Do You See What You Mean? Using Predictive Visualizations to Reduce Optimism in Duration Estimates

Morgane Koval, Yvonne Jansen

Do You See What You Mean? Using Predictive Visualizations to Reduce Optimism in Duration Estimates

Morgane Koval
morgane.koval@inria.fr
Sorbonne Université, CNRS, ISIR
Paris, France

Yvonne Jansen
yvonne.jansen@cnrs.fr
Sorbonne Université, CNRS, ISIR
Paris, France

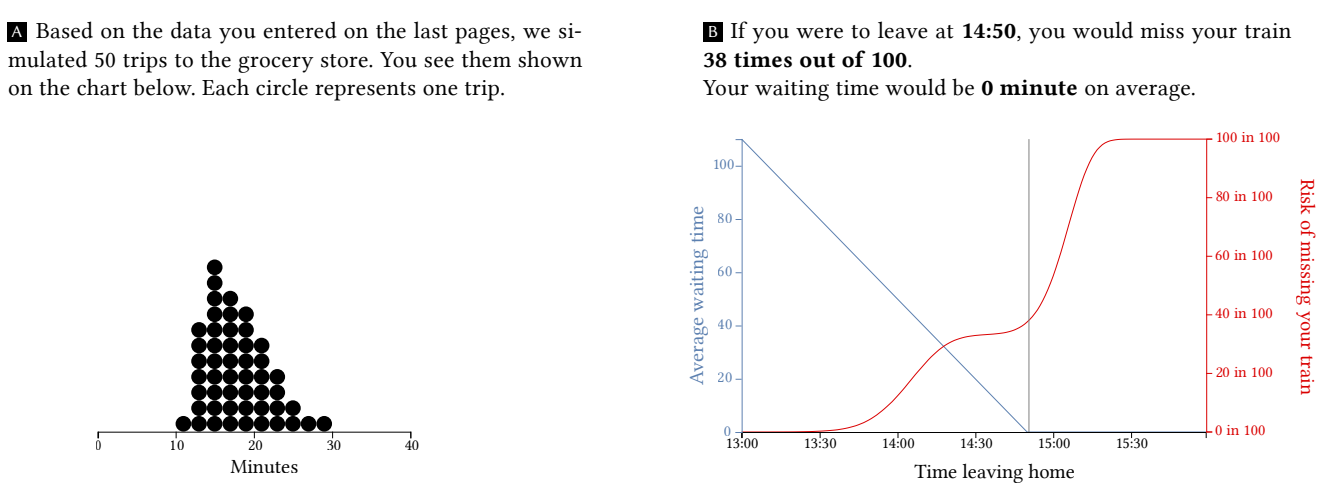


Figure 1: A 50-dot quantile plot (A) and an interactive linechart (B), as used in the two parts of our experiment. (A) is a static visualization showing likely task durations based on task properties provided by participants. (B) is an interactive visualization to support a decision-making exercise on when to leave from home to do the task from (A) and then catch a train. Moving the cursor across the x-axis in (B) updates the text displayed above the visualization.

ABSTRACT

Making time estimates, such as how long a given task might take, frequently leads to inaccurate predictions because of an optimistic bias. Previous attempts to alleviate this bias, including decomposing the task into smaller components and listing potential surprises, have not shown any major improvement. This article builds on the premise that these procedures may have failed because they involve compound probabilities and mixture distributions which are difficult to compute in one's head. We hypothesize that *predictive visualizations* of such distributions would facilitate the estimation of task durations. We conducted a crowdsourced study in which 145 participants provided different estimates of overall and sub-task durations and we used these to generate predictive visualizations of the resulting mixture distributions. We compared participants' initial estimates with their updated ones and found compelling evidence that predictive visualizations encourage less optimistic estimates.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

CCS CONCEPTS

• **Human-centered computing → Empirical studies in visualization**.

KEYWORDS

Planning fallacy, predictive visualization

ACM Reference Format:
Morgane Koval and Yvonne Jansen. 2022. Do You See What You Mean? Using Predictive Visualizations to Reduce Optimism in Duration Estimates. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3502010

1 INTRODUCTION

Predicting how long a given task might take usually yields optimistic estimates [7]. This effect is called the *planning fallacy* and can be defined more precisely as a phenomenon in which *under*estimation of future tasks' duration is almost systematic and not lessened by the knowledge of how long past tasks have taken to be completed. In other words, even informed time predictions present signs of irrationality.

The building of the Sydney Opera House is an often-cited example of high-stakes situations where even the prospect of dramatic

economic consequences did not prevent the project from prompting over-optimistic predictions. Initially expected to be finished in 1963 for 7 million dollars, a smaller-than-planned structure was finally presented ten years later, with almost 100 million dollars of additional costs. Yet, professional settings or group contexts are not the only situations in which time estimates matter. One might need to estimate at what time to leave to catch a train, or how long retrieving a parcel at the post office could take to pick up the children at school on time. Essentially, inaccurate duration estimates can impact both private and public lives, corporations or individuals.

The planning fallacy has been studied for decades, its causes still being under investigation to this day, making it even more intricate to devise appropriate debiasing methods. Various theories have been proposed, most notably the distributional versus singular information theory [42], and the memory bias theory [66]. The first one opposes general, aggregated, or population-based information, such as data about the *overall* punctuality of a train line, to situational, particular, or very precise clues, such as when one *specific* train left its last station. The second theory states that the *memory* of previous task durations is already erroneous. These theories underlie the debiasing methods suggested in the literature, but the difficulty in untangling the causes of the planning fallacy reflects on the difficulty in coming up with universally effective strategies. Common in past attempts to design suitable procedures is, however, that they only relied on oral or written materials. In this paper, we introduce a procedure engaging visualizations to study how these can lead to more considered time estimates. We argue it facilitates the estimation process: mentally envisioning a task is demanding and would be alleviated by concrete representations of one's beliefs on the task's proceedings. In order to gather the data needed to compute such visualizations, we integrate in our approach the task decomposition [8, 9, 17] and surprise listing [8, 37] methods.

We conducted a crowdsourced study involving 145 participants to investigate the effect of simulating and visualizing estimates of task durations. Our experiment was divided into two parts: the first one required estimating a specific task's duration and the second one making a decision. For both parts, we asked participants for their estimates and decisions before and after applying debiasing methods and showing visualizations. We used quantile dot plots [16, 44] as *predictive visualizations* (*i.e.* uncertainty visualizations showing predictions of plausible future event outcomes) to represent mixture distributions of participants' beliefs in the first part of the study, and compared interactive feedback in text form with visualization in the second one. Comparing initial against final estimates, results suggest the procedure induced a change in beliefs expressed by a reduced sense of optimism and an increased sense of uncertainty. Furthermore, the decision-making process was enhanced by both visual aids and textual additions, which suggests estimating on one's own is difficult and prone to inaccuracy.

As our three main contributions, we expose (1) a new debiasing approach combining classic methods and visualization, (2) study results providing compelling evidence, based on a pre-registered analysis and a reproducible experimental protocol, and (3) prospects for future research on time-related bias alleviation through the use of predictive visualization.

2 BACKGROUND

The planning fallacy is a cognitive bias whose definition highlights its resistance to consider prior knowledge, making its alleviation a challenging objective. Everyone is inclined to incorrectly predict the duration of personal or professional tasks, potentially impacting morale, other tasks' completion or even other people's lives.

This section reviews plausible causes for this bias, the debiasing methods which were suggested in line with those theories, and visualization techniques related to time and planning issues.

2.1 Underlying Causes of Inaccurate Estimates

The pioneer contributors to the knowledge we currently have on this bias are Kahneman and Tversky. They demonstrated how intuition could harm accuracy in estimation exercises [75] and formulated the distributional versus singular information theory [42] (or inside/outside account) which is often cited as motivation or inspiration for debiasing techniques. Distributional data exploits information on a whole population while singular data only involves one person's information. The latter one is described as yielding less accurate estimates and is thus the least desired. This theory implies the planning fallacy issue is caused by estimators not relying enough on distributional information. This could be due to personality traits, which would justify differences among individuals. Yet, many such factors have been studied, including dispositional optimism and the propensity to procrastinate [6], anxiety [45, 46], gender [33, 35, 40, 47, 56] and expertise [37], and none has revealed either a consistent effect, or one that could then be replicated. Moreover, the very definition of the planning fallacy expresses optimistic predictions or estimates. Thereby, it may be induced by a dismissal of pessimistic scenarios: people often ignore pertinent pieces of information on similar events as irrelevant. This hypothesis was formally tested, and the results suggest predicting for oneself leads to the disregard of negative past experiences [7]. Furthermore, *temporal discounting* is a significant component in matters related to time management: events that are too far in the future are overlooked along with their potential benefits [50, 74]. Additionally, even when people rely on former experiences, those memories might already be subject to erroneous time perception, leading to distorted recollections of how long previous tasks have taken. Biased memory was stressed as a potential element that could rationalize why most procedures have been giving unsatisfactory results [66].

The distributional/singular theory instructed many of the debiasing methods reported in psychology papers, yet it is not the only plausible explanation, for other theories have been proposed. The difficulty in establishing causes that are acknowledged by the whole scientific community undoubtedly shapes the difficulty in coming up with a truly effective debiasing method.

2.2 Debiasing Methods for the Planning Fallacy

Among the multiple debiasing procedures for task duration estimates reported in the literature, we identified four categories: methods that either propose (a) focusing on the proceedings of the task, (b) decomposing the task, (c) changing the estimation's instructions, or (d) reviewing past elements. Table 1 presents a brief overview of these categories with some relevant studies' findings.

Table 1: Brief overview of outcomes from experiments on debiasing methods. For a more nuanced and detailed description of the findings, refer to the relevant papers mentioned in the last column.

Strategy	Outcomes	References
Focusing on the task's proceedings		
Alternate scenarios	No effect / Generating pessimistic scenarios is not an effective strategy	Newby-Clark et al., 2000 [59]
	No significant difference	Byram, 1997 [8]
	Reduction in over-tightness of estimates	Connolly and Dean, 1997 [9]
Listing surprises	No significant difference	Byram, 1997 [8]
	No improvement for experts, a small one for intermediate users and novices	Hinds, 1999 [37]
Decomposing the task		
Decomposing the task	Still optimistic but more accurate predictions / Other factors such as the elicitation procedure might influence the efficacy of the approach	Connolly and Dean, 1997 [9]
	Sub-components sum up to a greater value than whole task's estimation	Forsyth and Burt, 2008 [17]
	No significant difference	Byram, 1997 [8]
Unpacking	Longer, less biased estimates / The efficiency of unpacking depends on the task's complexity	Kruger and Evans, 2004 [51]
	The typicality and where a sub-task is situated in the whole task influence the unpacking method's efficiency	Hadjichristidis et al., 2014 [29]
Altering the instructions		
Individual probabilities vs. aggregate frequencies	No improvement in accuracy	Griffin and Buehler, 1999 [26]
Sentence formulation	Small tasks turn into underestimates and longer ones into overestimates (*magnitude bias*)	Halkjelsvik et al, 2011 [30]
	'Why' instead of 'how' formulations prompt smaller estimates for simple tasks, and longer ones for complex tasks	Siddiqui et al., 2014 [69]
	Pushing the estimators to focus on speed or duration influences the direction of bias	Løhre and Teigen, 2014 [55]
	The perceived amount of time allocated to a task impacts the bias' effect	Sanna et al., 2005 [68]
	Tendency to overestimate with alternate format	Henry, 1994 [33], Henry and Sniezek, 1993 [34]
Reviewing the past		
Observers vs. actors	Pessimistic estimates, no improvement in accuracy / Observers do not underestimate others' performance time	Buehler et al., 1994 [7]
	Observers consider pessimistic outcomes when predicting for others	Newby-Clark et al, 2000 [59]
	No significant difference	Byram, 1997 [8]
Remembering the past	Mixed (both optimistic and pessimistic responses), no improvement in the accuracy / People use singular rather than distributional information	Buehler et al., 1994 [7]
	No effects on experts	Hinds, 1999 [37]
Feedback	Experienced runners with feedback on a daily basis are better at estimating and predicting performance durations	Tobin and Grondin, 2015 [73]
	Better estimates with accurate measures than when relying on memories	Roy et al., 2008 [67]
	No improvement	Gruschke and Jørgensen, 2008 [27], Jørgensen and Gruschke, 2009 [41]

2.2.1 Focusing on the Task's Proceedings. The planning fallacy commonly results in overly optimistic estimates. Thereby, driving estimators to concentrate on how a given task might unfold could stress pessimistic events and thus guide towards higher estimations of the overall duration. Scenarios encouraging the exploration of pessimistic fallout may reveal all types of potential events that could occur during a task's proceedings. These can be manipulated for various purposes, including playing on pessimism or optimism levels [8], or even helping improve and enlarge predictions' ranges [9]. Even though this strategy did not yield any improvement in estimates' accuracy, it helped uncover that pessimistic alternatives are often seen as less plausible by both estimators and external reviewers [59]. Another way to inflict pessimism on optimistic predictions is to list the multiple surprises that could happen. Neither when participants were asked to establish this list themselves [8] nor

when they were given one [37] did this approach produce satisfying results: it lacked consistency as both over and underestimation were observed.

In the case of underestimation, emphasizing pessimistic events of a task could benefit from being accentuated with other techniques such as visualization tools so that estimators' excessive optimism is better noticeable. Moreover, surprises' likelihoods were only evaluated with a discrete plausibility scale. We suggest more details on how a given event might unfold could be obtained if likelihoods were retrieved through frequencies.

2.2.2 Decomposing the Task. Strategies involving the decomposition of a task build on the idea that a problem is simpler when divided into smaller problems and therefore, smaller components of a task could be easier to estimate. Similarly, the *segmentation effect*, described by Forsyth and Burt, refers to the aggregated subcomponents of a task summing up to a greater duration than the whole task estimate [17], which is the desired effect in the case of an underestimation. In line with those ideas, breaking up a task into smaller components could balance over and underestimation out and thus induce a more calibrated prediction. Unfortunately, the various attempts to showcase the potential of this strategy were inconsistent: the estimates were either still optimistic [9], overestimated but more accurate [17], or equally biased [8]. It was also argued that sub-components might be beneficial thanks to the better view these provide on the overall task, not because of the size of the smaller steps. More precisely, how atypical a sub-task is, constitutes the key component influencing the direction of the estimation [29]. This has disclosed promising results in at least five experiments and revealed that the effect was even more significant as the tasks gained in complexity [51].

These procedures rely on two different motivations while being quite similar. Making use of such methods could, therefore, maximize the possibility for a practically relevant effect. Additionally, capturing information on sub-components confers a better understanding of how the whole task is perceived by the estimator.

2.2.3 Altering the Estimation's Instructions. If this bias is too intricate to overcome, another approach can be to act on the instructions of the estimating task to modify the perceived purpose of the exercise [76]. In line with the distributional versus singular data theory, Griffin and Buehler investigated the influence of individual probabilities compared to aggregate frequencies on the type of information used to draw inferences and predict some tasks' duration [26]. They hypothesized that requesting frequencies would reveal an increased usage of statistics when predicting. Nevertheless, no improvement in accuracy was reported. More generally, different wordings could generate different effects. Inverting a question by emphasizing how much work can be done rather than how long it could take, can be used as a medium to turn underestimates into overestimates [30, 33, 34]. Additionally, driving participants to focus on why, instead of how, also influences the direction of the cognitive bias, depending on the original task's duration [69]. This outcome can even be obtained by focusing either on the speed or the duration [55] or by influencing the perception of the amount of time granted for a task [68]. More theoretically, words can be used to alter time perspectives [5], and thus estimates.

While different wordings were studied, we lack concrete knowledge about the influence of inputs' format on bias mitigation. Estimates are generally requested as precise points in time, yet it might be confusing to provide this kind of value for uncertain events.

2.2.4 Reviewing Past Elements. The past is where most knowledge about a given task can be captured. When doing so while relying simply on memory, some elements are pertinent but overlooked while others are judged important but may lead to inaccurate estimates. Asking observers to predict a task's duration, rather than the actual performer of the task, prevents the use of personal beliefs and most generally optimistic ones: actors (*i.e.*, those who actually perform the task) often tend to justify their underestimated predictions. More precisely, they usually see past events as accidental misfortunes that are very unlikely to happen again, or fully trust in their ability to perform better than previous times. Still, even though observers provide more pessimistic estimates, they are not more accurate than actors [7, 59]. In at least one experiment [8], this strategy has failed to generate a measurable effect, also suggesting it may not be systematic. Yet, some elements might be relevant to integrate into the estimation. In line with this idea, Buehler and colleagues asked participants explicitly past-related questions to guide their focus [7]. The resulting predictions were, however, divergent and approximately as pessimistic as optimistic. Furthermore, when experts were requested to concentrate on their memories as novices, it did not yield the coveted results either [37].

Another theory is that humans might not need to be *tricked* into thinking differently, they could be *trained*. Feedback could gradually help people gain better prediction or estimation skills. Tobin and Grodin, for example, found that professional athletes are better at both predicting and estimating their performance time than are casual runners, which they explained as a result of the daily feedback they get during their training [73]. Further evidence supports the efficacy of this strategy: people seem to do better with more accurate measures of their own or others' past performances [1, 64, 67]. Yet, the way feedback is delivered may be crucial as some studies failed to replicate this effect [27, 41].

These procedures are motivated by an erroneous conception of the past and the relevance of past events, as the source of this bias. If combined with strategies inspired by the inside/outside account, their influence would cover most of the plausible theories on the causes of the planning fallacy.

The described attempts constitute inspiration to devise an effective approach in order to mediate biased time estimates, for these rely on multiple aspects of the estimation process. Additionally, these methods allow us to retrieve visualizable beliefs on task durations when operated as an elicitation procedure.

2.3 Visualization Approaches for Time-Oriented Issues

The visualization community has proposed several approaches to reduce the effort for individuals to properly plan. These visualization techniques aim to provide overviews and analysis tools to alleviate the associated cognitive load.

PlanningLines is a visualization technique aiming to support projects' planning with temporal uncertainties [2]. It offers an overview of a task's plausible duration using glyphs, which has

led to faster assessments of the presented information. While this system addresses time and planning issues, most reported visualization approaches focus on broader matters. Hartl described a 3D calendar visualization that allows for data comparison in a planning context, along with other features such as the discovery of patterns [31]. In fact, pattern detection has been the point of interest of numerous studies investigating different visualization techniques that could enhance active analysis of data sets. More precisely, DecisionFlow [21], EventThread [28], Frequence [63] and LifeFlow [82] are systems designed for event sequences' exploration and facilitate information readability so that trends are more easily spotted. Even though these visualizations are not closely related to time estimates, their use can be linked to several debiasing procedures. Indeed, this sequencing feature can compare to the method of the decomposition of tasks. World Lines, on the other hand, matches the 'alternate scenarios' approach, for it provides simulation runs to assist decision making [80]. Besides, time-oriented data calls for specific visualization representations. Simple Gantt charts have usually been used to portray this type of information but this solution is space-consuming and gets harder to read when the number of displayed events increases. Blurring and collapsing some parts of those charts were found to be beneficial in time visualizations [54]. Similarly, LiveGantt was introduced as a system demonstrating better scalability than original Gantt charts, while also enabling re-schedulability [39], and TimeBench explicitly allows for several time visualization options to be examined in order to select the most appropriate one [65].

Additionally, correctly fathoming levels of uncertainty related to a given task's duration is a major challenge when using visualization support. There is a growing body of work on related questions such as which visualization techniques best communicate uncertainty in estimates [10, 16, 24, 44], what kind of reasoning strategies are used to make judgments about (uncertain) effect sizes [43] or effects of aggregated vs. distributional representations [60], and even on how to support uncertainty interpretation with Bayesian assistance [48]. This past work focused notably on communicating uncertainty in already existing, externally gathered data, such as weather forecasts, bus schedules, or effect sizes in scientific publications. It informed the design of the approach studied in this article, as we detail in the next section.

3 STUDY RATIONALE

The debiasing methods introduced previously rely mostly on oral or written materials, which consequently require people to mentally compute how their beliefs on uncertainties around the possible proceedings of a task sum up. We argue that easing the cognitive load to make task duration estimations, by computing and visualizing people's beliefs, should result in better estimates. Indeed, in addition to prediction bias, humans are limited in their information processing capacities and can generally keep less numbers in their working memory than even long outdated computing devices [11, 57]. This also correlates with bounded rationality theory [70], in which cognitive limitations are acknowledged in the rationalization of human biases in decision making under uncertainty. As such, we aimed to design a procedure eliciting the data required to compute and visualize an estimator's beliefs on a given task and compare through three measures over time how their different estimates compare: the first one captures people's intuitive estimate of a somewhat familiar task; the second is made after experiencing debiasing techniques, and the third is made after seeing a predictive visualization of their beliefs. By comparing the second to the first estimate, we can measure the potential effects of the used debiasing techniques, and by comparing the third to the second, we can measure if visualizing people's estimates provides any additional benefit over the debiasing techniques alone. Because estimating the duration of a task with precision involves uncertainty, it should appear on the estimation input. Our design uses intervals, as opposed to single-point estimates, to account for this aspect.

Before going into the details of our study design, we describe our rationale for choosing the selected debiasing technique, for computing the probability distributions of estimators' beliefs, and for selecting visualization techniques.

3.1 Selecting Debiasing Methods

To be able to generate visualizations about the plausible values of a task's duration while also emphasizing its uncertainty, we must first elicit the necessary data from participants. More precisely, we want to model participants' views on a task's duration and doing so requires us to elicit their beliefs, *i.e.*, how they think the given task might unfold. Two known debiasing methods are well suited for this purpose: task decomposition and surprise listing.

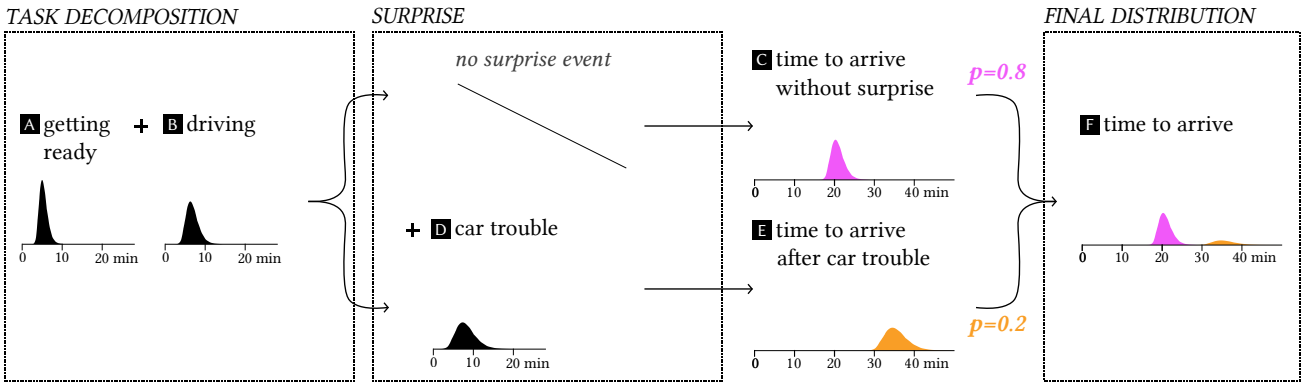
Task decomposition refers to a process in which the main task is divided into smaller components [8, 9]. For example, a morning routine can be divided into smaller steps such as getting up, brushing one's teeth, taking a shower, and having breakfast. This was investigated as a potential debiasing method for the planning fallacy, motivated by the idea that small durations are overestimated [17], which can help adjust underestimated task durations.

Surprise listing, on the other hand, emphasizes the possibility of unexpected events which can disturb the proceedings of a task [8, 37]. For example, continuing the above example, a morning routine could encounter surprises such as being out of coffee, finding the shower occupied, or getting into a lengthy argument with someone. In contrast to the sub-steps resulting from the *task decomposition* technique, surprises do not only have an estimated duration but also a likelihood with which they may occur. As most debiasing strategies aim to increase the overall task duration, surprises are usually events that can slow the completion of a task. Since we do not intend to sway estimators towards optimistic or pessimistic outcomes, our surprise listing method includes both events that may slow and events that may quicken a task.

Both methods on their own have resulted in little to no overall change in previous studies [8, 9, 37], and we expect the same outcome in ours. Their purpose is to help us acquire detailed data on estimators' beliefs and consequently to enable us to generate personalized visualizations of their beliefs.

3.2 Computing Probability Distributions of Estimators' Beliefs

The responses to the elicitation procedures serve as input to establish a probability distribution using Monte Carlo sampling where each instance represents a possible duration of the task, sampling



Each decomposed task's duration is considered to follow a log-normal distribution. The whole task's duration is the sum of the resulting distributions.

If a surprise event happens, then its distribution, which is also encoded as a log-normal one, sums up with the distribution of the whole task's duration.

In the case of a slowing event, the resulting distribution attests to an increased duration.

Overall the two distributions mix where the weight of the surprise event distribution depends on the likelihood of the event.

Figure 2: The probability distribution of the duration of a compound task (here two sub-tasks and a possible surprise event). Without any surprise, all distributions simply sum up ($A+B=C$). In this example, a surprise event (D) results in a much longer task duration ($A+B+D=E$). The likelihood of this surprise event is estimated at 0.2. The overall time to arrive (F) is therefore a mixture of C and E resulting in a bimodal distribution (F).

from the elicited range estimations of a planner. To be able to compute probability distributions, we elicit responses in the form of *prediction intervals*, that is, people's estimation of how long a future occurrence of an event may take based on their past experience. For example, one person could estimate that taking a shower takes them 5 to 10 minutes, while for another person it is 10 to 15 minutes. However, to draw samples from such intervals, we also need to make assumptions about the elicited data, or rather, we need to elicit the data such that our assumptions hold consistently for different planners.

Since we are dealing with duration data, all values are positive. We further make the assumption that task duration data can be approximated through log-normal distributions, which has been shown to be the case for various types of duration data, such as the duration of surgical procedures [71], the underestimated onset duration of acoustic stimuli [18], and users' time spent dwelling on websites [83]. We then model each sub-step and each potential surprise as an independent random variable, each one log-normally distributed. The parameters of the distribution of each random variable can be derived from the prediction interval of the event, if said prediction interval uses a consistent size. For example, for a *95% prediction interval*, which is defined as the interval containing 95% of likely future events, we can derive the parameters of a log-normal distribution as following:
$\mu = \frac{\log(\text{lower})+\log(\text{upper})}{2}$ and $\sigma^2 = \frac{\log(\text{upper})-\log(\text{lower})}{2*z}$ where z is the 0.975 quantile of the standard normal, and *lower* and *upper* are the bounds of the estimator's 95% prediction interval.

To approximate the probability distribution of a planner, we run Monte Carlo experiments on the sum of random variables while considering the likelihood of surprise events (which may or may not be included in any particular instance; see Figure 3). The sum of log-normally distributed random variables is no longer strictly log-normally distributed but shares essential properties such as a single mode and skewness to the right. Taking surprise events into consideration can affect the overall shape of a task's probability density function considerably due to the likelihood of these events. Figure 2 illustrates how sub-steps and surprise events sum up for a simple case with only two sub-steps and one potential surprise event. The overall probability density is generated as a mixture of individual components. For n possible surprise events, the mixture is composed of 2^n components: one for each subset of possible surprise events. The weight of a component within the mixture is determined by the compound probabilities of the surprise events included in that component. The resulting mixture distribution can have multiple modes depending on the number of potential surprise events and their likelihoods.

Past research has shown that people have difficulties properly evaluating compound probabilities of independent events [4] which has been proposed as one of the drivers of the planning fallacy [75]. This justifies the interest of combining the above debiasing techniques with visualizations to depict plausible durations of the task.

3.3 Selecting a Visualization Technique

The main purpose of visualization in this context is to increase planners' awareness of uncertainties and how they sum up, thereby easing overall estimations and decision-making processes. By doing so, we avoid the need for categorical proofs ascertaining whether the inside/outside theory is true, or if it is a memory bias, for we are not relying on one's abilities to remember the past or to individually choose which piece of information is relevant or not [14].

Displaying uncertainty has been shown to positively impact individuals' assessment and evaluation of conflicting information [23].

Because the planning fallacy implies such discrepancies between beliefs and true values, we argue uncertainty should be represented. Past work identified function graphs of probability density functions [24] as the technique supporting optimal decision making of non-experts in the context of uncertainty in weather forecasts. Later work identified more specifically *quantile dot plots* to outperform continuous function graphs [16, 44]. Conceptually, quantile dot plots can be interpreted such that each dot represents a simulated instance of a task. Importantly, any random draw of 20 or 50 samples would likely not be a representative sample of a distribution. Quantile dot plots account for this fact by using *much* larger samples which are then binned so as to select a *representative subset* of data points to visualize. As too many dots make a dot plot resembles a density chart [44], the recommended number of dots is 20 or 50 (as shown in Figure 3). Likewise, displaying too much uncertainty does not lead to better decisions as estimators can be overwhelmed by the amount of information [22]. More precisely, 50-dot quantile plots are described as yielding more consistent decisions, but 20-dot plots minimize bad performances [16].

To study the effect of predictive visualizations on planning behavior, we therefore propose a strategy in which estimators first decompose the task and estimate the duration of the sub-steps before focusing on surprise events and estimating each of these. The various estimates then serve to specify the parameter ranges for the Monte Carlo experiments, and the output of these then provides the data for a predictive visualization, a quantile dotplot. As we are also interested in how predictive visualizations may assist in decision-making situations, we add a second task where we extend the previous scenario by introducing a "deadline" such that estimators need to decide when to start the task they estimated previously to meet that time limit.

The visualization of uncertainty has been an active area of research for multiple decades [62], and there are many different types and needs for such visualizations. The specific ones on which we build our work have been shown to be effective for predictive tasks [16, 24, 44]. We therefore decided to employ the term *predictive visualization* instead of the broader term *uncertainty visualization* to emphasize that we use these visualizations to show predictions of plausible future event outcomes.

4 STUDY DESIGN

We designed a crowdsourced study including two parts: a time estimation task requiring to consider ranges of plausible durations, and a decision-making task requiring to select when to perform a specific action given a cost function. These two parts correspond to two questions one frequently is confronted with when planning a task: "how long will it take me?" (duration estimation) and "when do I have to start to finish in time?" (decision making). Our study design seeks to answer the following research questions:

RQ1 What is the effect of simulating and visualizing data on people's task duration estimates?

RQ2 Do quantile dot plots help indicate 95% prediction intervals? Does the number of dots matter?

RQ3 What is the effect of feedback on people's decision-making? Does the type of feedback (interactive text versus text-and-visualization) matter?

We investigate these research questions through a mixed design where we compare three within-subject measures taken over time to measure how estimates evolve after exposure to debiasing and feedback techniques. Since multiple options exist for the design of feedback techniques, we include two between-subject factors (with two levels each) to compare different visual choices.

4.1 Task and Background Story

In the context of an online experiment, we chose not to ask participants to actually perform the task and measure their performance time—similar to Hayes-Roth [32]. While it would have been informative to be able to compare their estimates against a ground truth, we could only have asked them to complete a filler task whose length they would only be able to estimate if they were already familiar with the type of task, which is difficult to assure for an online experiment soliciting a diverse population. Still, we needed to create an appropriate context for a task all potential participants of a crowdsourced study could relate to. Furthermore, this task had to be long enough for sub-tasks to be relevant and for surprise events to possibly happen. We thus selected a shopping task as the time estimation task and combined it with a train-catching scenario for the decision-making part.

4.1.1 First Part: Estimating a Duration. The shopping task had to be both simple and generic enough for participants with various cultural backgrounds and lifestyles to easily picture. We presented the following context to participants:

> *Imagine that you have friends coming over soon and you realize that you forgot to buy three items which will be essential to spend a pleasant evening together. Imagine three concrete items which can be bought in a grocery store.*

The associated task and instructions were detailed as follows:

> *Your (imagined) task:*
> *You have to go to a grocery store near you in order to buy the necessary 3 items and then get back home.*
>
> *If you were to leave right now, how long do you think it would take you to perform this task?*

After estimating the overall duration of the task, participants were asked to provide duration estimates for all sub-tasks and all surprise events as well as their respective likelihoods. We used predetermined sub-steps and surprise events to avoid situations where two events are joint, or else dependent, which would have complicated both elicitation and computation[1]. We selected these based on a pre-study with 21 participants. This pre-study was also crowdsourced and simply asked participants how they would break down the above task into sub-steps, what kind of events could make them slower or faster, and how likely these events would be to occur. We selected the most common responses for a final set of five sub-tasks and eight surprise events. These surprise events include four slowing events (*i.e.*, events that lead to an increase of the overall duration of the task) and four quickening events (*i.e.*, events that

[1] Allowing participants to manually decompose the task or indicate what could disturb their hypothetical task performance would have also required them to specify which events were inter-related, resulting in a rather complex experiment design.

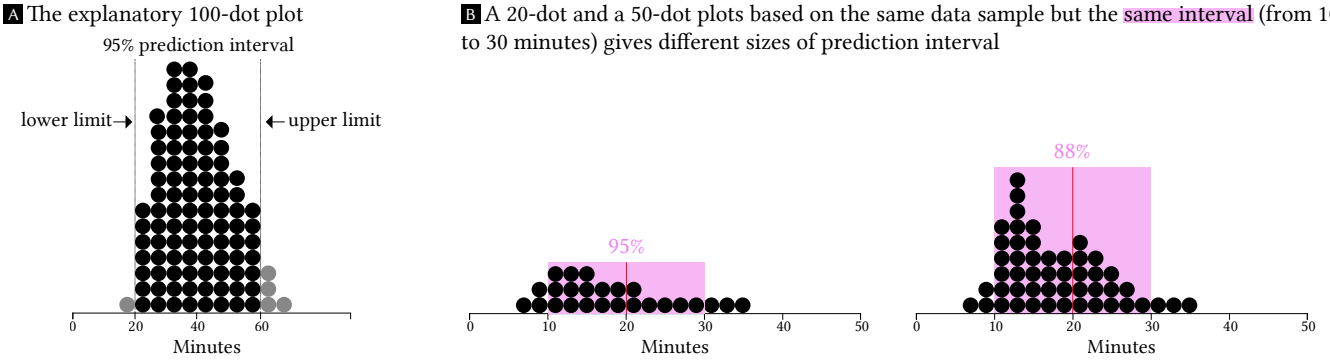


Figure 3: Prediction intervals: (A) the dot plot presented on the explanatory page to introduce 95% prediction intervals to participants; and (B) an example with two quantile dot plots based on the same distribution with a prediction interval's actual size (pink area) of 95% for the sample based on the 20 dots (left), 88% for the sample based on the 50 dots (right) and 87% for the complete sample: it shows how the samples influence the measure of the prediction interval's size.

lead to a decrease in the overall duration of the task). Details are provided in the supplemental material.

4.1.2 Second Part: Decision-Making Task. The decision-making task builds on the previous task to add circumstances that would create a situation in which deciding on the best departure time is not obvious. An appropriate context would be one in which leaving too early is approximately as undesirable as leaving too late. We thus introduced the following storyline:

> *Imagine that you have to catch a train which will be leaving at 4:00 pm, but, before that, you still have to go buy the three items from task 1 in the same grocery store as previously. Once you leave home, you will go to the grocery store and then directly to the train station. As in real life, you do not want to miss your train, but you would rather not arrive at the train station too early either.*

4.2 Stimuli / Visualization Designs

We used participants' estimations for the sub-tasks and surprise events to calculate their personal probability density for the duration of the shopping task as described in section 3.2. Participants saw the density in the form of a quantile dot plot with either 20 or 50 dots, depending on which level of the factor *number of dots* they were assigned to (see Figure 1-A for an example and Figure 3-B for a comparison of showing the same data with 20 and 50 dots). The text above the visualization stated that each dot on the plot symbolizes a simulated trip to the grocery store. Both 20- and 50-dots quantile plots are based on a sample of 100,000 draws and therefore, both are representative of the distribution, but 50-dot plots have better chances of including more extreme values and consequently give a more precise representation of the whole probability distribution (see Figure 3).

For the decision-making task, we provided textual feedback which could be adjusted to explore the outcome of different possible decisions using either a slider tool or an interactive linechart (similar to exploring a statistical multiverse [15]), thus adding the factor *feedback-type*. The slider was a simple gray bar that allowed participants to explore a fixed range of possible departure times by dragging a darker cursor on it (see Figure 4-A). The linechart used a dual-axis design to include both the waiting time (as a blue line) and the risk of missing the train (as a red line) plotted against possible departure times (Figure 4-B). Hovering over the visualization moved a vertical gray line, and clicking on it fixed a second vertical line. This gray line acted in the same way as the cursor for the slider did and indicated which value was selected as the preferred departure time. For both feedback types, textual information was also displayed. All numbers visible in Figure 4 updated according to the current position of the cursor on the visualization or the slider, inspired by the concept of *explorable explanations* [77], thereby detailing the precise departure time which was currently picked, along with the risk of missing a train, expressed as how many trains would be missed out of 100, and the average waiting time. Note that while the average waiting time decreased linearly, the risk of missing a train did not necessarily increase in a linear fashion and could include plateau areas (as shown in Figure 4). The information content of the two feedback types was in theory equivalent, although with the slider, participants would have to look at the entire range of the slider to extract the included information whereas the visualization provided an overview. To make sure that participants realized that they could explore different decisions, interaction with the slider or the visualization was required to be able to continue.

In summary, our design included 2 factors, one in each part of the study, resulting in 4 conditions: 20-dot and slider, 20-dot and linechart, 50-dot and slider, 50-dot and linechart. Participants were randomly assigned to any of the possible four conditions.

4.3 Dependent and Outcome Variables

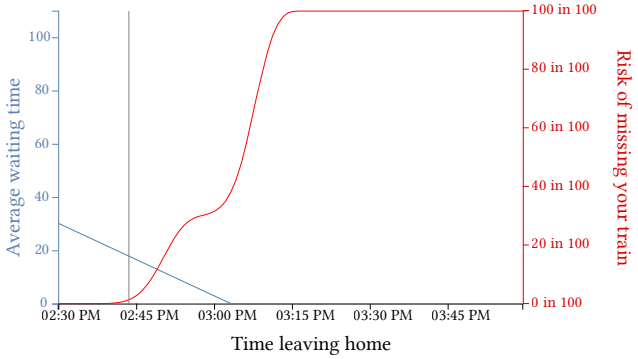
We captured the following dependent variables:

(1) three estimates of the duration of the whole task as an expected value (or central tendency) plus a symmetric interval size (x +/- y minutes) taken
   - after reading the task description (E1),
   - after going through the debiasing techniques (E2),
   - after seeing the quantile dot plot (E3);

A If you were to leave at **2:43 PM**, you would miss your train **1 time out of 100**.
Your waiting time would be **20 minutes** on average.

Your selected departure time: 2 : 43 PM
By leaving at that time, you would miss 1/100 train(s).
You would have to wait 20 minutes on average.

B If you were to leave at **2:43 PM**, you would miss your train **1 time out of 100**.
Your waiting time would be **20 minutes** on average.



Your selected departure time: 2 : 43 PM
By leaving at that time, you would miss 1/100 train(s).
You would have to wait 20 minutes on average.

Figure 4: The two feedback conditions, the slider (left) and the linechart visualization (right) are based on the same values and set at a departure time of 2:43 PM. The text above the linechart changes when hovering over it while the one below corresponds to the selected value (when clicked). For the slider, both text sections are adjusted when moved.

(2) the first 10,000 samples generated for the quantile dotplot[2];
(3) two decisions when to leave for the train, logged as the number of minutes before the train leaves, with associated waiting times and likelihoods of missing the train according to a participant's personal distribution taken
   - after reading the train scenario description and
   - after interacting with the assigned feedback type;
(4) time spent on the final departure-time-selection page, overall and interacting with the feedback technique.
(5) self-reported lateness on a 7-point scale.

From these we computed the following outcome variables to answer our research questions:

RQ1: increase between estimates, both for the point estimate and the symmetric interval around it
   - overall (E3-E1)
   - attributable to debiasing techniques (E2-E1)
   - attributable to visualization (E3-E2)

   We normalized these differences as *percent increase* ($\frac{B-A}{B}$) since each participant responded on the basis of different locations of supermarkets in their vicinity, where some may live right next to one whereas others would need to take a car for 15 minutes. Past work suggests that people tend to underestimate task durations, thus we should expect an increase in normalized differences after participants see a visualization.

RQ2: actual sizes of the elicited prediction intervals for each of the three estimates (E1, E2, E3)
   We computed the actual size by counting how many of the 10,000 logged simulated instances of the task fall inside the intervals participants provided for the three different estimates. If quantile dot plots help indicate 95% prediction intervals, then the final estimate should be closer to a 95% interval than the first two. As Figure 3-B illustrates, the same underlying data can suggest different 95% prediction intervals, especially for distributions that are wide relative to the bin width. Thus, we also test whether participants assigned to a plot using 50 dots are closer to a 95% interval than those assigned to a 20-dot plot.

RQ3: difference between the initial and final likelihood of missing a train, and waiting times.
   We computed both signed and absolute differences since both are needed to give a full picture: the signed difference can be close to 0 if there are as many participants who leave earlier after receiving feedback as there are participants who leave later, whereas the absolute difference would be considerably different from 0 in that case. Using only an absolute difference would hide in which direction participants change which is particularly informative in the case of how many trains are missed; that number should not increase following our intervention. Indeed, we are interested to learn if any type of feedback improves people's decision-making. We consider that decisions are overall improved if, *after* seeing feedback, fewer trains are missed (signed difference) without increasing waiting time considerably (signed difference).

[2]We realized during pilot studies that the logging of the complete sample of 100,000 floating-point numbers, sent without compression on participants' machines through a simple HTML form POST, resulted in considerable amounts of data being sent for each participant which was disproportionate compared to what was required to calculate our outcome variables. Based on simulations, we determined that 10,000 samples rounded to two decimal places were largely sufficient to calculate the desired variables while requiring only about 60 kB of data to be transmitted.

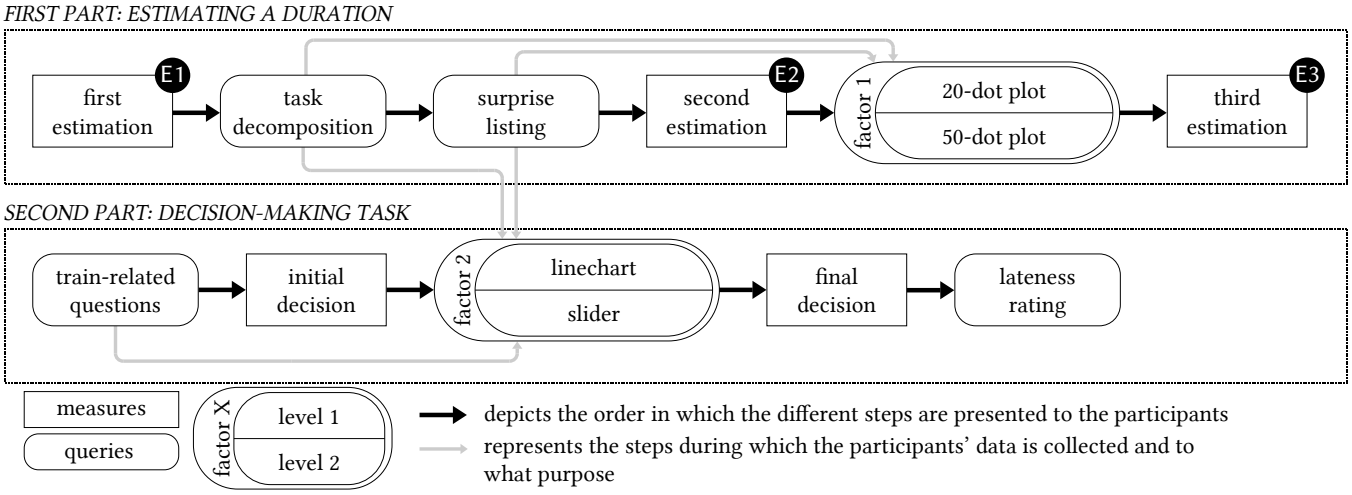


Figure 5: The experiment is divided into two parts: the first one is composed of three measures, E1, E2 & E3, two types of queries (the debiasing methods used to retrieve participants' personal beliefs) and one factor, *number of dots*, with two levels, while the second one comprises two measures (based on a decision-making task), one query (that informs the following intervention), another query for additional information on participants and one factor with two levels.

Finally, we test whether the type of feedback matters by comparing responses for the interactive-text group with those from the text-and-visualization group.

Additionally: a correlation coefficient between self-reported lateness and estimated waiting times for initial and final departure decisions. We use this measure for additional analysis to test if preferred waiting time could be predicted from someone's self-reported lateness.

4.4 Additional Logged Variables

Apart from the dependent variables, we saved additional data from participants, including the time spent on each page of our study and the overall duration. We also logged the dot plot generated for each participant and, for the visualization feedback group, the linechart each time it was clicked which shows the current selection (all of these are included in the supplemental material).

4.5 Procedure

Before the actual experiment started, all participants were asked to read an information sheet and give informed consent. We then presented the concept of 95% prediction intervals on an explanatory page including a short definition and an example using a dot plot representation with 100 dots so that a 95% interval corresponds to 95 dots (see Figure 3-A). We solicited estimates of intervals using two formats: either a central tendency plus/minus an interval or as lower and upper bounds. The first type was used for all dependent measures where we wanted participants to focus separately on their intuition (the central tendency) and then consider how much variation they would expect. For the sub-tasks and surprise events we wanted them to only focus on bounds, the lowest and highest values they considered reasonable, hence the second type of input.

The experiment consisted of two parts, as mentioned before: the first one centered around a time estimation task (*how long will it take?*), and the second part introduced an additional decision-making scenario (*when do I have to leave?*). Figure 5 provides an overview of the entire experiment. Participants performed both parts without interruption.

4.5.1 First part: estimating a duration. Participants were presented with the context and details of the task and had to provide a first estimation for the whole duration of the given scenario. Then they went through the debiasing technique where they had to fill in a lower and upper bound for all sub-tasks, the four slowing surprise events, and the four quickening events. On the next page, participants were asked with which frequency they expected each of the previous eight surprise events to happen (*e.g.*, it happens 3 times out of 10). Participants then provided a second estimation to measure the effect of the debiasing technique alone. Finally, they saw either a 20- or 50-dot quantile plot computed from their previous responses using the procedure described in section 3.2. Underneath this visualization, they were asked to provide a third estimation of the task's duration.

4.5.2 Second part: decision-making task. To prepare the new simulation for the decision-making section, we asked participants for lower and upper boundaries on how long it would take them to go from the grocery store they pictured for the first task, to the nearest inter-regional train station. Then we asked how much time would be added should a "major disturbance" occur during their trip and how likely such an event was. Additionally, they were asked whether they preferred the 24- or 12-hour timekeeping system to show the feedback in a format they intuitively recognize. The next page asked them to provide their preferred departure time. These data were then used to compute and present their expected waiting time and likelihood for missing their train given their stated preferred time. They were then prompted to interact, depending on their assigned group with a slider bar or the function chart, in

order to explore the data and change, if desired, their preferred departure time (see Figure 4). Under the feedback display, there was an optional comment box designed to let participants explain their decision if they wished to. This component was added following pilot tests, so that participants could feel freer to select a departure time that fits their preference, rather than what they believed was expected of them. They were asked, on the last page, to self-report their lateness on a 7-point Likert scale with the following values: (1) always late; (2) often late; (3) sometimes late; (4) rather on time; (5) sometimes early; (6) often early; and (7) always early. A text field was available for additional comments, right above the button to finish the study.

We included three attention checks throughout the different pages of the experiment according to the attention check instructions provided by the Prolific platform[3]. These checks required participants to enter or select the value indicated in the text, such as "enter 1 in both fields". The system was set up to automatically exclude participants who failed more than one attention check. A static printout of all experiment pages is included in the supplemental material and available here.

4.6 Participants

This study was approved by the Comité d'Éthique de la Recherche at Sorbonne Université (avis N° CER-2020-65). We recruited 160 participants on the Prolific platform. The only criteria used were fluency in the English language, having at least a 95% acceptance rate for previous contributions, and having completed at least 10 prior experiments. Participants had access to the title of the experiment, *Study on estimating durations and probabilities in the context of everyday tasks*, and a small description of what will be expected of them. They were offered £1.80 for an estimated duration of 12 minutes, averaging at £9/h or £0.15/min.

Out of the 160 contributors who provided either a completion code or marked the task manually as completed, our system recorded data for 157 participants. Our pre-registration specifies that we would exclude participants for whom any data is missing, which was the case for 12 participants, resulting in a final 145 valid responses. Specifically, out of the 12 excluded participants, our log files showed impossible values for 9 of them[4]. For three of the twelve excluded participants, the visualization shown to them was missing data lines or axis labels such that the visualization could not be interpreted). Six participants failed one attention check without being excluded (as per our pre-registration), but no one failed two (which would have led to automatic exclusion).

The research questions and analysis plan for this study were pre-registered on OSF, along with the data and analysis script.

5 RESULTS

As specified in our pre-registration, we report our results using estimation statistics [12], that is, we use confidence intervals to evaluate the strength of evidence for the investigated effects [13]. Our inferences are based on graphically-reported point estimates and confidence intervals; the exact numbers can be found in the supplemental material. We use 95% BCa bootstrap confidence intervals which have been shown to provide good estimates without distributional assumptions as long as sample sizes exceed 20 [49] (the smallest analyzed group in our study has 63 data points).

We performed an attrition analysis to determine whether one condition caused participants to drop out more than the others [84]. We found a slight difference between 50-dot condition groups (visualization vs. slider) of 19% (95% confidence interval: [2.7% - 36%]). However, this difference is not due to participants failing to complete the study, which would have been a thread to its validity. Rather, of the twelve participants whose submissions we had to remove from the final data set due to missing data (as detailed in section 4.6), nine were in the visualization condition. As such, the difference in group size is not surprising and is unlikely to indicate participants voluntarily dropping out at different rates in different conditions.

5.1 RQ1: The Effect of Predictive Visualizations on Duration Estimates

To investigate our first research question, we examine to what extent our intervention, showing a predictive visualization, induced a change in people's duration estimation. As described in section 4.3, we operationalized this question through the normalized difference from the initial to the last estimate, a within-subject measure. Figure 6 summarizes our results; the upper part focuses on increases of point estimates (that is, how much higher people's last estimate is relative to their first) and the lower part on the level of uncertainty participants indicated (that is, how much wider participants' final interval is relative to their initial).

We observe an overall considerable increase from the initial to the final estimate for both the point estimate (70%, 95%CI [48%, 100%]) and the uncertainty interval around it (62%, 95%CI [41%, 110%]), which suggests that, in average, participants realized, after seeing the visualization, that their initial estimates and sense of uncertainty were too optimistic. Looking at the sources of the increase, we find that for the point estimate, the debiasing method alone led to only a modest increase, which is consistent with results reported in previous work introducing these methods. The effect of the visualization seems to be at least twice as big, that is, on average participants increased their estimates more after seeing their visualization than after only reflecting on task composition and surprise events.

To get a better sense of the variety of responses and individual differences concerning time estimates, Figure 6 also includes histograms of the measures. Notably, we can see that the mode of all response distributions is at 0, which illustrates that some participants did not change their responses at all. Figure 6 shows analyses for point estimates and intervals separately, which does not permit us to show that about 25% of participants kept *both* parts of their initial estimate constant across all three estimates[5]. The histograms also show that all distributions are heavily right-skewed – indicating that some participants increased their responses multi-fold, with some extreme cases providing estimates six or seven times

[3]https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Using-attention-checks-as-a-measure-of-data-quality

[4]Some values suggested unexplained logging issues (*e.g.*, 102 years spent interacting with our tool). We removed those participants because we could not ensure the integrity of the other values.

[5]This and other descriptive results are part of our complete analysis report (accessible on OSF).

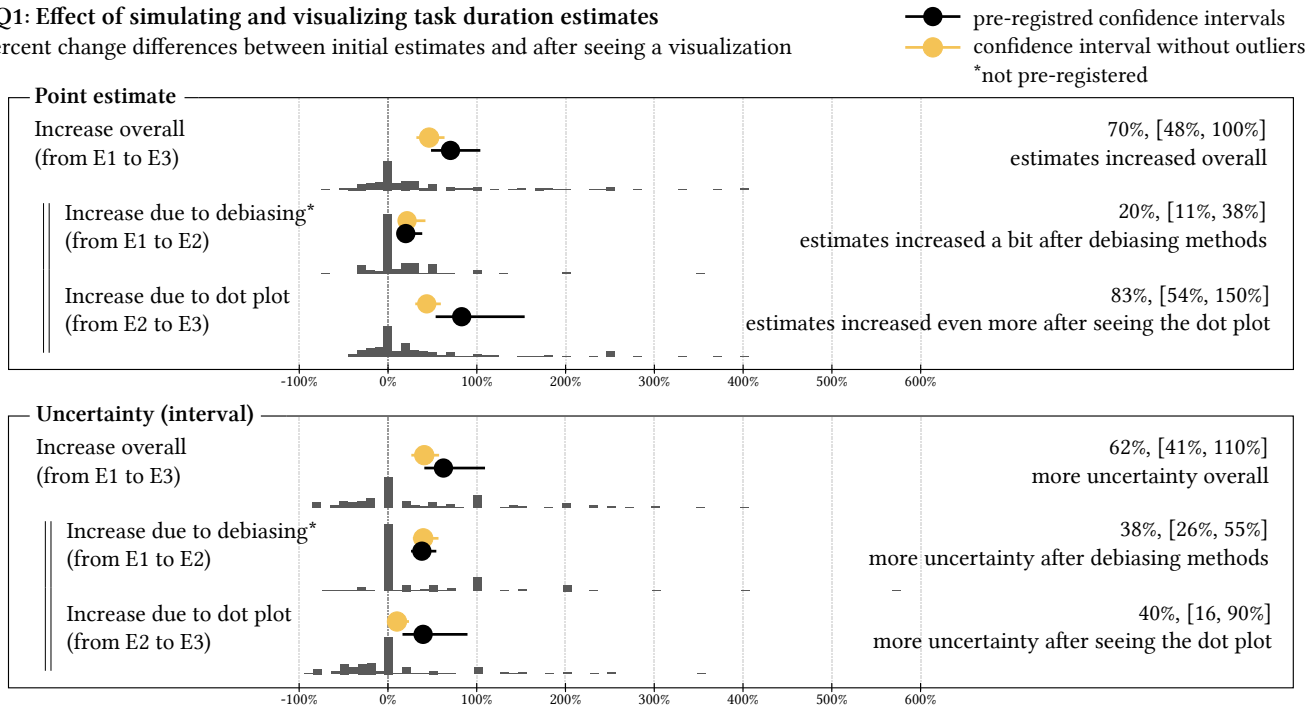


Figure 6: Results for research question 1. Error bars show 95% bootstrapped confidence intervals for both pre-registered measures and those without outliers. Histograms include all values that were pre-registered as admissible. For the point estimate, we can observe that most of the overall increase is due to the latest estimate (that is to say, after seeing the dot plot). The effects of the debiasing methods and the dot plot on the width of the interval (used to evaluate the uncertainty) is harder to attest for with a smaller difference: both led to approximately the same increase in intervals' width.

higher than their initial interval while downward adjustments were more subtle. Since these few extreme cases may inflate our reported effect size, we additionally report in Figure 6 estimates where extreme outliers (more than 3 standard deviations from the mean) are removed (this part of the analysis was not pre-registered and is shown in yellow in Figure 6). The direction of all effects reported above holds under the exclusion of outliers, but the adjustment results in some cases in smaller and likely more accurate effect sizes. Notably, filtering out extreme outliers suggests that the debiasing techniques may be effective to get (at least some) people to reconsider their uncertainty and increase at least the *interval* around their estimate (bottom part of Figure 6). However, to increase their overall estimated time (top part of Figure 6), the visualization was more effective than the debiasing technique.

5.2 RQ2: Effect of Dotplots on Prediction Intervals

We instructed participants to provide estimates aiming for 95% prediction intervals throughout the whole study. Since participants only imagined the task, we do not have actual durations to determine the accuracy of their estimations. However, we have *simulated* instances of the task based on participants' responses to the subtasks and surprise event questions[6]. We use these simulated trips to determine the proportion falling into participants' overall estimates E1, E2 & E3 (for illustration refer to Figure 3-B which shows the coverage of prediction intervals overlaid on quantile dot plots). A perfect estimate would result in 95% of all simulated instances falling within that estimate.

The results for RQ2 are summarized in Figure 7. We find a coverage below 50% for the first estimate, a small improvement for the second estimate and, on average, a rather large increase in the coverage of the interval for the final estimate, after participants saw the visualization. The included histograms uncover three relevant observations: (1) a high number of participants provided first and second estimates which are completely outside the range of the samples for their data, illustrated by the peaks at 0%, which is only the case for a few participants for their third estimate, (2) we do not observe a clear peak at 95% for any of the three estimates, and (3) the number of people who "overshoot" and provide intervals which cover 100% of the simulated data roughly doubles from the second to the third estimate, but remains lower than the peak at 0% for the first estimates. The histograms also suggest that while the debiasing alone did not have much of an effect on those participants whose

[6]These simulated instances are those used to generate the quantile dot plot. They are a representative sample (n = 10,000) taken from the mixture distribution elicited through the debiasing techniques as detailed in section 3.2.

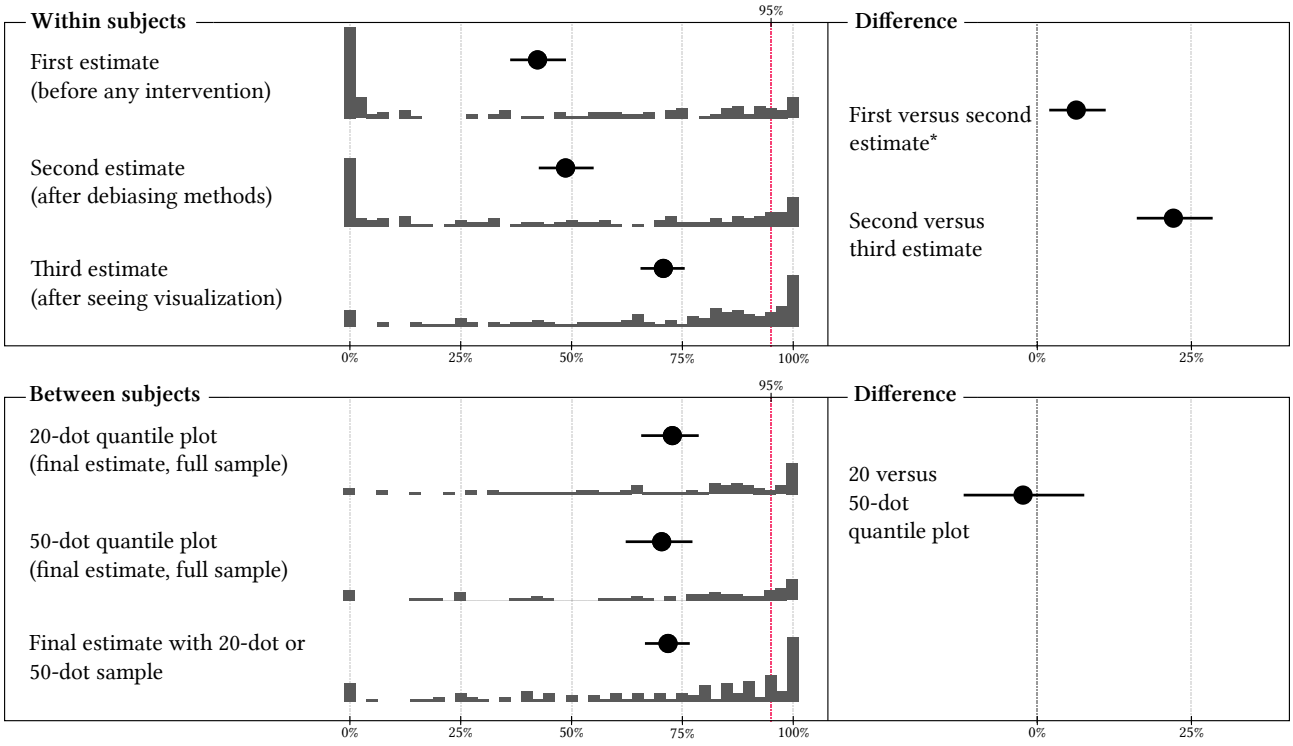


Figure 7: Results for research question 2. Error bars show 95% bootstrapped confidence intervals. On the left are estimates for the individual measures and on the right are differences between pairs of measures on the left. The further a difference is from 0, the bigger the effect. The within-subject measures (top) show participants get closer to a 95% prediction interval after seeing the dot plot (third estimate) even though many do not get close to the 95% mark. The between-subject measures (bottom), used to compare the 20-dot quantile plot to the 50-dot one, do not attest to any relevant difference between the two choices when it comes to prediction intervals.

coverage was 0% for the initial estimate, the density upwards of 75% increases somewhat for the second estimate. As the differences on the right-hand side of Figure 7 illustrate, we found strong evidence for an increase in interval coverage from the second to the third estimate. Concerning the question of whether plots with 20 or 50 dots lead to better coverage of the estimate, our data do not support the hypothesis of one being more suitable than the other to provide 95% prediction intervals.

5.3 RQ3: Effect of Two Types of Feedback on Decision-Making

Our third research question investigates how feedback affects the decision of when to leave to catch a train. Participants received feedback in text form spelling out their expected waiting time and the corresponding likelihood to miss their train, both as a function of departure time. They could explore how these two outcomes changed depending on different departure times using either a slider or a visualization (as shown in Figure 4). We analyze the difference in missed trains (in percent) and waiting time (in minutes) between their initial selected departure time (before any feedback) and their final selected departure time (after interacting with feedback). Results are summarized in Figure 8.

Most notably, the *signed* differences for the two outcomes suggest that participants overall optimized their decisions, that is, across participants, there are fewer missed trains while waiting times increase only slightly. Nonetheless, the *absolute* difference in waiting time indicates that more participants changed their departure time than the signed difference suggests: those whose initial departure times resulted in very high waiting times decided to leave later, and those who probably would have missed their trains decided to leave earlier. The larger effect on the reduction of missed trains over the increase in average waiting time can be interpreted as an overall optimization of behavior: fewer trains are missed and the average waiting time increases only a bit. Concerning the type of feedback, slider versus visualization, our data suggest that the two groups behaved similarly. This might be due to all relevant information being in the text which was identical in the two conditions. However, we observed a difference in the time taken to come to a decision between the linechart (122s, 95% CI [108s, 137s]) and the slider (79s, 95% CI [70s, 90s]), meaning the visualization group spent more time on the final decision page.

RQ3: Effect on the decision when to leave for a train after receiving feedback based on simulation data
How much do estimates change and how much does feedback type matter?

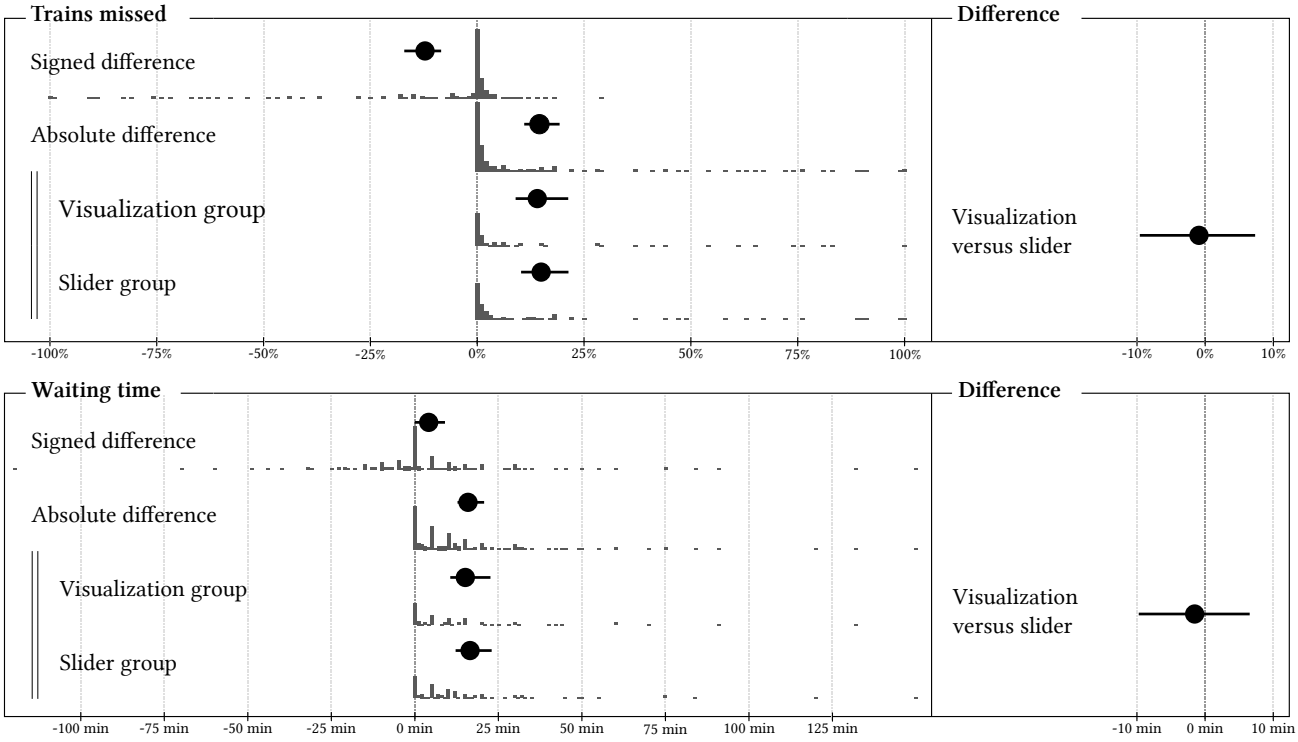


Figure 8: Results for research question 3. Error bars show 95% bootstrapped confidence intervals and the underlying histogram shows the corresponding raw differences for each participant (within-subject measure). Signed differences emphasize changes across the simulation (less trains missed, only slightly increased waiting times) whereas absolute differences emphasize individual changes irrespective of the direction of the change. The right side shows the computed differences between the visualization and slider group (between-subject measure).

5.4 Additional Analyses

We also performed additional analyses, allowing us to compare our results to those of psychology papers, probe individual preferences in the decision-making process and explore the different strategies participants used when optimizing. Additionally, we considered respondents' optional comments.

5.4.1 Previous Work in Psychology. Our measurements also allowed us to determine whether our results are consistent with previous work in psychology. Specifically, we analyzed how similar participants' responses remained between their first and second estimates (included in Figure 6). According to the literature, the planning fallacy commonly indicates estimates are too optimistic. Consistent with past work, we did observe evidence for overly-optimistic initial estimates with an increase in central tendency and interval size. We also noticed that about a quarter of participants did not change their initial estimate at all. The average absolute variation is 28% (95% CI: [20%, 47%]) of the first estimates' central tendency, which suggests the combination of the decomposition and surprise listing methods had an influence on the estimation. More precisely, the average signed variation is an increase of 20% (95% CI: [11%, 38%]) of the first estimate, attesting to a noticeable decrease in optimism. As for the interval size, the variation represents, on average, an increase of 38% (95% CI: [26%, 55%]), which indicates larger intervals and thus an increase in uncertainty[7].

Simultaneously, we were expecting the sub-tasks to undergo the same segmentation effect as described by Forsyth and Burt [17]. Indeed, the motivation behind task decomposition is that smaller tasks are easier to estimate for and thus prompt less optimism. We therefore anticipated the sum of the sub-task estimates would be greater than the overall estimate. Nevertheless, it was not the case: when we summed the central tendencies of the sub-tasks (without the surprise events) and compared the result to the central tendency of the first estimate, the most optimistic of all three estimates, then we find that the sum of the sub-tasks was smaller (with a difference in mean of -0.41, 95% CI [-0.5,-0.29]), the opposite direction. This compares better with the experiment from Connolly and colleagues [9], in which optimism persisted.

Furthermore, while looking at the raw data, we noticed that participants often rounded their estimates to the nearest multiple

[7]Refer to OSF for the full statistical analysis report.

of 5[8], a behavior regularly observed in perception experiments (*e.g.*, see Talbot et al. [72]). This impacts both overall accuracy (*i.e.*, how their estimates compare to their actual performance time), and the precision of their expressed beliefs, yet it likely reflects how people behave in real life.

5.4.2 Individual Preferences. Participants presented diverse preferences in the decision-making task, that is, they chose different tradeoffs between waiting time and risk to miss one's train. We were interested to learn if there was a simple mechanism to predict these choice preferences and thereby computed a Kendall's τ correlation coefficient to test if their self-rated lateness would predict whether they optimize according to the waiting time or the risk to miss the train. However, we found no evidence for any correlation for either the initial ($\tau = -0.08$, 95% CI $[-0.2, 0.05]$) or the final decision ($\tau = -0.1$, 95% CI $[-0.2, 0.01]$).

5.4.3 Heuristics for the Train Decision. We used participants' remarks from the comments field and their interaction traces, such as screenshots of their visualization, to identify heuristics participants may have used during their decision process. We identified the following strategies:

- **Fixed waiting time**: Adjusting to arrive at their usual waiting time (one participant indicated explicitly in the comments to always arrive 20 minutes in advance), visible through a rounded waiting time such as 15 minutes and not rounded risk and departure time such as 2:39 PM. 18 out of 145 participants used waiting time to decide.
- **Fixed departure time**: 35 out of 145 participants did not change their departure time from initial to final decision (one indicated always leaving around the same time before going to the train station).
- **Line intersection**: We observed through the logged *svg* files of the participants in the visualization condition that 6 of them (out of 63) chose the point where the two lines intersected. Note that in most cases, this intersection point is a rather risky decision. 12 more participants explored this point but decided in the end on a different option.
- **Latest safe point**: Some participants (10 out of 145) selected the latest point where the probability to miss their train was indicated as 0[9], and 14 participants selected a departure time with a risk indicated as less than 1 out of 1000, out of a total of 35 participants who optimized according to the likelihood of missing a train (mostly visible for the visualization condition group for whom each click was recorded).

For the remaining 50 participants we could not conclude which factor was the decisive one as their behavior was less consistent or did not exhibit an identifiable heuristic.

5.4.4 Participants' Comments. Two participants left additional comments at the end motivating the implementation of a similar system to be used more generically:

> *"[...] I would love for everyone to have this experiment, so they could apply these considerations on their everyday life."*

and:

> *"This was eye opening. Should turn it into a planing app or sum."*

Others also described the study as "eye-opening", "fun" and "interesting".

6 DISCUSSION

Results suggest that predictive visualizations prompt considerable reductions in optimism and an increased sense of uncertainty. Concerning the decision-making part of our study, we observed what can be interpreted as an overall optimization of decisions: fewer missed trains and only a small increase in average waiting time.

Surprisingly, many participants made initial estimations which are completely inconsistent with their responses to the sub-tasks and surprise listing, that is, not a single instance simulated from these responses fell into their initial estimate nor in their updated estimate, which they gave just after providing these responses (visible in the spikes at 0 in Figure 7). While a few participants continued to insist on their inconsistent estimate even after seeing their predictive visualization, most adjusted their estimate upwards with the bulk of responses being between 75% and 100%. It should be noted that we also found an increase in the proportion of participants overshooting, that is, providing a too high or wide interval. While overestimation can incur a considerable cost, it is reasonable to assume that people would be able to calibrate themselves over time [67, 73].

Furthermore, about a quarter of all participants kept the same estimates from beginning to end. Various reasons can explain this behavior, including the confidence in one's ability to accurately estimate the task's duration, or the *continued influence effect*, defined as one's persistence in considering and relying on a rejected hypothesis [36]. Estimators might also lack trust in the displayed predictive visualization. Such limited confidence has previously been shown in predictive decision-making with machine learning systems and is further exacerbated by cognitive load [3, 85, 86].

6.1 Interactive Text versus Visualization

Interestingly, we did not find a difference between the slider and the visualization feedback type in the second part of our experiment. This is likely due to all task-relevant information being inserted in the text at pertinent places, so that it was not necessary for a participant to decipher the information contained in the visualization. We did observe that participants in the visualization condition spent more time interacting with the visualization than those who interacted with the slider feedback. We can only speculate why that might be. It could be that it was more difficult, but it could also be that the overview provided them with additional insights which they wished to explore further. Participants may have also been able to gain a better understanding of the non-linear relationship between the likelihood to miss one's train and the expected waiting time. Since we did not provide indications on how to read the linechart, participants in that condition may have struggled to

[8]Only 3 participants never used a multiple of 5 in any of their three estimates (E1, E2 and E3) and approximately 41% exclusively provided multiples of 5 for their three estimates.

[9]Which means that, given the data they provided, none of the simulated trips took longer than the time available between their chosen departure time and the time the train would leave

interpret what they were seeing. Future research is needed to better understand the mechanisms at play.

Additionally, anecdotal evidence from participants' comments (section 5.4.4) suggests that there may be an added learning benefit which could help them in their future time planning needs. More studies looking specifically into learning effects will be necessary to determine and quantify the potentially added benefit of the used interactive visualization over the interactive text.

6.2 Eliciting Estimations

It might be intricate to retrieve people's complete beliefs and to apply the same assumptions to everyone. We asked participants to enter durations as ranges (lower and upper bounds) in respect of a 95% prediction interval and we then assumed those values followed a log-normal distribution. While the latter choice was informed by prior work [18, 71, 83], we are unaware of work studying explicitly the relationship between elicited prediction intervals of durations and how to best model them. Additionally, providing 95% prediction intervals is not an easy task and untrained people tend to provide intervals that are closer to 50% [38, p. 112], which is consistent with our findings for the initial and second estimates.

More elaborate user interfaces to elicit people's beliefs could therefore be useful. For example, web-based tools to elicit probability distributions have already been developed [25, 58] and could be integrated into our procedure. Entering data would be more time-demanding but advanced inputs could also prompt reflection, and thus more considered and well-thought-out estimates. A potential risk is still the mathematical knowledge that is necessary to make full use of such tools.

6.3 Accuracy

It is essential to determine whether our strategy provokes overestimates or maintains underestimates, or whether individual differences are reduced or increased. To investigate these questions, comparison with concrete performance times is paramount. However, requesting participants to actually perform a task after predicting also raises issues: if there is no experimenter to record the duration, participants can measure inaccurately or misreport their durations, and even if an experimenter is timing them, they can still adjust their behavior, that is, try to speed up or slow down, to match their prediction.

6.4 Biases and Resource Rationality

One potential limitation of our approach could be human biases as a model mechanism [79]: if people's inputs are already biased, then the distribution function is calculated on erroneous values and presents therefore incorrect information. Past work suggests that estimating sub-tasks is easier and leads less likely to underestimations [17]. That, however, is not consistent with our data where sub-tasks without the inclusion of surprise events led to *more optimistic* estimations than even the initial estimates. Our approach of combining sub-tasks *and* surprise events may be a promising approach to elicit more realistic responses, yet, this remains a hypothesis to test in future work as we are not able to determine this without collecting actual performance times of a task.

As visual analytic systems also suffer from human biases, metrics evaluating to what extent the visualizations in our approach undergo similar effects could be integrated [78]. This would help design appropriate visual representations that diminish those analysts' biases. We should beware, however, not to subdue positive bias, or fall into the bias bias [20]. Additionally, based on bounded rationality theory, a resource-rational analysis, as suggested by Lieder, Wesslen and colleagues [52, 53, 81], could help better understand decision making with data visualization as it takes cognitive limitations into account in the evaluation, rather than solely irrational biases.

The goal of work like ours is in the end to help people make better decisions and avoid social or financial problems due to erroneous estimates. Yet, when studying this problem space, one needs also to consider that such estimates may not necessarily be due to an incorrigible bias but that other dynamics may play a role, similar to Gigerenzer's suggested alternative explanations for various biases reported in the literature, such as intuition about randomness and paying attention to the *framing* of a problem [20]. Similarly, people might use heuristics to simplify complicate decisions [19], such as when one always aims to arrive at the train station 20 minutes in advance (as reported in section 5.4.3). The data supplied with this article supports the hypothesis that the planning fallacy may, at least in part, be due to the difficulty of deriving estimates for compound probabilities and mixture distributions. More work will be needed to test this hypothesis in real-life planning scenarios and in the context of different cost functions.

6.5 Incentives

In the second part of our study, we asked participants to select a departure time so that they would not miss their train but also not arrive too early (we wanted both cases to be equally discouraging). In real life, missing a train is more punishing than arriving early and having to wait, and as such, incentives are often asymmetrical. We opted for a scenario with a *light* asymmetry (that is, there is often a later train that could be taken) which corresponds to a variety of everyday constraints, whereas a scenario with a *strong* asymmetry (high costs when missing one's plane) may result in overall more pessimistic responses.

Besides asymmetric incentives there may be other reasons for people to make biased estimations when trying to plan a task. An important factor to take into account is, for example, the attached cost function. In many cases there can be hidden costs to making more realistic (which often means longer) estimates, for example, in a competitive setting: if multiple companies compete for a contract to build an opera house, then a company claiming to be faster and cheaper might have better chances to win the contract. Such situations create *incentives to make inaccurate estimations*. Hidden incentives are also often at play in social settings, for example, when someone promises to finish a task someone else depends on, then they are more likely to provide an optimistic estimate to reassure the other person: at the time of the promise, there may be *just enough* time to finish the task in time, and there is no incentive for the planner to provide a more realistic estimate which would attract the other person's anger prematurely.

6.6 Demand Characteristics

As we used a within-subject design and the same instructions were repeated multiple times, participants may have understood what results we were hoping to collect and responded accordingly. This phenomenon is referred to as *demand characteristics* and expresses a participant's will to positively help a researcher by providing the responses they anticipate [61]. However, our analysis shows a difference between how respondents estimated for the second time (after the debiasing methods) and the third (after seeing the dot plot) in terms of prediction intervals (see section 5.2), that is, even if they guessed that we were expecting less optimistic estimates, they were not able to correct their second estimate to match the desired 95% prediction interval, but were closer to it the third time, using the visualization at their disposal. In other words, our data supports our assumption that computing how long a given task might take in one's head is a difficult exercise. The only nuance in our findings this can have an impact on is how the dot-plots are trusted: it is indeed possible that participants did not believe in their third estimate but chose nonetheless to put an answer that matched the visualization that was given. A possible way to mitigate demand characteristics could be monetary incentives: it has already been manipulated in a similar decision-making exercise requesting participants to provide a departure time, according to visualizations presenting signs of uncertainty, and with a reward based on simulated trips [16]. This method could be even more efficient if actual completion times were recorded.

6.7 Generalizability

In our study, we only used predictive visualizations in the form of 20- or 50-dot quantile plots. We have not tested whether different choices of data representations lead to different outcomes in the same context but chose this type of visualization based on past research which reported quantile dot plots as the most effective way to communicate uncertainty for decision-making in a transit scenario [44]. Our claim that predictive visualizations can reduce optimism and increase the overall sense of uncertainty in a time-related decision-making task depends on the choice of a visualization type capable of communicating distributions effectively to wide audiences.

7 IMPLEMENTATIONS

The online tool *Guesstimate* provides a spreadsheet-like interface to help people make estimates in uncertain settings using Monte Carlo sampling. As such, it already supports task duration estimates similar to our experiment (see this model for an example) but it requires some mathematical knowledge that can prevent lay people from using it. Moreover, the outputs are literal numbers (an interval and a point estimate) with an unlabeled distribution visible underneath while we used apparently labeled quantile dot plots to communicate the results with our participants.

The approach we described throughout this paper can easily be extended into a system for generic tasks, sub-tasks, and surprise events. We introduce a first version of such a tool in our online supplemental material available here[10].

[10]The full link is https://timeestimator.github.io.

In contrast to the tool used for our study, our time estimator includes a *live update* feature, that is, the visualization is constantly updated based on the currently set data, and the data can easily be adjusted by directly dragging over the numbers, using Bret Victor's *Tangle* JavaScript library created for his *Explorable Explanations*. This allows to explore more easily the influence of different parameters and enables planners to simulate different scenarios. Additionally, the use of such a tool on a regular basis would also enable training by feedback. As time and performance estimates have already been found to be enhanced by such a strategy [1, 64, 67, 73], a generic tool could enable a long-term or field study with the potential to advance our knowledge on multiple aspects including an evaluation of accuracy, a measurement of human biases, alternate belief inputs and different study design's strategies to reduce experimental biases.

Altogether, there remains a lot of room for devising strategies to mitigate the planning fallacy, within and outside our approach. The knowledge on this cognitive bias is still growing and should keep offering new opportunities for research.

8 CONCLUSION

With the goal to explore visualization in relation to duration estimation biases, we introduced a procedure involving two debiasing methods, namely task decomposition and surprise listing, which we combined with predictive visualizations in the form of quantile dot plots and linecharts. Our crowdsourced study's results provide compelling evidence that (1) quantile dot plots prompted changes in participants' task duration estimates, which were expressed by signs of reduced optimism and increased uncertainty, and that (2) feedback, either through a linechart or interactive text, helped improve decisions, expressed through a larger effect on reducing cost (fewer trains missed) than on increasing waiting time. While more work is needed to better understand people's planning behavior, our approach generated promising changes in participants' estimates and provides rich directions for future work.

ACKNOWLEDGMENTS

We thank Pierre Dragicevic, Petra Isenberg, Lijie Yao, Gilles Bailly, Oleksandra Vereschak, Elodie Bouzbib, Ignacio Avellino and Clara Rigaud as well as the rest of the HCI Sorbonne Team for their help and feedback. This work is part of the ANR Ember project, supported by a grant from Agence Nationale de Recherche (ANR-19-CE33-0012).

REFERENCES

[1] Pekka Abrahamsson and Karlheinz Kautz. 2002. Personal Software Process: Classroom Experiences from Finland. In *Software Quality — ECSQ 2002*, Jyrki Kontio and Reidar Conradi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 175–185. https://doi.org/10.1007/3-540-47984-8_21

[2] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. 2005. PlanningLines: novel glyphs for representing temporal uncertainties and their evaluation. In *Proceedings of the 9th International Conference on Information Visualisation (IV'05)*. 457–463. https://doi.org/10.1109/IV.2005.97

[3] Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating User Confidence for Uncertainty Presentation in Predictive Decision Making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (Parkville, VIC, Australia) *(OzCHI '15)*. Association for Computing Machinery, New York, NY, USA, 352–360. https://doi.org/10.1145/2838739.2838753

[4] Maya Bar-Hillel. 1973. On the Subjective Probability of Compound Events. *Organizational Behavior and Human Performance* 9 (jun 1973), 396–406. https://doi.org/10/btwqxt

[5] Marilyn G. Boltz and Yen Na Yum. 2010. Temporal concepts and predicted duration judgments. *Journal of Experimental Social Psychology* 46, 6 (2010), 895–904. https://doi.org/10.1016/j.jesp.2010.07.002

[6] Roger Buehler and Dale Griffin. 2003. Planning, personality, and prediction: The role of future focus in optimistic time predictions. *Organizational Behavior and Human Decision Processes* 92, 1 (sep 2003), 80–90. https://doi.org/10.1016/S0749-5978(03)00089-X

[7] Roger Buehler, Dale Griffin, and Michael Ross. 1994. Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times. *Journal of Personality and Social Psychology* 67, 3 (1994), 366–381. https://doi.org/10.1037/0022-3514.67.3.366

[8] Stephanie J. Byram. 1997. Cognitive and motivational factors influencing time prediction. *Journal of Experimental Psychology: Applied* 3, 3 (1997), 216–239. https://doi.org/10.1037/1076-898X.3.3.216

[9] Terry Connolly and Doug Dean. 1997. Decomposed Versus Holistic Estimates of Effort Required for Software Writing Tasks. *Management Science* 43, 7 (1997), 1029–1045. https://doi.org/10.1287/mnsc.43.7.1029 arXiv:https://doi.org/10.1287/mnsc.43.7.1029

[10] M. Correll and M. Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[11] Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24, 1 (2001), 87–114. https://doi.org/10.1017/S0140525X01003922

[12] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond.* Routledge.

[13] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern statistical methods for HCI*. Springer, 291–330.

[14] Pierre Dragicevic and Yvonne Jansen. 2014. Visualization-Mediated Alleviation of the Planning Fallacy. IEEE VIS 2014. https://hal.inria.fr/hal-01500560

[15] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. *Increasing the Transparency of Research Papers with Explorable Multiverse Analyses.* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300295

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

[17] Darryl K. Forsyth and Christopher D. B. Burt. 2008. Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & Cognition* 36, 4 (jun 2008), 791–798. https://doi.org/10/cqf6q2

[18] Björn Friedrich and Peter Heil. 2017. Onset-Duration Matching of Acoustic Stimuli Revisited: Conventional Arithmetic vs. Proposed Geometric Measures of Accuracy and Precision. *Frontiers in Psychology* 7, 2013 (2017). https://doi.org/10.3389/fpsyg.2016.02013

[19] Gerd Gigerenzer. 2004. Fast and frugal heuristics: The tools of bounded rationality. *Blackwell handbook of judgment and decision making* 62 (2004), 88.

[20] Gerd Gigerenzer. 2018. The Bias Bias in Behavioral Economics. *Review of Behavioral Economics* 5, 3-4 (2018), 303–336. https://doi.org/10.1561/105.00000092

[21] D. Gotz and H. Stavropoulos. 2014. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (dec 2014), 1783–1792. https://doi.org/10.1109/TVCG.2014.2346682

[22] Miriam Greis, Passant El. Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. 2016. Decision-Making under Uncertainty: How the Amount of Presented Uncertainty Influences User Behavior. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16).* Association for Computing Machinery, New York, NY, USA, Article 52, 4 pages. https://doi.org/10.1145/2971485.2971535

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

[24] Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. 2015. Investigating Representation Alternatives for Communicating Uncertainty to Non-experts. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 256–263. https://doi.org/10.1007/978-3-319-22723-8_21

[25] Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. 2017. Input Controls for Entering Uncertain Data: Probability Distribution Sliders. *Proceedings of the ACM on Human-Computer Interaction* 1, EICS, Article 3, 17 pages. https://doi.org/10.1145/3095805

[26] D. Griffin and R. Buehler. 1999. Frequency, Probability, and Prediction: Easy Solutions to Cognitive Illusions? *Cognitive Psychology* 38, 1 (1999), 48–78. https://doi.org/10.1006/cogp.1998.0707

[27] Tanja M. Gruschke and Magne Jørgensen. 2008. The Role of Outcome Feedback in Improving the Uncertainty Assessment of Software Development Effort Estimates. *ACM Transactions on Software Engineering Methodology* 17, 4, Article 20 (aug 2008), 35 pages. https://doi.org/10.1145/13487689.13487693

[28] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. 2018. EventThread: Visual Summarization and Stage Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 56–65. https://doi.org/10.1109/TVCG.2017.2745320

[29] C. Hadjichristidis, B. Summers, and K. Thomas. 2014. Unpacking estimates of task duration: The role of typicality and temporality. *Journal of Experimental Social Psychology* 51 (2014), 45–50. https://doi.org/10.1016/j.jesp.2013.10.009

[30] Torleif Halkjelsvik, Magne Jørgensen, and Karl Halvor Teigen. 2011. To read two pages, I need 5 minutes, but give me 5 minutes and I will read four: how to change productivity estimates by inverting the question. *Applied Cognitive Psychology* 25, 2 (mar 2011), 314–323. https://doi.org/10.1002/acp.1693

[31] P. R Hartl. 2008. *Visualization of Calendar Data.* Master's thesis. Vienna University of Technology.

[32] Barbara Hayes-Roth. 1980. Estimation of Time Requirements During Planning: Interactions Between Motivation and Cognition.

[33] Rebbeca A. Henry. 1994. The Effects of Choice and Incentives on the Overestimation of Future Performance. *Organizational Behavior and Human Decision Processes* 57, 2 (1994), 210–225. https://doi.org/10.1006/obhd.1994.1012

[34] Rebecca A. Henry and Janet A. Sniezek. 1993. Situational Factors Affecting Judgments of Future Performance. *Organizational Behavior and Human Decision Processes* 54, 1 (1993), 104–132. https://doi.org/10.1006/obhd.1993.1005

[35] Rebecca A. Henry and Oriel J. Strickland. 1994. Performance self-predictions: The impact of expectancy strength and incentives. *Journal of Applied Social Psychology* 24, 12 (1994), 1056–1069. https://doi.org/10.1111/j.1559-1816.1994.tb02373.x

[36] Richards J. Heuer Jr. 1999. Psychology of Intelligence Analysis. Washington, DC.

[37] Pamela J. Hinds. 1999. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied* 5, 2 (1999), 205–221. https://doi.org/10.1037/1076-898X.5.2.205

[38] Douglas W Hubbard and David Drummond. 2011. *How to measure anything.* Wiley Online Library.

[39] J. Jo, J. Huh, J. Park, B. Kim, and J. Seo. 2014. LiveGantt: Interactively Visualizing a Large Manufacturing Schedule. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (dec 2014), 2329–2338. https://doi.org/10.1109/TVCG.2014.2346454

[40] R. A. Josephs and E. D. Hahn. 1995. Bias and Accuracy in Estimates of Task Duration. *Organizational Behavior and Human Decision Processes* 61, 2 (1995), 202–213. https://doi.org/10.1006/obhd.1995.1016

[41] M. Jørgensen and T. M. Gruschke. 2009. The Impact of Lessons-Learned Sessions on Effort Estimation and Uncertainty Assessments. *IEEE Transactions on Software Engineering* 35, 3 (may–jun 2009), 368–383. https://doi.org/10.1109/TSE.2009.2

[42] Daniel Kahneman and Amos Tversky. 1982. *Intuitive prediction: Biases and corrective procedures.* Cambridge University Press, 414—-421. https://doi.org/10.1017/CBO9780511809477.031

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

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

[45] W. E. Kelly. 2000. *Conscientiousness and the prediction of task duration: Evidence of the role of personality in time prediction.* Ph.D. Dissertation. ProQuest Information & Learning, US.

[46] W. E. Kelly. 2002. Anxiety and the Prediction of Task Duration: A Preliminary Analysis. *The Journal of Psychology: Interdisciplinary and Applied* 136, 1 (2002), 53–58. https://doi.org/10.1080/00223980209604137 arXiv:https://doi.org/10.1080/00223980209604137

[47] William E. Kelly. 2004. College Students' Accuracy and Perceptions of Accuracy in Predicting the Duration of an Academic-Related Task. *Individual Differences Research* 2, 3 (2004), 225–230.

[48] Yea-Seul Kim, Paula Kayongo, Madeleine Grunde-McLaughlin, and Jessica Hullman. 2021. Bayesian-Assisted Inference from Visualized Data. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 989–999. https://doi.org/10.1109/TVCG.2020.3028984

[49] Kris N Kirby and Daniel Gerlanc. 2013. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods* 45, 4 (2013), 905–927.

[50] Cornelius J. Koch and Martin Kleinmann. 2002. A stitch in time saves nine: Behavioural decision-making explanations for time management problems. *European Journal of Work and Organizational Psychology* 11, 2 (jun 2002), 199–217. https://doi.org/10.1080/13594320244000120

[51] J. Kruger and M. Evans. 2004. If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology* 40, 5 (2004), 586–598. https://doi.org/10.1016/j.jesp.2003.11.001

[52] Falk Lieder and Thomas L. Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43 (2020), E1. https://doi.org/10.1017/S0140525X1900061X

[53] Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, and Noah D. Goodman. 2018. The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review* 25, 1 (feb 2018), 322–349. https://doi.org/10.3758/s13423-017-1286-8

[54] Saturnino Luz and Masood Masoodian. 2010. Improving Focus and Context Awareness in Interactive Visualization of Time Lines. In *Proceedings of the 24th BCS Interaction Specialist Group Conference* (Dundee, United Kingdom) *(BCS '10)*. BCS Learning & Development Ltd., Swindon, GBR, 72–80.

[55] Erik Løhre and Karl Halvor Teigen. 2014. How fast can you (possibly) do it, or how long will it (certainly) take? Communicating uncertain estimates of performance time. *Acta Psychologica* 148 (may 2014), 63–73. https://doi.org/10.1016/j.actpsy.2014.01.005

[56] W. McCown, T. Petzel, and P. Rupert. 1987. An experimental study of some hypothesized behaviors and personality variables of college student procrastinators. *Personality and Individual Differences* 8, 6 (1987), 781–786. https://doi.org/10.1016/0191-8869(87)90130-9

[57] George A. Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81–97. https://doi.org/10.1037/h0043158

[58] David E. Morris, Jeremy E. Oakley, and John A. Crowe. 2014. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software* 52 (2014), 1–4. https://doi.org/10.1016/j.envsoft.2013.10.010

[59] Ian Newby-Clark, Michael Ross, Roger Buehler, Derek Koehler, and Dale Griffin. 2000. People focus on optimistic scenarios and disregard pessimistic scenarios when predicting task completion times. *Journal of Experimental Psychology: Applied* 6, 3 (oct 2000), 171–182. https://doi.org/10.1037//1076-898X.6.3.171

[60] F Nguyen, X Qiao, J Heer, and J Hullman. 2020. Exploring the Effects of Aggregation Choices on Untrained Visualization Users' Generalizations From Data. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 33–48.

[61] Martin T. Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17, 11 (1962), 776–783. https://doi.org/10.1037/h0043424

[62] Alex T Pang, Craig M Wittenbrink, and Suresh K Lodha. 1997. Approaches to uncertainty visualization. *The Visual Computer* 13, 8 (1997), 370–390.

[63] Adam Perer and Fei Wang. 2014. Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2557500.2557508

[64] L. Prechelt and B. Unger. 2001. An experiment measuring the effects of personal software process (PSP) training. *IEEE Transactions on Software Engineering* 27, 5 (may 2001), 465–472. https://doi.org/10.1109/32.922716

[65] A. Rind, T. Lammarsch, W. Aigner, B. Alsallakh, and S. Miksch. 2013. TimeBench: A Data Model and Software Library for Visual Analytics of Time-Oriented Data. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (dec 2013), 2247–2256. https://doi.org/10.1109/TVCG.2013.206

[66] Michael Roy, Nicholas Christenfeld, and Craig Mckenzie. 2005. Underestimating the Duration of Future Events: Memory Incorrectly Used or Memory Bias? *Psychological bulletin* 131, 5 (2005), 738–756. https://doi.org/10.1037/0033-2909.131.5.738

[67] Michael M. Roy, Scott T. Mitten, and Nicholas J. S. Christenfeld. 2008. Correcting memory improves accuracy of predicted task duration. *Journal of Experimental Psychology: Applied* 14, 3 (sep 2008), 266–275. https://doi.org/10.1037/1076-898x.14.3.266

[68] Lawrence J. Sanna, Craig D. Parks, Edward C Chang, and Seth E. Carter. 2005. The Hourglass Is Half Full or Half Empty: Temporal Framing and the Group Planning Fallacy. *Group Dynamics: Theory, Research, and Practice* 9, 3 (sep 2005), 173–188. https://doi.org/10.1037/1089-2699.9.3.173

[69] Rafay A. Siddiqui, Frank May, and Ashwani Monga. 2014. Reversals of task duration estimates: Thinking how rather than why shrinks duration estimates for simple tasks, but elongates estimates for complex tasks. *Journal of Experimental Social Psychology* 50 (2014), 184–189. https://doi.org/10.1016/j.jesp.2013.10.002

[70] Herbert A. Simon. 1957. *Models of man; social and rational.* Wiley, Oxford, England. Pages: xiv, 287.

[71] David P. Strum, Jerrold H. May, and Luis G. Vargas. 2000. Modeling the Uncertainty of Surgical Procedure Times: Comparison of Log-normal and Normal Models. *Anesthesiology* 92, 4 (04 2000), 1160–1167. https://doi.org/10.1097/00000542-200004000-00035 arXiv:https://pubs.asahq.org/anesthesiology/article-pdf/92/4/1160/401694/0000542-200004000-00035.pdf

[72] J. Talbot, V. Setlur, and A. Anand. 2014. Four Experiments on the Perception of Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (dec 2014), 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320

[73] Simon Tobin and Simon Grondin. 2015. Prior task experience affects temporal prediction and estimation. *Frontiers in Psychology* 6 (jul 2015), 916. https://doi.org/10.3389/fpsyg.2015.00916

[74] Yaacov Trope and Nira Liberman. 2003. Temporal Construal. *Psychological review* 110 (jul 2003), 403–421. https://doi.org/10.1037/0033-295X.110.3.403

[75] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:https://science.sciencemag.org/content/185/4157/1124.full.pdf

[76] A Tversky and D Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211, 4481 (jan 1981), 453–458. https://doi.org/10.1126/science.7455683 arXiv:https://science.sciencemag.org/content/211/4481/453.full.pdf

[77] Bret Victor. 2011. Explorable Explanations. Online. http://worrydream.com/ExplorableExplanations/.

[78] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, Los Alamitos, CA, USA, 104–115. https://doi.org/10.1109/VAST.2017.8585669

[79] Emily Wall, Leslie M. Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. *Four Perspectives on Human Bias in Visual Analytics.* Springer International Publishing, Cham, 29–42. https://doi.org/10.1007/978-3-319-95831-6_3

[80] J. Waser, R. Fuchs, H. Ribičič, B. Schindler, G. Blöschl, and E. Gröller. 2010. World Lines. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (nov–dec 2010), 1458–1467. https://doi.org/10.1109/TVCG.2010.223

[81] Ryan Wesslen, Doug Markant, Alireza Karduni, and Wenwen Dou. 2020. Using Resource-Rational Analysis to Understand Cognitive Biases in Interactive Data Visualizations. arXiv:2009.13368 [cs.HC]

[82] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. 2011. LifeFlow: Visualizing an Overview of Event Sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1747––1756. https://doi.org/10.1145/1978942.1979196

[83] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is Also Evidence: Interpreting Dwell Time for Recommendation from Psychological Perspective. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. Association for Computing Machinery, New York, NY, USA, 989–997. https://doi.org/10.1145/2487575.2487663

[84] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology* 111, 4 (oct 2016), 493–504. https://doi.org/10.1037/pspa0000056

[85] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. 2017. Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In *Human-Computer Interaction – INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer International Publishing, Cham, 23–39.

[86] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of Influence on User Trust in Predictive Decision Making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312962