other factors, etc. The target in producing model estimates of π^B is to achieve a good prediction of the score distribution expected for behavioral agents if the experiment were to be repeated many times on a new sample from the same population. In general, *generative* statistical models that model the joint probability distribution $p(x,y)$ and use Bayes rule to compute $p(y|x)$ are preferable. For example, in our demonstrations below, we use Bayesian regression models. However, our approach is compatible with sampling from observed results directly or using non-generative models (e.g., Frequentist regression), as long as push-forward transformations to the outcome space can be simulated using fitted model parameter estimates. Regardless of the specific modeling approach, experimenters should keep in mind that the value of the rational agent framework for gaining insight into a design or set of results depends on how well the behavioral scores predict expected performance in that experiment. Scores produced by a modeling approach that overfits to the particular observed behavior in the experiment (e.g., overfit to the particular combination of participants as shown in the example by [32]) will produce overfit benchmarks.

4 Demonstrations

We apply the rational agent framework to two visualization experiments.Both experiments won awards for their rigorous design at the conferences at which they were published, making them a conservative choice for demonstrating the interpretive value added by the framework.

4.1 Effect size judgments and decisions [19]

Kale et al. [19] use an online crowdsourced experiment to investigate the extent to which visualization design impacts people's use of heuristics based on the central tendency in judging effect size [4], a measure of the "signal" in a distributional comparison relative to the noise.

4.1.1 Experiment design

Kale et al.'s mixed design experiment compares judgments and decisions across four approaches to visualizing a pair of distributions: quantile dotplots (QDPs) [21], hypothetical outcome plots [14], 95% containment intervals, and density plots, assigned between subjects. Each participant does trials where the means are visually annotated and where they are not. The distributions are framed as predicted scores in a fantasy sports game for a team with and without a new player. Participants are tasked with using the visualizations for a binary decision task: whether to pay to add the new player to their team, knowing that doing so increases their chance of winning a monetary award but costs money. Additionally, on each trial an unincentivized probability of superiority (PoS) judgment is elicited, representing the participant's belief about the probability that a random draw from the score distribution with the new player will be greater than one from the distribution without. This allows us to calculate belief and optimization loss for both a belief and a decision question.

Scoring rule Table 5 summarizes the decision problem under our framework. The action space is $A = \{0,1\}$ for the participant or equivalently $A = \{\text{not hire}, \text{hire}\}$. There are two fine-grained random states, one x_0 indicating the score without a new player, and the other one x_1 indicating the score with a new player. The agent wins a game if the realized score is above 100, i.e. $\theta_i = \mathbf{1}[x_i \geq 100]$. The payoff function is defined by

$$S(a,\theta) = \begin{cases} 0 & \text{if } a=0, \theta_0 = 0 \quad \textit{lose without hiring} \\ 3.17 & \text{if } a=0, \theta_0 = 1 \quad \textit{win without hiring} \\ -1 & \text{if } a=1, \theta_1 = 0 \quad \textit{lose with new player} \\ 2.17 & \text{if } a=1, \theta_1 = 1 \quad \textit{win with new player} \end{cases}$$

where the unit is millions of dollars in the simulated account. The simulated accounts are initialized with 108M dollars. At the end of the experiment, the agents are rewarded \$0.8 per 1M more than 150M in their simulated accounts.

Stimuli generation and optimal decision strategy The probability $\Pr[\theta_0 = 1]$ of winning without a new player is fixed at 50%. The experiment varies the probability $\Pr[\theta_1 = 1]$ of winning with a new player at 8 levels above 50%, corresponding to 8 ground truth PoS sampled in log space from 0.55 to 0.95. The score x_0 and x_1 follow a Gaussian distribution with identical standard deviations of either 5 or 15. x_0 has a mean fixed at 100; the target PoS for each trial is realized by varying the mean of x_1. Each block of trials the participant completes presents these eight levels twice, once with the lower standard deviation and once with the higher standard deviation.

The realized score in the fictional sports game (used to determine the participant's payoff for a trial) is simulated using Monte Carlo method. The agent faces a decision problem of hiring the new player or not, where his expected utility is as follows:

$$\begin{cases} 3.17 \cdot \Pr[x_0 \geq 100] & \text{if he does not hire;} \\ 2.17 \cdot \Pr[x_1 \geq 100] + (-1) \cdot \Pr[x_1 < 100] & \text{if he hires.} \end{cases}$$

When the rational agent believes that $3.17 \cdot \Pr[x_1 \geq 100] \leq 2.17 \cdot \Pr[x_1 \geq 100] + (-1) \cdot \Pr[x_1 < 100]$, or equivalently that $\Pr[x_1 \geq 100] \geq 81.5\%$, her optimal decision is to choose to hire a new player and vice versa.

As mentioned above, on each trial behavioral agents are asked for an unincentivized PoS judgment $\Pr[x_1 \geq x_0]$. Under the choice to fix the mean of x_0 at 100, the PoS judgment maps to a unique probability of winning with a new player, thus mapping to a unique optimal decision. As a result, the PoS judgment represents beliefs associated with the incentivized decision.

Rational Agent On any given trial, the agent is presented with a probability $\Pr[x_1 \geq 100]$ of winning with a new player, randomly drawn from the 8 predetermined levels $p_1, p_2, \ldots, p_8$. Without getting any additional information (i.e., seeing any visualizations), the rational agent has prior belief $\Pr[x_1 \geq 100] = \frac{1}{8}\sum_{i=1}^{8} p_i = 80.5\%$, so the optimal decision is always not to hire a priori.

The rational agent knows the distributions of scores shown in the visualization follow Gaussian distributions which are parameterized by mean and variance. Different visualization strategies have the same value to the rational agent, regardless of whether means are added or not[2]. Hence, any visualization in the experiment is equivalent for the rational agent to show the probability of the team winning with the new player. After seeing the visualization, the rational agent knows $\Pr[x_1 \geq 100] = p_i$ for some i, and makes the optimal decision.

Dotted lines in Figure 3 show the rational baseline ($R_\varnothing$, left) and rational benchmark (R_V, right).

4.1.2 Pre-experimental Analysis

We calculate the rational agent baseline and benchmark for a single decision task, in simulated account dollars in millions.

[2]A rational agent will spend infinite time looking at HOPs, to fully understand the distribution of scores.

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

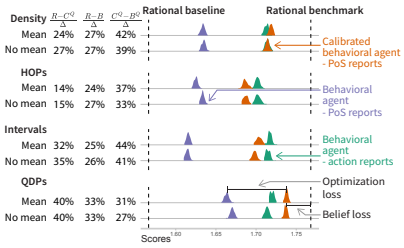

Fig. 3: Estimated payoffs under the scoring rule used in Kale et al. [19] for 100 simulated experiments in which behavioral agents make decisions (**behavioral decision score** B, green) and report PoS judgments (**PoS raw score**, purple, and adjusted **calibrated PoS score**, orange) by visualization condition with means added and without. The rational agent benchmark R_V and the rational agent baseline $R_\varnothing$ are shown as dotted lines.

Rational baseline: $R_\varnothing = 1.57$. The rational agent achieves $R_\varnothing$ by selecting any fixed action, or arbitrarily randomizing over the actions.

Rational benchmark / visualization optimal: $R_V = 1.77$ for all visualization formats.

Value of information: $\Delta = R_V - R_\varnothing = 0.20$.

The information loss is 0 for all visualization strategies.

When we translate these scores through the conversion rate to real dollars received by the participant ($f(r) = \$1 + \max\{0, \$0.08(r - 150\text{M})\}$ for each 1M over 150M in the account where r is in millions), we get the total incentive that an agent has to consult the visualization, shown in Table 6[3]. This incentive seems reasonable for encouraging agents to consult the visualization, as it is nearly a third of the guaranteed payment from choosing any fixed action.

$f(R_\varnothing)$	$f(R_V)$	Δ_f	$\Delta_f / f(R_\varnothing)$
\$1.66	\$2.17	\$0.51	30.72%

Table 6: $f(R_\varnothing)$ shows the expected payment to a rational agent, $f(R_V)$ shows the expected payment to a rational agent who reads the visualization, while $\Delta_f = f(R_V) - f(R_\varnothing)$ is the incentive to consult the visualization.

One point worth acknowledging is that Kale et al. do not provide participants with the prior, as is frequently true in visualization experiments. This is not necessarily a flaw in the design. In this example, there are reasons why we would expect behavioral agents to achieve scores higher than $R_\varnothing$ in the experiment design despite not explicitly being given the prior. For this example, the prior score can be obtained by taking the same action in any trial or arbitrarily randomizing over actions. Additionally, participants were given feedback, and a participant who was randomizing but watching feedback is arguably in a position to approximately learn the prior over the course of the experiment.

4.1.3 Post-experimental Analysis

The original results presented by Kale et al. [19] include a consistent but very small impact of annotating means on bias in PoS judgments, and some disparity between what visualizations appear to perform best for PoS judgments versus incentivized decisions: QDPs perform relatively well across the two tasks, but performance with intervals and densities varies across tasks. The authors advise visualization researchers to be cautious in assuming that perceptual accuracy feeds directly into decision-making, because a user's internal sense of effect size is not necessarily identical when they use the same information for different tasks. The authors speculate that the decoupling of performance may result from users relying on different heuristics to judge the same data for different purposes. (e.g., Kahneman and Tversky's [17] suggestion of a distinction between perceiving an event's probability and weighting the probability in decision-making), or from not incentivizing the PoS question. By applying the rational agent framework post-experimentally, we further investigate their results and this ambiguity.

In our post-experimental analysis, we first empirically estimate the expected payoff B for decisions. Because the study hypothesis in Kale et al. concerned the comparison between performance with means annotated versus not annotated, we calculate the expected **behavioral score for the decision task** for each of the four visualization strategies crossed with the means manipulation, resulting in eight total scores with uncertainty (Figure 3, **green**). We report scores from 100 simulated experiments, with binary decisions sampled from the posterior predictive distribution of the Bayesian logistic regression model in Kale et al. [19]. These scores indicate that the behavioral agents' decisions achieved a payoff higher than the rational agent with prior and fairly close to the rational agent with posterior, which we further analyze below. A detailed discussion of our estimation is deferred to **??**.

Kale et al. [19] elicit responses on a finer space $Q = \Delta(\{0,1\})$ - the PoS reports, which is more informative than their decision task in that each PoS corresponds to a unique belief on the winning probability. We apply our framework by calculating the scores from PoS reports. To calculate expected **behavioral scores B^Q for the PoS task**, we simulate decisions by applying the optimal decision rule to reported PoS, however this time we sampled from the posterior predictive distribution of the authors' linear-in-log-odds model for PoS judgments (Figure 3, **purple**). Scores for the PoS task are closer to the prior than those for the decision task. Similar to Kale et al.'s results, for both the decision task and PoS task we see only a slight difference in expected behavioral scores with and without the addition of means.

Finally, we calculate the calibrated behavioral scores. The calibrated scores for decisions are the same as the expected payoff B; recall this is because for a binary decision where the behavioral score is above $R_\varnothing$, calibration cannot improve the score. We follow the same approach to calibrate PoS reports and calculate the **calibrated behavioral scores C^Q for the PoS task** by discretizing the PoS report space (Figure 3, **orange**). We discretize the space into intervals of length 0.02 so that we can calculate the empirical Bayesian posterior of state θ_1 without overfitting.[4]

Belief Loss Recall that belief loss measures the extent to which a behavioral agent can distinguish between stimuli by consulting the visualization, and is quantified by taking the difference between the rational benchmark and the calibrated behavioral responses, $R_V - R_B$, and normalizing by Δ. Because calibrating the decision scores does not improve upon the behavioral scores for Kale et al's decision task, belief loss is equivalent to $\frac{(R_V - B)}{\Delta}$ in Figure 3.

We next consider belief loss for the PoS task as $\frac{R_V - C^Q}{\Delta}$ in Figure 3. QDPs induce the least belief loss and HOPs the most. This may be because agents will often not watch the HOPs animation for long, and hence are lossy information processors compared to the rational agent [19]. The ranking we observe across visualization conditions resembles that observed in the Just-Noticeable-Difference (JND) estimates in Kale et al.'s model of participants' decisions. JNDs measure how sensitive behavioral agents are to the evidence in making decisions.

Optimization Loss Recall that optimization loss is calculated as $\frac{(R_B - B)}{\Delta}$. This loss is 0 for the decision task because expected scores were above $R_\varnothing$. When we evaluate optimization loss for the PoS task, we observe fairly substantial gaps between the behavioral and calibrated behavioral scores (**purple** and **orange** distributions). The normalized optimization loss is shown as $\frac{C^Q - B^Q}{\Delta}$ in Figure 3. These scores indicate 1) that the behavioral agents are struggling to report their beliefs but getting information from the visualizations, and 2) the behavioral agents are getting a fair amount of information from the visualizations: the calibrated scores are obtaining a relatively high percentage of the rational benchmark.

When we look at decision scores, and compare them to calibrated PoS, we see that the behavioral agents are making nearly optimal

[3]With high probability, the simulated payoff falls over 150M. f can be considered linear here, so we write the expected real payment as $f(R_V)$.

[4]Note that discretization induces an unavoidable discretization error to the estimation of calibrated score.

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

decisions given the information they have (to hire the new player or not). This is because we can expect the PoS reports to capture the agents' perceived probability of winning with the new player (due to the one-to-one mapping between PoS and probability of win by design). This suggests agents are understanding the experiment task fairly well.

The fact that behavioral scores for the PoS report are considerably improved by calibrating indicates that agents struggled to use the information they had obtained to report their beliefs. Kale et al. acknowledge that they cannot disambiguate the reason for the disparity in the PoS versus decision results they observe, and speculate it may stem from the PoS question not being incentivized or from a difference between probability perception and weighting [17]. However, our comparison between expected scores for the binary decision task and the PoS task suggests that agents *were* consulting the visualizations and extracting much of the information.

Alternative reasons agents may have struggled with reporting for the PoS question is that while Kale et al.'s design cleanly maps PoS to probability of winning with the new player, the latter is the more directly relevant information to the decision at hand. PoS is also harder to read from the visualizations that the participants were provided relative to the probability of winning. Our analysis calls into question the possible explanations proffered in the paper for explaining differences observed in how visualizations perform between PoS and decision tasks. Had the experiment asked a directly payoff-related question like *What is the improvement in the probability of winning by hiring a new player?* the comparison the work makes between beliefs and decisions may have been more informative for assessing conjectures like Kahneman and Tversky's notion of differences in probability perception and weighting [17].

4.2 Transit decisions [6]

Fernandes et al. [6] compare approaches to presenting bus arrival time predictions–including textual descriptions of one-sided probability intervals, containment intervals, QDPs, CDFs, density plots, density plots with intervals, and only a point estimate (no uncertainty control)–for making transit decisions about when to leave for the bus stop.

4.2.1 Experiment Design

Payoff-relevant state	$\theta \in [0,30]$ bus arrival time
Data generating model	θ from Box-Cox t distribution
Signal (visualization)	$v \in V$ visualizing θ
Agent's action	$a \in [0,30]$ time to go to bus stop
Scoring rule (payoff)	$S(a,\theta)$

Table 7: Decision problem for Fernandes et al. [6]

Fernandes et al.'s mixed design experiment compares incentivized decisions across twelve visualization strategies that are assigned between subjects. Each participant is presented with 40 total trials parameterized by bus arrival time distributions. Participants are randomly assigned one of three decision scenarios representing a hypothetical real-world decision with an associated (unique) scoring rule.

The decision problem is summarized in Table 7. The agent takes action from $A = [0,30]$, a time to arrive at the bus stop. The payoff-relevant state is $\theta \in [0,30]$, the time the bus arrives at the bus stop. When $a > \theta$, the agent does not catch the bus. If he misses the bus, he is guaranteed to catch a second bus that arrives at $\theta' + 30$, where θ' follows the same arrival distribution as the first bus. In each of the three decision scenarios, the agent gains a bonus $r_0 > 0$ for each minute of activities before arriving at the bus stop, $r_w < 0$ for each minute waiting at the bus stop, and a bonus $r_d > 0$ for each minute spent at the destination with a maximum time of T spent. The payoff can be formulated as follows:

$$S(a,\theta) = \begin{cases} r_0 a + r_w(\theta - a) + r_d \cdot T & \text{if } a \le \theta \\ \text{catching bus} & \\ r_0 a + r_w(\theta' + 30 - a) + r_d \cdot [T - (\theta' - \theta)] & \text{else} \\ \text{not catching bus} & \end{cases}$$

For each decision scenario, payoffs are generated as in Table 8.

Stimuli generation and optimal decision strategy Each trial corresponds to a Box-Cox t distribution generated from a model of real bus arrival predictions [21]. Fixing a belief distribution p where the arrival time θ is drawn, if the agent chooses action a, his expected payoff is

$$\mathbb{E}_{\theta\sim p}[S(a,\theta)] = \sum_{\theta \le a} \Pr[\theta]\,[r_0 a + r_w(\theta - a) + r_d \cdot T] + \sum_{\theta > a} \Pr[\theta]\,\left[r_0 a + r_w(\mathbb{E}_{\theta'\sim p}[\theta'] + 30 - a) + r_d \cdot [T - (\mathbb{E}_{\theta'\sim p}[\theta'] - \theta)]\right].$$

Scenario ID	r_0	r_w	r_d	T
1	8	-14	14	90
2	14	-14	14	60
3	8	-17	17	120

Table 8: Payoffs of decision tasks for different scenarios.

4.2.2 Pre-experimental Analysis

We calculate the rational agent baseline, visualization optimal, and rational benchmark for a single trial in the unit of simulated coins.

Scenario ID	1	2	3
$R_\varnothing$	1078.7	767.5	1850.2
R_V full information (interval, pdf+interval, QDPs, pdf, cdf, none)	1171.8	852.0	1919.4
R_V text60	1170.3	851.5	1918.7
R_V text85	1171.0	851.6	1918.3
R_V text99	1165.0	848.1	1914.9
R_V^R	1171.8	852.0	1919.4
Δ	93.1	84.6	69.3

Table 9: The **rational baseline** $R_\varnothing$ for different scenarios, the **visualization optimal** R_V for different scenarios and visualization strategies, the **rational benchmark** R_V^R for different scenarios, and the **value of information** $\Delta = R_V^R - R_V$ for different scenarios.

Table 9 summarizes the **rational baseline** $R_\varnothing$, the **visualization optimal** R_V, the **rational benchmark** R_V^R, and the **value of information** $\Delta = R_V^R - R_V$. The visualization strategies are informationally equivalent to the rational agent and equivalent to knowing the bus arrival distribution, except for the text displays. This is because, with the exception of text displays, there is a one-to-one mapping between the distribution visualization on a trial and the bus arrival distribution. Note that this is also true for no uncertainty displays (control). The no uncertainty condition visualization displays the mean of the bus arrival distribution. Each bus arrival distribution in the experiment has a distinct mean, so the rational agent fully knows the bus arrival distribution after seeing the mean. After seeing the visualization, the rational agent knows the bus arrival distribution D, thus is able to make the optimal decision. For the text probability interval displays, however, the rational agent is not able to distinguish between distributions that map to the same text, leading to a lower expected score.

We quantify this information asymmetry by information loss.

Information loss We calculate the information loss in Table 10.

Scenario ID	1	2	3
full information (interval, pdf+interval, QDPs, pdf, cdf, none)	0	0	0
text60	1.6%	0.7%	1.2%
text85	0.9%	0.6%	1.6%
text99	7.3%	4.7%	6.5%

Table 10: The information loss $(R_V^R - R_V)/\Delta$ for different scenarios and visualization conditions.

All types of visualizations have an information loss $\sim 1\%$, except for text99 which induces a small information loss $\sim 7\%$.

We calculate the cumulative incentive for the rational agent (Δ) across 40 trials. In the experiment, each 1000 coins translate into a $\$d$ bonus in real payment, with another \$1.25 as a guaranteed base payment, i.e. the payment conversion rule is $f(r) = \frac{d}{1000}r + \$1.25$. $d = 0.01698, 0.08228, 0.016076$ for scenarios $1,2,3$, respectively. The value of information for a rational agent in real dollars is shown in Table 11. Since the information loss for text displays is small ($\le 7\%$), we omit the payoff calculation for text displays.

Across the three scoring rules, the incentive for the rational agent to consult a visualization is always less than 10% of the guaranteed

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

Scenario ID	$f(\mathrm{R}_\varnothing)$	$f(\mathrm{R}_V)$	Δ_f	$\Delta_f/f(\mathrm{R}_\varnothing)$
1	\$1.983	\$2.046	\$0.063	3.12%
2	\$3.776	\$4.054	\$0.287	7.37%
3	\$2.440	\$2.484	\$0.044	1.82%

Table 11: $f(\mathrm{R}_\varnothing)$ shows the expected payment to a rational agent who takes the optimal fixed action, $f(\mathrm{R}_V)$ shows the expected payment to a rational agent who reads the visualization, while $\Delta_f = f(\mathrm{R}_V) - f(\mathrm{R}_\varnothing)$ is the incentive to consult the visualization.

payment of choosing an optimal fixed action (Table 11). The incentive is not well designed if the goal is to encourage agents to consult the visualizations.

To improve incentives, we suggest subtracting f_0 from all payments, where f_0 is a threshold that any behavioral agent's score is unlikely to fall below. For example, one obvious choice of f_0 is $30 \cdot r_0$, obtained by a strategy to always arrive at the bus stop at 30 minutes.

Additionally, Fernandes et al. [6] conclude from the results of their experiment that with the dot50 visualization, *50% of decisions will be above 95% of optimal, about 80% of decisions will be above 90% of optimal, and more than 95% of decisions will be above 80% of optimal*. However, we find that the baseline is able to achieve a 92.1%, 90.1%, and 96.4% of the optimal for each scenario, respectively, calculated assuming the agent does not look at the visualization. This pre-experimental analysis therefore calls into question how impressive the dot50 performance reported by the original work is, illustrating how without a baseline to compare with, statements based on the proximity of observed behavior to optimal can mislead.

4.2.3 Post-experimental Analysis

In our post-experimental analysis, we empirically estimate the behavioral expected payoff B for the 10 visualization conditions in Fernandes et al. We fit a mixed-effects Bayesian regression model to predict agents' actions (i.e. chosen arrival time) from visualization condition, scenario, and bus arrival distribution.

Our results show that behavioral payoffs are above or close to the rational baseline. We conclude that optimization loss is the main source of loss in decision-making, while belief loss reduces the difference compared to raw behavioral scores. In other words, visualizations appear to allow users to obtain a good proportion of the available information in the visualization. A good visualization helps agents make the optimal decision from the information.

A detailed presentation of our results is deferred to **??** in supplemental materials.

5 Discussion

We contribute rational agent benchmarks for assessing 1) the potential for an experiment to incentivize participants and show differences between visualizations and with best attainable performance, and 2) the sources of error that explain observed results from behavioral agents. As our demonstrations on two celebrated visualization studies show, our framework can be applied to identify improvements in designs and to deepen understanding of results even when the original research was rigorously done. A key feature is that it provides well-defined comparison points for any given visualization, reducing reliance on rough, relative ordering information that is often used to interpret visualization experiment results.

Returning to the questions posed in Section 1, by applying our framework we can expect to answer them as follows:

- How hard is the task? The value of information, the difference between rational baseline and benchmark, captures the "room" for improvement on the task.
- How incentivized are participants? Through pre-experimental analysis, we calculate the expected increase in payment that the participants can get from consulting the visualization.
- To what extent do the differences in performance stem from informational asymmetries? This difference is quantified by the information loss.
- What are the reasons for sub-optimal decisions from behavioral agents? We separate the sources of loss into a) the belief loss, the loss from not perceiving the information, and b) the optimization loss, the loss from not properly using the information.

There are many other practical advantages to the rational framework, which we observed in conducting analyses for our demonstrations. For example, having the ability to compare results from different tasks in score space, as we did for Kale et al. [19], can sidestep the challenges associated with trying to interpret and compare findings between models that estimate different parameters, often under different mathematical transformations that must be inverted to get any perspective on performance from results. Additional benefits will arise on a case-by-case basis, as demonstrated in our examples.

Integrating measures of the value of information into visualization is an important step forward in the pursuit of more rigorous theoretical foundations for visualization-based inference, as van Wijk called for years ago, and researchers continue to call for today [5, 9, 12, 31]. By providing a widely applicable definition of a decision task and associated analyses identifying the value of information, our work makes possible deeper connections between information economics and design with data visualization. There are many exciting extensions to the rational agent framework to be explored in future work. For example, for certain decisions tasks, such as binary decisions which are amenable to complete characterization, it is likely possible to provide more prescriptive guidelines that can point visualization researchers to the right task to study in the first place given a high-level research goal (e.g., evaluate visualization alternatives for election forecasts).

Another direction worth pursuing is to integrate the rational agent benchmarks into the sample size calculations that experimenters use to ensure that an experiment design is capable of assessing performance differences. We might ask, What sample size is needed to resolve performance with a visualization relative to the value of information to the task? Alternatively, scoring rules could be designed to obtain the same value of information with fewer samples, cf. Li et al. [26] It may also be useful to use quantities from the rational agent framework to contextualize target effect sizes (e.g., in units of Δ) or assumed noise from measurement error (e.g., in units of the standard deviation in scores across trials given the data-generating model) in fake data simulation for power analysis.

5.1 Limitations

Conducting pre-experimental analysis is less useful if the experimenter doubts the value of performance incentives (e.g., Mason and Watts [27]). Pre-experimental analysis will also not offer actionable guidelines if the experiment assumes a flat or no reward scheme. Although, choosing to provide no clear incentive to use the visualizations may signal that the experimenter trusts that participants will try their best. In such cases, analyzing the value of information is still well-motivated for making sure a study design provides enough room for seeing performance differences and assessing information gain from a visualization.

The relationship between the rational baseline $\mathrm{R}_\varnothing$ and what a participant would do in the actual experiment if they did not look at the visualizations is nuanced. As we describe above, the purpose of $\mathrm{R}_\varnothing$ is not to predict how randomizing behavioral agents will score, though in some cases it may. The framework is also not intended as a theory of how behavioral agents make decisions: the benchmarks are useful for evaluating the quality of decisions of behavioral agents who act differently from a rational one. While a rational agent would update their beliefs based on the empirical joint distribution over signals and states and then choose the optimal action under those beliefs, no intermediate measurement of beliefs is typically made of the behavioral agent and so his optimization loss cannot be similarly decomposed. In many experiments the behavioral agent is not informed of the prior hence the Bayesian update is not well defined., with the lack of prior information accounted for in the optimization loss.

One of the biggest impediments to applying the framework widely is insufficient reporting of study details in empirical papers. For example, full information about the scoring rule used in a study may not be reported, such as when there are exclusion criteria like performance on an attention check that led to non-payment for a task but not mentioned in the paper.

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] M. Agrawal, J. C. Peterson, and T. L. Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020. doi: 10.1073/pnas.1915841117 2

[2] K. Button, J. Ioannidis, C. Mokrysz, B. Nosek, J. Flint, E. Robinson, and M. Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14, 04 2013. doi: 10.1038/nrn3475 5

[3] 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, 1984. doi: 10.1080/01621459.1984.10478080 5

[4] R. Coe. It's the effect size, stupid. In *British Educational Research Association Annual Conference*, vol. 12, p. 14, 2002. 6

[5] E. Dimara and J. Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2022. doi: 10.1109/TVCG.2021.3114813 2, 9

[6] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718 1, 8, 9

[7] D. Fudenberg, J. M. Kleinberg, A. Liang, and S. Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130:956 – 990, 2022. doi: 10.1086/718371 2

[8] C. Gonzalez and V. Dutt. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, 118(4):523, 2011. doi: 10.1037/a0024558 5

[9] C. Heine. Towards modeling visualization processes as dynamic bayesian networks. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1000–1010, 2020. doi: 10.1109/TVCG.2020.3030395 9

[10] J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021. doi: 10.1038/s41586-021-03659-0 2

[11] J. Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 26(1):130–139, 2019. doi: 10.1109/TVCG.2019.2934287 5

[12] J. Hullman and A. Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3), 2021. doi: 10.1162/99608f92.3ab8a587 2, 9

[13] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889 2

[14] 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, 2015. doi: 10.1371/journal.pone.0142444 3, 6

[15] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 6. ACM, 2008. doi: 10.1145/1377966.1377974 2

[16] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013. doi: 10.1109/TVCG.2013.126 2

[17] D. Kahneman and A. Tversky. *Prospect Theory: An Analysis of Decision Under Risk*, chap. 6, pp. 99–127. [Wiley, Econometric Society], 2013. doi: 10.1142/9789814417358_0006 7, 8

[18] A. Kale, Z. Guo, X. Qiao, J. Heer, and J. Hullman. Evm: Incorporating model checking into exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics (forthcoming)*, 2023. 2

[19] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021. doi: 10.1109/TVCG.2020.3030335 1, 5, 6, 7, 9

[20] A. Kale, Y. Wu, and J. Hullman. Causal support: Modeling causal inferences with visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1150–1160, 2022. doi: 10.1109/TVCG.2021.3114824 2

[21] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 5092–5103. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858558 6, 8

[22] Y.-S. Kim, P. Kayongo, M. Grunde-McLaughlin, and J. Hullman. Bayesian-assisted inference from visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):989–999, 2021. doi: 10.1109/TVCG.2020.3028984 2

[23] C. Kinkeldey, A. M. MacEachren, and J. Schiewe. How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51(4):372–386, 2014. doi: 10.1179/1743277414Y.0000000099 2

[24] D. Knill and R. Whitman. *Perception as Bayesian Inference*, chap. 7, pp. 825–837. MIT Press, 1996. 2

[25] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279 2

[26] Y. Li, J. D. Hartline, L. Shan, and Y. Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, p. 988–989. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3490486.3538338 9

[27] 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, p. 77–85. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1600150.1600175 9

[28] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111 2

[29] S. Savelli and S. Joslyn. The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, 27(4):527–541, 2013. doi: 10.1002/acp.2932 2

[30] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, p. 1–7. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1168149.1168158 2

[31] J. van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pp. 79–86, 2005. doi: 10.1109/VISUAL.2005.1532781 9

[32] T. Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022. doi: 10.1017/S0140525X20001685 6

[33] T. Zuk and S. Carpendale. Theoretical analysis of uncertainty visualizations. In R. F. Erbacher, J. C. Roberts, M. T. Gröhn, and K. Börner, eds., *Visualization and Data Analysis 2006*, vol. 6060, p. 606007. International Society for Optics and Photonics, SPIE, 2006. doi: 10.1117/12.643631 2

A SUPPLEMENTAL MATERIALS

- Supplemental material 1 - appendix to the paper, including:
  - an example of visualization strategies used for the two demonstrations, and
  - a detailed discussion of results for transit decision experiment.
- Supplemental material 2 - source codes and guided walkthrough to the framework, including:
  - code for demonstrations under `./demonstrations/`, and
  - a guided walkthrough of our framework, with application to the hypothetical weather forecasting experiment. See `./guideline/guideline.pdf`.

Authorized licensed use limited to: Northwestern University. Downloaded on August 31,2024 at 17:10:26 UTC from IEEE Xplore. Restrictions apply.