

# The Rational Agent Benchmark for Data Visualization

Yifan Wu<sup>\*</sup>   Ziyang Guo<sup>†</sup>   Michalis Mamakos<sup>‡</sup>   Jason Hartline<sup>§</sup>  
 Jessica Hullman<sup>¶</sup>

July 4, 2023

## 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.

**Keywords:** evaluation, decision-making, rational agent, scoring rules

## 1 Introduction

Writing in 2005, van Wijk famously asked, What is the value of visualization? [van Wijk, 2005]. Nearly twenty years later, the common usage of empirical studies in modern visualization research might seem to answer van Wijk’s query by providing evidence about which visualization best supports a task. Design guidelines driven by intuition are increasingly being replaced with data-driven recommendations based on visualization studies, whether the task is trend estimation trends [Correll and Heer, 2017], telling causation from correlation [Xiong et al., 2020], or using a display to inform decisions in a strategic game [Kayongo et al., 2022].

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 process. To understand how well people performed with a visualization, or how important an observed performance difference is, we must understand what sorts of performance differences

---

<sup>\*</sup>Department of Computer Science, Northwestern University. Email: yifan.wu@u.northwestern.edu.

<sup>†</sup>Department of Computer Science, Northwestern University. Email: ziyanguo2027@u.northwestern.edu.

<sup>‡</sup>Department of Computer Science, Northwestern University.

Email: michailmamakos2022@u.northwestern.edu

<sup>§</sup>Department of Computer Science, Northwestern University. Email: hartline@northwestern.edu

<sup>¶</sup>Department of Computer Science, Northwestern University. Email: jhullman@northwestern.edu

an experimental scenario admits. 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, in addition to conventional design decisions like how many participants to target and how they will 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 to produce certain results. For example, one canonical form of analysis when the goal is to detect an effect such as a difference between strategies is to ensure that the study design provides sufficient statistical power to detect an effect of the hypothesized size. More generally, an ideal approach to interpreting the results of a study comparing visualization strategies would help a reader answer questions like the following to contextualize the results:

- How hard is the task? For example, how well could we expect someone do without consulting the forecast 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 compared (e.g., providing 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?
- To what extent might observed differences in performance be driven by "luck of the draw" in allocating ground truth distributions across visualization conditions?

The above questions highlight how results of visualization research studies often lack clear comparatives, or *benchmarks* that can aid in their design and interpretation. Answering such questions contextualizes what was learned from observing the performance of any single visualization in absolute terms defined on the experiment design. Additionally, a good set of benchmarks is necessary to assess the fitness of the experiment design itself for studying a given visualization research question. Without clear benchmarks, readers and authors alike tend to draw conclusions from coarse, *relative* information like visualization performance rankings.

We contribute a rational agent framework based on quantifying the value of information to a judgment or decision problem to provide 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 [Kale et al.,

2021] and one on the impact of visualization design on transit decisions [Fernandes et al., 2018]. 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.

## **2 Related Work**

### **2.1 Visualization Evaluation**

Our work extends a larger body of work aimed at improving evaluation methods in visualization. Researchers have contributed overviews of qualitative and quantitative approaches [Isenberg et al., 2013, Lam et al., 2012, Zuk and Carpendale, 2006] and conceptual models and approaches for ensuring that one selects an evaluation that is appropriate for a given task, context, or contribution type [Isenberg et al., 2008, Munzner, 2009, Shneiderman and Plaisant, 2006].

When visualizations are meant to support inference in addition to merely describing an observed dataset, as they often are [Hullman and Gelman, 2021], then the evaluation approach should define a standard for assessing the quality of the inference. However, several recent surveys of evaluative studies for visualizations [Dimara and Stasko, 2022] and uncertainty visualizations specifically [Hullman et al., 2019, Kinkeldey et al., 2014] suggest that the use of well-defined decision or belief 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 [Hullman et al., 2019, Kinkeldey et al., 2014]. This has led some researchers to advocate for adopting Bayesian inference as a normative model for conceiving effective reactions to visualizations [Hullman and Gelman, 2021, Kale et al., 2022, Kim et al., 2021]. These models use the deviation of human performance from the normative Bayesian ideal as a means of better understanding human judgment biases, and for inspiring new design approaches. Similarly, ideal observer analysis, used in psychophysics, theoretically upperbounds behavioral performance by a Bayesian agent in the same situation to reason about factors influencing human perception [Knill and Whitman, 1996]. Our framework defines the baseline performance in addition to the upperbound, and hence provides a “scale” for interpreting behavioral performance, including separating sources of loss in decision-making. 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 then the design suggestions they lead to will not be effective. In contrast to prior applications of Bayesian theory to visualization, the value of the rational agent framework for assessing experiment results and improving experiment designs does not depend on actual humans acting like rational agents.

### **2.2 Interpreting experiment results**

Guidance on designing rigorous empirical studies can be found in a wide range of literatures, such as statistics, economics, and psychology, and often comes in the form of design calculations like power analyses that are done before an experiment is run. In experiments that use quantitative measures to compare performance on judgment or decision tasks, as we are concerned with here, a researcher’s ability to detect or estimate the magnitude of an effect (such

as between visualizations) will be offset by sampling error from a limited number of subjects, inherent randomness in stimuli generation (e.g., random sampling from a ground truth distribution to generate visualized samples, random assignment of stimuli to trials) as well as measurement error from imprecise responses (a.k.a., “a shaky hand”). Designing an informative experiment typically requires explicit steps toward controlling expected error rates (e.g., false positive or false negative rates via setting  $\alpha$  and  $\beta$  in a hypothesis testing framework).

To interpret effect estimates and use them to inform real-world decisions like when to deploy an intervention, consumers of experiments must grapple with a number of ambiguities, many of which are not addressed by improving the transparency or form of statistical reporting alone as researchers in visualization have advocated [Dragicevic, 2015, Kay et al., 2016a, in Human–Computer Interaction Working Group, 2019]. Some ambiguities stem from underspecification of sources of variance affecting results. For example, the same average treatment effect can map to very different patterns at the level of individual units [Gelman et al., 2023], and the population from which participants are sampled and to which the researcher hopes to generalize is often underspecified [Gigerenzer and Marewski, 2015], as is the sampling of stimuli [Gigerenzer, 2022, Wells and Windschitl, 1999]. Interpretation is further challenged by ambiguity in the mapping between the experimental situation and target real-world scenario. We should expect experiment results to paint an optimistic picture of the effect under study, because it is in the experimenter’s best interest (and arguably the benefitting population’s best interest) for them to study the range of conditions that they deem most likely to show a hypothesized difference between visualizations.<sup>1</sup>

One way to address some of these ambiguities is by formalizing expectations about an experiment. Our work is related to recent integrative modeling [Hofman et al., 2021] approaches to benchmarking the irreducible variance in data used for modeling [Agrawal et al., 2020, Fudenberg et al., 2022]. For example, the explanatory power of theories embedded in behavioral models can be assessed by quantifying irreducible error inherent in an experimental task [Fudenberg et al., 2022], 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

---

<sup>1</sup>Consider an experimenter who hypothesizes that a new drug will help treat a chronic disease. It would be pointless to test it on healthy people and equally pointless to test it on people who are clearly minutes from death. If the experimenter believes that the drug will have an effect, they will feel obligated to test it on conditions that will show its effect.

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.

### 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  $\Theta$  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*  $S : A \times \Theta \rightarrow \mathbb{R}$  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.,

$$a^* = \arg \max_{a \in A} S(a, p). \quad (1)$$

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

$$\hat{S}(p, \theta) = S(\arg \max_{a \in A} S(a, p), \theta). \quad (2)$$

**Example** We illustrate the framework with a hypothetical weather forecast experiment, loosely inspired by Savelli and Joslyn[Savelli and Joslyn, 2013]. 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  $\theta$  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:

$$S(a, \theta) = \begin{cases} 0 & \text{if } a = 0, \theta = 0 & \text{no salt, not freezing} \\ -100 & \text{if } a = 0, \theta = 1 & \text{no salt, freezing} \\ -10 & \text{if } a = 1, \theta = 0 & \text{salt, not freezing} \\ 0 & \text{if } a = 1, \theta = 1 & \text{salt, freezing} \end{cases} \quad (3)$$

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)	$S : A \times \Theta \rightarrow \mathbb{R}$

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

### 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.

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  $\pi_1, \pi_2, \dots, \pi_k$ , 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) [Hullman et al., 2015] (plotting probability as frequency).

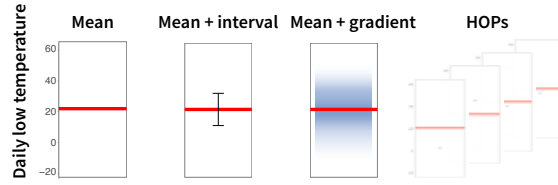


Figure 1: Example visualizations for a hypothetical weather forecast task.

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  $\mu = 5^\circ\text{C}$  and standard deviation  $\sigma$ . The standard deviation  $\sigma$  is uniformly drawn from  $\{2, 3, 4, 5\}$ .

For visualization strategies that depict uncertainty (CI, gradient, HOPs), the signal  $v$  is  $(\mu, \sigma)$ ; for the visualization of the mean, the signal  $v$  is deterministically  $\mu$ .

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  $\theta$  and  $\sigma$  in Section 3.2.

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

	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
$\theta = 0$	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.

Payoff-relevant state	$\theta \in \{0, 1\}$ = {not freezing, freezing}
Data generating model	<ul style="list-style-type: none"> <li>daily low temperature  <math>t \sim N(\mu, \sigma^2)</math>;  <math>\Pr[\theta = 1] = \Pr[t \leq 0]</math>;  <math>\mu = 5</math> fixed;  <math>\sigma</math> uniformly from <math>\{2, 3, 4, 5\}</math>.</li> <li>equivalently,  <math>\Pr[\theta = 1]</math> uniformly from  0.62%, 4.78%, 10.56%, 15.87%.</li> </ul>
Agent’s action	$a \in \{0 = \text{no salt}, 1 = \text{salt}\}$
Signal (visualization)	$v^{\text{vis}} \in V^{\text{vis}}$ , vis = visualization strategies vis $\in \{\text{mean, CI, gradient, HOPs}\}$ of temperature
Scoring rule (payoff)	$S(a, \theta)$ (see eq. (3))

Table 3: Notation for the freezing-salting example.

### 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  $\pi$  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*:

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

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

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

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  $\pi$ , 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.

$$R_{\emptyset} = \mathbf{E}_{\theta \sim p}[\hat{S}(p, \theta)]. \quad (6)$$

**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.

$$R_V = \mathbf{E}_{(v,\theta) \sim \pi}[\hat{S}(\pi(\theta|v), \theta)]. \quad (7)$$

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:

$$\Delta = R_V - R_\emptyset.$$

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

### 3.3.1 Multiple Visualization Strategies

Our framework can be applied to comparing a set of  $k$  different visualization strategies, with information structures  $\pi^1, \dots, \pi^k$ .

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

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

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  $1 \dots k$ , inducing information structures  $\pi^1, \dots, \pi^k$ . The rational benchmark is defined as

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

**value of information** Again, the value of information is defined as the difference between the rational benchmark and rational baseline.

$$\Delta = R_I - R_\emptyset. \quad (10)$$

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, i.e., studies where visualization strategies that provide varying amounts of information about the uncertain state are compared.

**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  $(R_I - R_V)/\Delta$  between the rational agent benchmark (the rational best performance across visualizations) and the visualization optimal for that 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  $p = 0.0796$ . 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 (3). 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  $p = 0.9$  where the decision-maker is indifferent between no-salt and salt actions.

The rational agent framework gives the following quantities:

**rational baseline:**  $R_\emptyset = -7.96$ .

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

**visualization optimal:**  $R_V^{\text{CI}} = R_V^{\text{gradient}} = R_V^{\text{HOPs}} = -5.69$ ;  $R_V^{\text{mean}} = -7.96$ .

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  $R_V = -5.69$ .

For the visualization of the mean, the rational agent has only the prior information and obtains  $R_V^{\text{mean}} = R_\emptyset = -7.96$ .

**rational benchmark:**  $R_I = \max_{\text{vis}} R_V^{\text{vis}} = -5.69$ , the best achievable across visualizations.

**value of information:**  $\Delta = R_I - R_\emptyset = 2.27$ .

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%).

$f(R_\emptyset)$	$f(R_V)$	$\Delta_f$	$\Delta_f/f(R_\emptyset)$
\$0.920	\$0.943	\$0.023	2.5%

Table 4:  $f(R_\emptyset)$  shows the expected payment to a rational agent,  $f(R_V)$  shows the expected payment to a rational agent who reads the visualization, while  $\Delta_f = f(R_V) - f(R_\emptyset)$  is the incentive to consult the visualization.

The information loss can also be calculated pre-experimentally.

**information loss** CI, gradient, and HOPs:  $(R_I - R_V)/\Delta = 0$ . Mean:  $(R_I - R_V^{\text{mean}})/\Delta = 100\%$ .

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

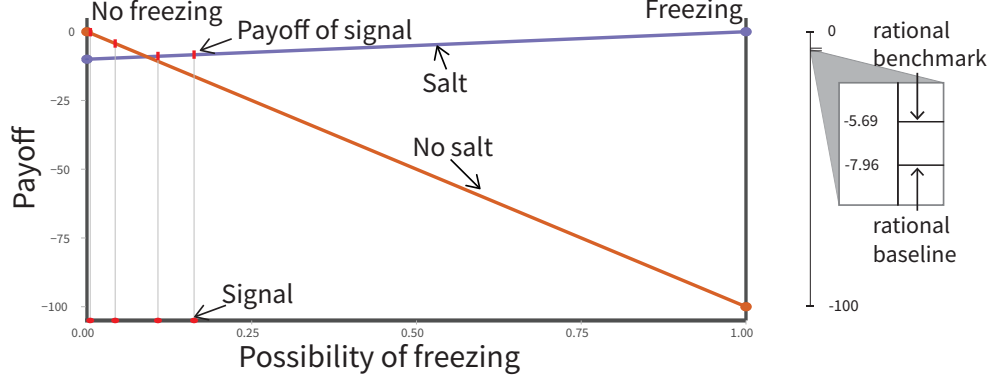


Figure 2: Score  $S(a, p)$  as a function of belief  $p \in [0, 1]$  as probability of freezing.

### 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.

$$B = \mathbb{E}_{(a, \theta) \sim \pi^B}[S(a, \theta)]. \quad (11)$$

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

$$\Delta^B = \max(B - R_\emptyset, 0).$$

The behavioral score  $B$  is always below the rational benchmark  $R_V$  and can be either above or below the rational baseline  $R_\emptyset$ . 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  $\pi^B$ .

$$R_B = \mathbf{E}_{(a, \theta) \sim \pi^B}[\hat{S}(\pi^B(\theta|a), \theta)]. \quad (12)$$

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 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  $(R_V - R_B)/\Delta$  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  $(R_B - B)/\Delta$  quantifies the magnitude to which the agent is unable to use the information they have obtained.

## 3.6 Calculations

Calculating the quantities above is based on

- taking expectations, for example, the behavioral score of equation (11) is

$$B = \sum_{(a, \theta) \in \pi^B} S(a, \theta) \pi^B(a, \theta), \quad (13)$$

- Bayesian updating to obtain posterior beliefs as defined in equation (5), and
- constructing the corresponding proper scoring rule  $\hat{S}$  via equation (2) and the optimal action for a belief as described in equation (1).

For example, calculating the rational benchmark  $R_V$  and the calibrated behavioral score  $R_B$  are the same calculations applied to information structures  $\pi$  and  $\pi^B$ , respectively, with the three subcalculations above.

## 3.7 Applying the Framework to Visualization Studies

### 3.7.1 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 [Cleveland and McGill, 1984]) 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 [Cleveland and McGill, 1984] 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  $R_B - B$  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 distribution over states.

### 3.7.2 $R_\emptyset$ as a simple baseline

$R_\emptyset$  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  $R_\emptyset$  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., 2021]. There is nothing wrong with using other simple baselines to estimate bad performance. However, the unique value of  $R_\emptyset$  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  $R_{\emptyset}$  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,  $R_{\emptyset}$  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  $R_{\emptyset}$  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, 2011] can inform speculation about particular repeated feedback experiment designs.

### 3.7.3 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  $\Delta$  can be calculated for each visualization condition tested. Different  $R_V$  and  $\Delta$  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  $\Delta$ . 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 [Hullman, 2019]. Whenever informationally-inequivalent visualizations are compared, the experimenter can use the differences in attainable performance to contextualize behavioral results, because differences in these benchmarks capture the maximum differences we expect under optimal use of the two visualizations.<sup>2</sup>

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  $\pi^B$  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.

---

<sup>2</sup>Additionally, we can use comparisons between  $R_V$  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 [Button et al., 2013].

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 [Yarkoni, 2022]) will produce overfit benchmarks.

## 4 Demonstrations

We apply the rational agent framework to two visualization experiments.<sup>3</sup> 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 [Kale et al., 2021]

Kale et al. [Kale et al., 2021] 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 [Coe, 2002], 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) [Kay et al., 2016b], hypothetical outcome plots [Hullman et al., 2015], 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** Section 4.1.1 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 random states, one  $X_0$  indicating the score without a new player, and the other one  $X_1$

---

<sup>3</sup>See “demonstrations/effect\_size/analysis.Rmd” and “demonstrations/transit\_decisions/analysis.Rmd” in our supplementary material for the complete analysis. Our supplement is available at [https://github.com/Guoziyang27/rational\\_framework](https://github.com/Guoziyang27/rational_framework)

Payoff-relevant state	<ul style="list-style-type: none"> <li>• <math>\theta_0 \in \{0, 1\}</math> = lose/win w/o. a new player</li> <li>• <math>\theta_1 \in \{0, 1\}</math> = lose/win w. a new player</li> </ul>
Data generating model	<ul style="list-style-type: none"> <li>• <math>X_0 \sim N(100, \sigma^2)</math> = score w/o. a new player</li> <li>• <math>X_1 \sim N(\mu, \sigma^2)</math> = score w. a new player</li> <li>• win: score higher than 100 <math>\Pr[\theta_i = 1] = \Pr[X_i \geq 100]</math></li> <li>• <math>\Pr[\theta_0 = 1] = 50\%</math></li> <li>• <math>\Pr[\theta_1 = 1]</math> uniformly drawn from <math>\{p_1, \dots, p_8\}</math></li> </ul>
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.[Kale et al., 2021] decision problem under our framework.

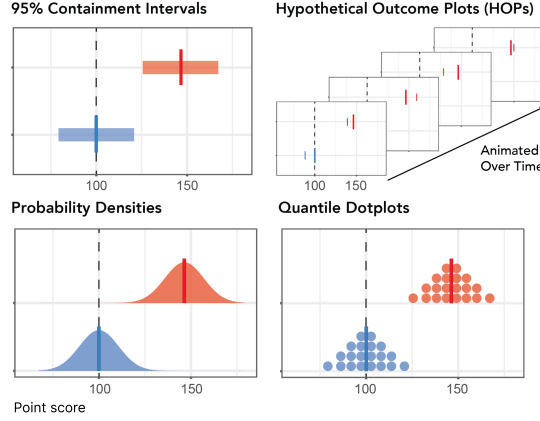


Figure 3: Stimuli from Kale et al. [Kale et al., 2021].

indicating the score with a new player. The agent wins a game if the realized score is above 100. The payoff function is defined by

$$S(a, X) = \begin{cases} 0 & \text{if } a = 0, X_0 < 100 & \text{lose without hiring} \\ 3.17 & \text{if } a = 0, X_0 \geq 100 & \text{win without hiring} \\ -1 & \text{if } a = 1, X_1 < 100 & \text{lose with new player} \\ 2.17 & \text{if } a = 1, X_1 \geq 100 & \text{win with new player} \end{cases} \quad (14)$$

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  $\Pr[X_0 \geq 100]$  of winning without a new player is fixed at 50%. The experiment varies the probability  $\Pr[X_1 \geq 100]$  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. Both  $X_0$  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:

$$3.17 \cdot \Pr[X_0 \geq 100] \quad \text{if he does not hire;} \quad (15)$$

$$2.17 \cdot \Pr[X_1 \geq 100] + (-1) \cdot \Pr[X_1 < 100] \quad \text{if he hires.} \quad (16)$$

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  $p_1, p_2, \dots, p_8$ . 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<sup>4</sup>. 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 4 show the rational baseline ( $R_\emptyset$ , 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.

**Rational baseline:**  $R_\emptyset = 1.57$ . The rational agent achieves  $R_\emptyset$  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:**  $\Delta = R_V - R_\emptyset = 0.20$ .

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 ( $f(r) = \$1 + \max\{0, \$0.08(r - 150M)\}$  for each 1M over 150M in the account

<sup>4</sup>A rational agent will spend infinite time looking at HOPs, to fully understand the distribution of scores.



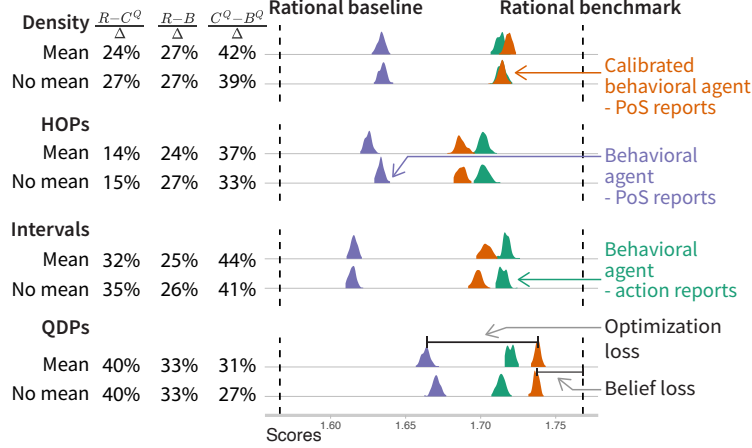


Figure 4: Estimated payoffs under the scoring rule used in Kale et al. [Kale et al., 2021] for 100 simulated experiments in which behavioral agents make decisions (**behavioral decision score B**, green) and report PoS judgments (**PoS raw score**, purple, and adjusted **calibrated PoS score**, orange) by visualization condition with means added and without. The rational agent benchmark  $R_V$  and the rational agent baseline  $R_\emptyset$  are shown as dotted lines.

where  $r$  is in millions), we get the total incentive that an agent has to consult the visualization, shown in Table 6<sup>5</sup>. 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.

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

Table 6:  $f(R_\emptyset)$  shows the expected payment to a rational agent,  $f(R_V)$  shows the expected payment to a rational agent who reads the visualization, while  $\Delta_f = f(R_V) - f(R_\emptyset)$  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  $R_\emptyset$  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. [Kale et al., 2021] 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

<sup>5</sup>With high probability, the simulated payoff falls over 150M.  $f$  can be considered linear here, so we write the expected real payment as  $f(R_V)$ .

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 [Kahneman and Tversky, 2013] 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 4, **green**).

Specifically, we calculate these scores by simulating binary decisions for the intended number of agents per combination of visualization approach and means manipulation (of eight) in the original experiment (160 people per visualization approach, each of which completed a block of 16 task trials with and without means).<sup>6</sup> For each condition we repeatedly sampled  $n = 160 \times 16$  simulated responses from the posterior predictive distribution of the Bayesian logistic regression model used by Kale et al. [Kale et al., 2021], balancing trial numbers and block orders according to the original experiment design. We report scores obtained from simulating results 100 times (Figure 4, **green**). 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.

Kale et al. [Kale et al., 2021] 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 4, **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  $R_\emptyset$ , 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 4, **orange**). We discretize the space into intervals of length 0.02 so that we can calculate the empirical Bayesian posterior of state  $\theta_1$  without overfitting.<sup>7</sup>

**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  $\Delta$ . Because calibrating the decision scores does not improve upon the

<sup>6</sup>In reality, less than 160 were achieved for some conditions in the original experiment. Replicating the missing data structure instead of using the intended cell count does not change our results.

<sup>7</sup>Note that discretization induces an unavoidable discretization error to the estimation of calibrated score.

behavioral scores for Kale et al.’s decision task, belief loss is equivalent to  $\frac{(R_V - B)}{\Delta}$  in Figure 4.

We next consider belief loss for the PoS task as  $\frac{R_V - C^Q}{\Delta}$  in Figure 4. 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 [Kale et al., 2021]. 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  $R_\emptyset$ . 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 4. 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 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 [Kahneman and Tversky, 2013]. 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 [Kahneman and Tversky, 2013].

## 4.2 Transit decisions [Fernandes et al., 2018]

Fernandes et al. [Fernandes et al., 2018] compare different approaches to presenting bus arrival time predictions—including textual descriptions of one-sided probability intervals, containment

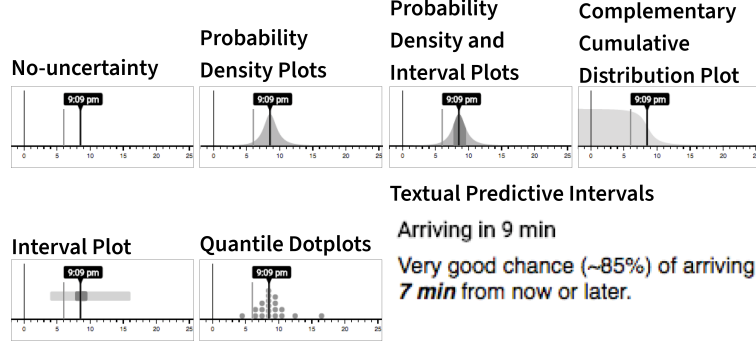


Figure 5: Stimuli from Fernandes et al. [Fernandes et al., 2018]

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	$\theta$ from Box-Cox $t$ distribution
Signal (visualization)	$v \in V$ visualizing $\theta$
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.[Fernandes et al., 2018]

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  $\theta'$  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:

$$S(a, \theta) = \begin{cases} r_0 a + r_w(\theta - a) + r_d \cdot T & \text{if } a \leq \theta \\ \text{catching bus} \\ r_0 a + r_w(\theta' + 30 - a) + r_d \cdot [T - (\theta' - \theta)] & \text{else} \\ \text{not catching bus} \end{cases} \quad (17)$$

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 [Kay et al., 2016b].

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.

Fixing a belief distribution  $p$  where the arrival time  $\theta$  is drawn, if the agent chooses action  $a$ , his expected payoff is

$$\begin{aligned} \mathbb{E}_{\theta \sim p}[S(a, \theta)] &= \sum_{\theta \leq a} \Pr[\theta] [r_0 a + r_w(\theta - a) + r_d \cdot T] \\ &+ \sum_{\theta > a} \Pr[\theta] [r_0 a + r_w(\mathbb{E}_{\theta' \sim p}[\theta'] + 30 - a) + r_d \cdot [T - (\mathbb{E}_{\theta' \sim p}[\theta'] - \theta)]] . \end{aligned} \quad (18)$$

**Rational Agent** The visualizations 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.

#### 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
$R_\emptyset$	1078.7	767.5	1850.2

Table 9: The rational baseline  $R_\emptyset$  for different scenarios.

**Rational baseline:** Table 9 summarizes the baseline  $R_\emptyset$ . The rational agent achieves  $R_\emptyset$  by selecting a fixed action.

**Visualization optimal:** Table 10 summarizes the visualization optimal  $R_V$ .

Scenario ID	1	2	3
$R_V$ 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

Table 10: The visualization optimal  $R_V$  for different scenarios and visualization conditions.

Scenario ID	1	2	3
$R_I$	1171.8	852.0	1919.4

Table 11: The rational benchmark  $R_I$  for different scenarios.

**Rational benchmark:** By taking maximum over visualization optimal, the rational benchmark is the rational agent with full information in Table 11.

**Value of information:** Table 12 summarizes the value of information  $\Delta = R_I - R_\emptyset$ .

Scenario ID	1	2	3
$\Delta$	93.1	84.6	69.3

Table 12: The value of information  $\Delta$  for different scenarios.

From these calculations, we first note that all visualization conditions have the same visualization optimal, except for the text displays. We quantify this information asymmetry by information loss.

**Information loss** We calculate the information loss induced in Table 13.

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 13: The information loss  $(R_I - R_V)/\Delta$  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 ( $\Delta$ ) 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 14. Since the information loss for text displays is small ( $\leq 7\%$ ), 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 payment of choosing an optimal fixed action (Table 14). 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. [Fernandes et al., 2018] 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.

Scenario ID	$f(R_\emptyset)$	$f(R_V)$	$\Delta_f$	$\Delta_f/f(R_\emptyset)$
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 14:  $f(R_\emptyset)$  shows the expected payment to a rational agent who takes the optimal fixed action,  $f(R_V)$  shows the expected payment to a rational agent who reads the visualization, while  $\Delta_f = f(R_V) - f(R_\emptyset)$  is the incentive to consult the visualization.

### 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. The authors fit a mixed-effects Bayesian regression model to predict the ratio *expected/optimal payoff* from visualization condition and trial number, with random effects of scenario and participant. Because the outcome ratio is an input to the model, predictions from this model cannot be used to predict expected behavioral scores under different scenarios. We therefore fit our own model to predict agents’ actions (i.e. chosen arrival time) from visualization condition, scenario, and bus arrival distribution. We include random intercepts by participant and random slopes to allow varying effects of trial number by participant. Full model details and model checks we performed to validate the model are available in supplemental material. We use predictions from this model in conjunction with the stated scoring rules in Fernandes et al. to calculate expected scores by scenario.<sup>8</sup> Because Fernandes et al. did not describe a target distribution of participants over visualization conditions, scenarios, and arrival time distributions, we estimate the behavioral scores by sampling arrival time decisions from our model for the same number of agents they analyzed data from per combination of scenario, visualization condition, and bus arrival distribution. We report scores from 100 simulated experiments and report the distributions of **behavioral scores** (Figure 6, **purple**). For each simulated experiment, we calculate the **calibrated behavioral scores**  $R_B$  (Figure 6, **orange**). In our simulations, we round predicted arrival decisions from our behavioral model to integers to match the format of responses used by behavioral agents in the original experiment.

Figure 6 shows that behavioral payoffs are above or close to the baseline. Specifically, they are above  $R_\emptyset$  for Scenario 1 and 2, and above  $R_\emptyset$  for Scenario 3 with the exception of the interval display which induces a payoff below but close to  $R_\emptyset$ .

The original paper evaluated visualization conditions in several ways: by plotting estimated learning effects by visualization condition and by ranking visualization conditions by estimated means and standard deviations of the ratio of expected to optimal payoff for the last trial participants completed. All analyses aggregated results across scenarios despite their varying scoring rules. Specifically, ranking visualizations by estimated mean ratio for the last trial resulted in dot50 as the best performing condition, followed by cdf, dot20, text99, text60, pdf-interval, pdf, interval, no uncertainty, and text85. Ranking visualizations by estimated standard deviation of the last trial resulted in similar rankings, with the first portion of the list matching the previous ranking (dot50, cdf, dot20, text99, text60, pdf-interval) but with no uncertainty performing better than pdf and interval in addition to text85. These rankings lead to the authors’ conclusion that dot50 and cdf are the top performing visualizations.

<sup>8</sup>Even with access to an extended repository containing more complete materials than the public version for the original study, we were not able to exactly reproduce the expected payoffs analyzed by Fernandes et al. However, the expected payoffs our method produces are within 100 simulated coins of their expected payoffs across scenarios.

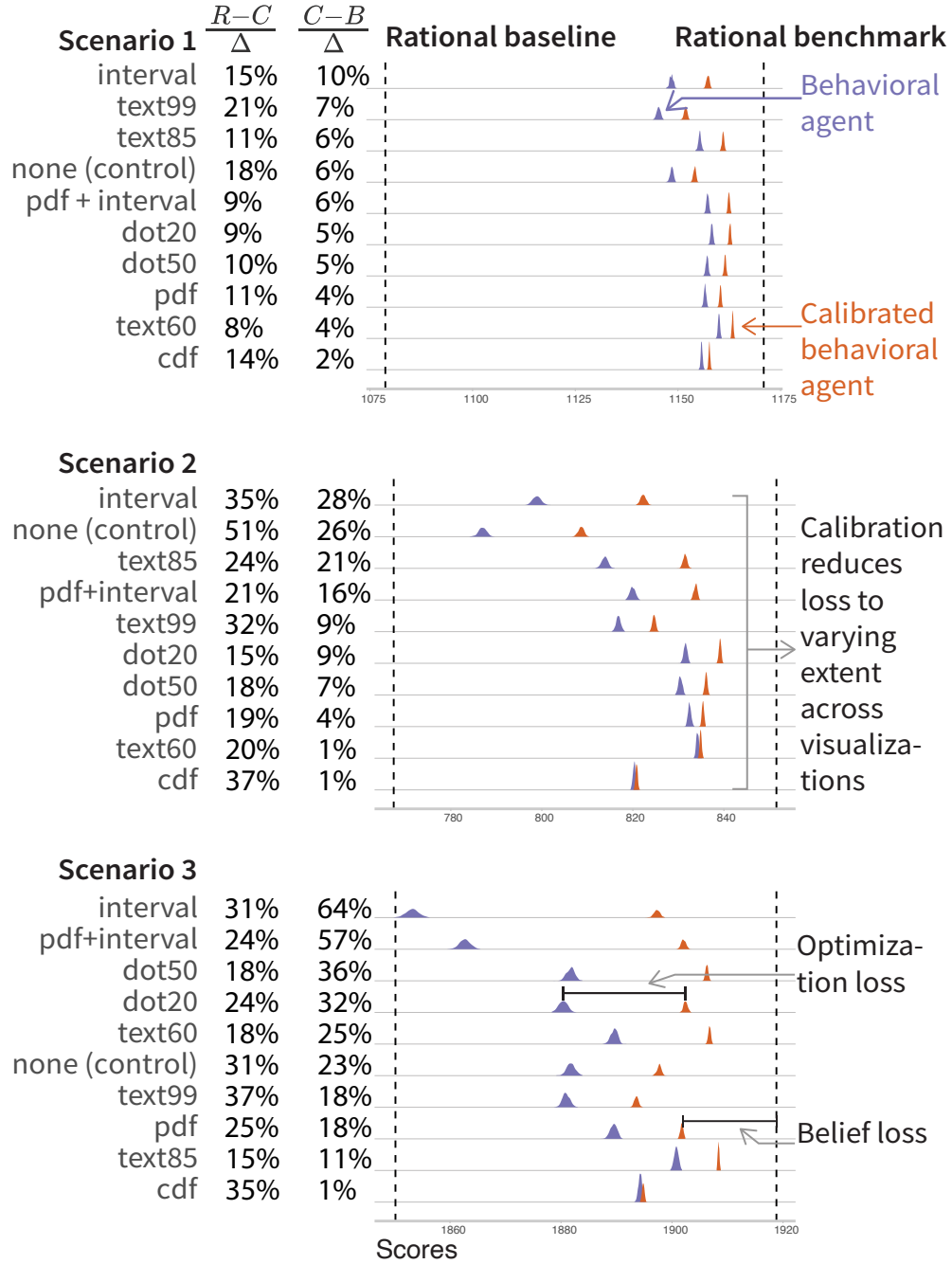


Figure 6: Estimated scores (in simulated coins) for each combination of visualization condition and scenario. Visualizations are ordered by optimization loss for each scenario. The rational agent benchmark  $R_V$  and baseline  $R_\emptyset$  are shown as dotted lines, **behavioral scores**  $B$  in purple, and **calibrated behavioral scores**  $R_B$  in orange.



In contrast, our analysis of behavioral scores shown in Figure 6 purple represents expected score over all trials by visualization condition separately by scenario. From these results, dot50 and cdf are not clearly better performing than multiple other visualization conditions (i.e., they are not furthest right in the plot). Ranking by expected behavioral score by scenario leads to text representations as top performing, with text60 ranking best for Scenarios 1 and 2 and text85 for Scenario 3. cdf is ranked sixth, fifth and second while dot50 is ranked fourth, fourth, and eighth for Scenario 1, 2, and 3, respectively. These differences compared to the original results may be partially attributable to the different modeling approach (our scores consider expected performance across all trials, not just the last trial) or to slightly differences in our computation of expected ratio compared to theirs, as we were not able to perfectly reproduce their model inputs from the available codebase despite using the equations they provided. Our ranking of conditions is clearly inconsistent to those of the original paper when it comes to the performance of dot50, which according to Fernandes et al.’s results performed consistently better in expected ratio across the earlier trials as well, with dot50 users starting and ending with higher estimated ratios than any other condition.

**Belief Loss** The differences between the calibrated score payoff  $R_B$  (orange) and  $R_V$  (right-most dotted line) show that in general, comparing visualizations by belief loss reduces differences between them compared to raw behavioral scores (purple), and that Scenario 1 leads to less belief loss than Scenarios 2 and 3. If anything, ranking visualizations by belief loss suggests that dot20 performs consistently well (ranked second in all Scenarios). In other words, these visualizations appear to allow users to obtain a good proportion of the available information in the visualization, even if they do not necessarily make the optimal decision from the information.

Visualizations convey over 80% and 61% of the information to the agents for scenarios 1,3, respectively, and over 65% of the information for scenario 2, with the exception of the no uncertainty control under scenario 2, which conveys 47% of the information ( $100\% - \text{belief loss } \frac{R_V - R_B}{\Delta}$  in Figure 6). We conclude that all visualization strategies provide reasonable support for detecting changes in the bus arrival time distributions. Belief loss is not the main source of loss in decision-making.

**Optimization Loss** The differences between the calibrated payoff  $R_B$  (orange) and behavioral payoff  $B$  (purple) suggest that optimization loss is a large source of loss in participants’ decision-making. Figure 6 sorts visualization conditions in decreasing order of optimization loss. We see that interval users have the hardest time optimizing their decisions, while cdf and pdf users are able to do so consistently well (cdf achieving first rank, pdf third rank across Scenarios 1, 2, and 3). Users of text60 optimize very well except for in Scenario 3, where their ranking falls from first to sixth.

## 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
  - the belief loss, the loss from not perceiving the information, and
  - the optimization loss, the loss from not properly use the information.
- To what extent are observed differences driven by “luck of draw”? Our Bayesian framework compares the expected payoff over the experiment design, avoiding the effect of random lucky draws.

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. [Kale et al., 2021], 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 [Dimara and Stasko, 2022, Heine, 2020, Hullman and Gelman, 2021, van Wijk, 2005]. 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. [Li et al., 2022] It may also be useful to use quantities from the rational agent framework to contextualize target effect sizes (e.g., in units of  $\Delta$ ) 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

Applying the rational agent framework to pre-experiment analysis is not as useful if the experimenter doubts the value of performance incentives, as some have for certain types of behavioral research like crowdsourced experiments (e.g., [Mason and Watts, 2009]). Pre-experiment analysis will not offer actionable guidelines if the experimenter has already predetermined they will provide a flat or no reward scheme. At the same time, choosing to provide no clear incentive to use visualizations in an experiment is usually a signal that the experimenter trusts that their 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 differences between visualization types and assessing the information gain from any visualization.

The relationship between the rational baseline  $R_\emptyset$  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  $R_\emptyset$  is not to predict how randomizing behavioral agents will score, though in some cases it may.

The rational agent framework is not intended as a theory of how behavioral agents make decisions. Instead, the benchmarks that the framework provides are valuable in evaluating the quality of decisions of behavioral agents who act differently from a rational one. While a rational agent would solve such a problem by updating 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 made of the behavioral agent and so his optimization loss cannot be similarly decomposed. In many experiments, in fact, the behavioral agent is not informed of the prior and, therefore, the Bayesian update is not well defined. This lack of prior information is also accounted for in the optimization loss.

One of the biggest impediments to applying the framework is not a lack of generalizability but a potential lack of transparent 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. This makes it difficult to analyze the experiment using the rule that the original research used.

## 6 Conclusion

We contribute a widely applicable analytical framework for benchmarking visualization performance. The approach uses the performance achievable by a rational agent doing the same visualization experiment as a comparison point for the estimated performance of behavioral agents. The framework distinguishes sources of error in results, like not being able to get the information versus not being able to choose the optimal decision given the information one has obtained. Applying the framework to two awarded visualization studies shows how it can identify ways to improve even rigorous decision experiment designs, and enhance the knowledge gained from observed behavioral performance.

## Acknowledgment

We are grateful to Steve Haroz for helpful comments and suggestions that improved this paper.

## References

- J.J. van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86, 2005. doi: 10.1109/VISUAL.2005.1532781.
- Michael Correll and Jeffrey Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1387–1396, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025922. URL <https://doi.org/10.1145/3025453.3025922>.
- Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2020. doi: 10.1109/TVCG.2019.2934399.
- Paula Kayongo, Glenn Sun, Jason Hartline, and Jessica Hullman. Visualization equilibrium. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):465–474, 2022. doi: 10.1109/TVCG.2021.3114842.
- Alex Kale, Matthew Kay, and Jessica 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.
- Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew 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, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173718. URL <https://doi.org/10.1145/3173574.3173718>.
- Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten 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.
- Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh 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.
- Torre Zuk and Sheelagh Carpendale. Theoretical analysis of uncertainty visualizations. In Robert F. Erbacher, Jonathan C. Roberts, Matti T. Gröhn, and Katy Börner, editors, *Visualization and Data Analysis 2006*, volume 6060, page 606007. International Society for Optics and Photonics, SPIE, 2006. doi: 10.1117/12.643631. URL <https://doi.org/10.1117/12.643631>.

- Petra Isenberg, Torre Zuk, Christopher Collins, and Sheelagh Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaluation methods for Information Visualization*, page 6. ACM, 2008. doi: 10.1145/1377966.1377974.
- Tamara 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.
- Ben Shneiderman and Catherine 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, page 1–7, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595935622. doi: 10.1145/1168149.1168158. URL <https://doi.org/10.1145/1168149.1168158>.
- Jessica Hullman and Andrew Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3), 2021. doi: 10.1162/99608f92.3ab8a587.
- Evanthia Dimara and John 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.
- Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew 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.
- Christoph Kinkeldey, Alan M. MacEachren, and Jochen 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. URL <https://doi.org/10.1179/1743277414Y.0000000099>.
- Alex Kale, Yifan Wu, and Jessica 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.
- Yea-Seul Kim, Paula Kayongo, Madeleine Grunde-McLaughlin, and Jessica 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.
- David Knill and Richards Whitman. *Perception as Bayesian Inference*, chapter 7, pages 825–837. MIT Press, 1996.
- Pierre Dragicevic. *HCI Statistics without p-values*. PhD thesis, Inria, 2015.
- Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4521–4532, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858465. URL <https://doi.org/10.1145/2858036.2858465>.

- Transparent Statistics in Human–Computer Interaction Working Group. Transparent Statistics Guidelines, Jun 2019. (Available at <https://transparentstats.github.io/guidelines>).
- Andrew Gelman, Jessica Hullman, and Lauren Kennedy. Causal quartets: Different ways to attain the same average treatment effect. *arXiv preprint arXiv:2302.12878*, 2023.
- Gerd Gigerenzer and Julian N Marewski. Surrogate science: The idol of a universal method for scientific inference. *Journal of management*, 41(2):421–440, 2015. doi: 10.1177/0149206314547522.
- Gerd Gigerenzer. We need to think more about how we conduct research. *Behavioral and Brain Sciences*, 45, 2022. doi: 10.1017/S0140525X21000327.
- Gary L. Wells and Paul D. Windschitl. Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25(9):1115–1125, 1999. doi: 10.1177/01461672992512005. URL <https://doi.org/10.1177/01461672992512005>.
- Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021. doi: 10.1038/s41586-021-03659-0.
- Mayank Agrawal, Joshua C. Peterson, and Thomas 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. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1915841117>.
- Drew Fudenberg, Jon M. Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130:956–990, 2022. doi: 10.1086/718371.
- Sonia Savelli and Susan 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: <https://doi.org/10.1002/acp.2932>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2932>.
- Jessica Hullman, Paul Resnick, and Eytan 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.
- William S. Cleveland and Robert 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. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10478080>.
- Cleotilde Gonzalez and Varun Dutt. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, 118(4):523, 2011. doi: 10.1037/a0024558.
- Jessica 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.

- Katherine Button, John Ioannidis, Claire Mokrysz, Brian Nosek, Jonathan Flint, Emma Robinson, and Marcus Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14, 04 2013. doi: 10.1038/nrn3475.
- Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022. doi: 10.1017/S0140525X20001685.
- Robert Coe. It’s the effect size, stupid. In *British Educational Research Association Annual Conference*, volume 12, page 14, 2002.
- Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean 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, page 5092–5103, New York, NY, USA, 2016b. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858558. URL <https://doi.org/10.1145/2858036.2858558>.
- Daniel Kahneman and Amos Tversky. *Prospect Theory: An Analysis of Decision Under Risk*, chapter 6, pages 99–127. [Wiley, Econometric Society], 2013. doi: 10.1142/9789814417358\_0006. URL [https://www.worldscientific.com/doi/abs/10.1142/9789814417358\\_0006](https://www.worldscientific.com/doi/abs/10.1142/9789814417358_0006).
- Christian 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.
- Yingkai Li, Jason D. Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC ’22, page 988–989, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538338. URL <https://doi.org/10.1145/3490486.3538338>.
- Winter Mason and Duncan J. Watts. Financial incentives and the ”performance of crowds”. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP ’09, page 77–85, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605586724. doi: 10.1145/1600150.1600175. URL <https://doi.org/10.1145/1600150.1600175>.