

Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making

Michael Fernandes¹, Logan Walls¹, Sean Munson¹, Jessica Hullman¹, and Matthew Kay²

¹University of Washington
Seattle, WA, USA
mfern, logan.w.gm, smunson, jhullman@uw.edu

²University of Michigan
Ann Arbor, MI, USA
mjskay@umich.edu

ABSTRACT

Everyday predictive systems typically present point predictions, making it hard for people to account for uncertainty when making decisions. Evaluations of uncertainty displays for transit prediction have assessed people’s ability to extract probabilities, but not the quality of their decisions. In a controlled, incentivized experiment, we had subjects decide when to catch a bus using displays with textual uncertainty, uncertainty visualizations, or no-uncertainty (control). Frequency-based visualizations previously shown to allow people to better extract probabilities (quantile dotplots) yielded better decisions. Decisions with quantile dotplots with 50 outcomes were (1) better on average, having expected payoffs 97% of optimal (95% CI: [95%,98%]), 5 percentage points more than control (95% CI: [2,8]); and (2) more consistent, having within-subject standard deviation of 3 percentage points (95% CI: [2,4]), 4 percentage points less than control (95% CI: [2,6]). Cumulative distribution function plots performed nearly as well, and both outperformed textual uncertainty, which was sensitive to the probability interval communicated. We discuss implications for realtime transit predictions and possible generalization to other domains.

Author Keywords

Uncertainty visualization; transit predictions; mobile interfaces; dotplots; cumulative distribution plots.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Mobile devices provide a way to quickly access quantitative predictions to inform everyday decisions. Predictive applications help people make quick decisions about what outfit to wear to suit the weather, how much time to allocate for a trip, or when to leave to catch a bus. People are aware of the potential for uncertainty when interacting with predictions in everyday domains like weather [17, 18, 16] or transit

[20]. However, many domains applications present quantitative predictions as point estimates of the most likely outcome, conflicting with users’ expectations and how events unfold in real life. A realtime transit application might predict a bus to come 10 minutes from now (a point estimate), but in reality there is uncertainty in this prediction: traffic might cause the bus to be late, location sensing error might mean the bus is actually closer or further away than predicted, and so on.

Communicating the uncertainty in a prediction—by conveying that outcomes other than the best point estimate are possible—can help people make better decisions in everyday situations. For example, when presented with uncertainty in a weather forecast, people make more economically appropriate decisions than those who receive weather forecasts alone [16], and a better understanding of uncertainty can also improve trust in a system [21]. However, for uncertainty information to help in everyday circumstances, it must be presented in ways that non-experts can understand. Displaying a probability distribution over possible bus arrival times may not necessarily improve people’s decisions, especially if they do not understand what is being represented or do not have time to incorporate it into their decisions. The design of uncertainty representations should also account for users’ needs to make quick, in the moment-decisions, such as when they glance at a mobile display [20]. Presenting too much information risks confusing people, rather than helping them make better decisions.

Prior work demonstrates that people can accurately extract probabilities relevant to realtime transit decisions from discrete outcome uncertainty representations called quantile dotplots [20, 34]. Quantile dotplots are particularly appropriate for space-constrained mobile predictive displays like bus arrival time applications, because they present an abstraction of a probability density plot that enables thinking about probabilities in terms of counts instead of areas (making it easier to answer questions like *what is the chance the bus will arrive 8 minutes from now or later?*). While *extracting* probabilities from these visualizations has been shown to be more precise [20], it is not known if quantile dotplots enable better *decisions* when compared to lower-fidelity representations of uncertainty such as intervals or text. For example, a simple text description that a bus has a high (e.g., 80%) chance of arriving 5 minutes from now or later may be all a user needs to make their decision. Different display types, which simplify uncertainty to different degrees, could vary in their effectiveness for supporting decision-making. Some may also be easier to learn to use

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI’18, April 21–26, 2018, Montreal, QC, Canada. Copyright is held by the owner/author(s). Publication rights licensed to ACM © 2018 ACM 978-1-4503-5620-6/18/04 \$15.00. DOI: <http://dx.doi.org/10.1145/3173574.3173718>

or adapt better to a range of circumstances—thereby better meeting the needs of users who rely on mobile predictions on a daily basis.

To address these questions, we present the results of a large crowdsourced online experiment in which we evaluated how ten different uncertainty representations affected people’s decisions in a realtime bus-catching scenario. Participants were given transit predictions and asked to decide when to catch a bus. The bus’s arrival then was simulated (via a random draw from a distribution) and the outcome of their choice displayed. To simulate decision-making with real-world stakes, we rewarded subjects based on events that occurred due to their decisions. For example, subjects could earn more money if they waited less time at the bus stop.

Our findings extend prior research on presenting uncertainty in everyday predictive systems in multiple ways. By comparing decisions made from various visual and textual representations of uncertainty to decisions made without uncertainty information, we address the question of **whether presenting uncertainty improves decision making over point estimates alone in a bus arrival time context**. We show that *uncertainty information can support more rational decisions despite the space constraints of mobile displays*. We find that uncertainty displays outperform no-uncertainty displays, with expected payoffs that are about 97% of the expected payoff under the optimal strategy, about 5 percentage points more than displays without uncertainty.

Our experiment included displays with a range of uncertainty representations: discrete outcome uncertainty displays (quantile dotplots), cumulative distribution function (CDF) plots, probability density function (PDF) plots, textual uncertainty displays, and intervals. This allows us to identify **whether higher fidelity uncertainty displays that allow people to better estimate probabilities also allow them to make better decisions in a bus arrival time context**. We find that several types of visualizations previously shown to support more precise probability *extraction*, including quantile dotplots [20] and CDFs [13], also lead to more optimal (i.e., accurate and consistent) *decisions*: these displays lead to expected payoffs that are higher than lower-fidelity and/or less perceptually effective representations, like text, intervals, or PDFs, by anywhere from about 1 to 5 percentage points. We find that CDFs, which have been suggested to be difficult for the public to interpret in past work [20, 13], perform better than PDFs and interval plots that are often thought to be simpler to interpret.

Quantile dotplots and CDFs may perform well because they allow more accurate estimation of probability intervals [20, 13]. If people can learn to correctly estimate uncertainty from these plots and correctly incorporate that uncertainty into their decisions, they *should* be able to make better decisions—but do they? We investigate how decision quality improves over time as a user calibrates their usage; that is, **if and how well people learn to more effectively incorporate uncertainty into their decisions by observing the outcomes of their prior decisions**. We examine how decision performance with those displays changes over time with displays of varying fidelity

and theoretical estimation accuracy. We find that those displays with the best theoretical estimation accuracy were also the best-performing displays: Decisions with these displays improved in mean decision quality by about 5 percentage points over the course of the study; decisions with these displays also became more consistent over time, having a reduction in standard deviation of about 4 percentage points.¹

Together, our results demonstrate that non-experts can learn to incorporate uncertainty into bus-catching decisions, making better decisions from a space-constrained display in a time-constrained context. We also demonstrate a methodology for assessing decision quality with uncertainty displays. Our results contribute some of the first (to our knowledge) evidence that displays that allow for accurate extraction of probabilities in a bus arrival context also lead to more optimal decisions, and that people can learn to use such displays over time. We also demonstrate that, by both raising average decision quality and reducing variance, transit uncertainty displays allow the majority of the population of users (not just the best and average users) to make better decisions, an important trait for displays designed for lay audiences.

BACKGROUND

Prior work has studied both the benefits and challenges to communicating uncertainty, and has developed various techniques aimed at communicating uncertainty in everyday contexts.

Challenges and Benefits of Communicating Uncertainty

People have well-documented tendencies to misinterpret uncertainty in systematic ways. A large body of research in judgment under uncertainty and promoted most famously by Tversky and Kahneman [32] shows, for example, how people have trouble understanding statistical principles that govern uncertainty like the relationship between sample size and variance [31]. In the context of presenting predictions for everyday decisions, uncertainty contributes ambiguity in predictions, which designers may perceive as confusing to users. As a result, designers may view communicating uncertainty in interfaces, especially in systems that are designed to support quick, everyday decisions, as daunting or even undesirable to users.

Various studies, however, have shown that communicating uncertainty can help people make better decisions than those made when uncertainty is not communicated (e.g., [18, 24, 29, 10, 19, 5]). In a scenario in which people were asked to make decisions that a city administrator would have to make when faced with potential inclement weather, people presented with information about uncertainty in the forecasts made better decisions than those not presented with that information [24].

Presenting only point estimates can give people a false sense of precision, leading them to believe an estimate or prediction is more precise than it truly is. For example, people have been shown to interpret measurements from body weight scales as overly precise [21]. In contrast, when people have a better

¹Since we consider increasing average decision quality and reducing variance in decision quality both to be important, we will generally refer to displays that improve either of these metrics to have better *performance*.

understanding of the uncertainty in their measured weight, their trust in the system improves, as users can better account for day-to-day fluctuations in weight [21]. In other contexts, such as transit, people understand that displayed predictions are uncertain and may wish they have access to information about that uncertainty to inform their decisions [20].

Techniques for Communicating Uncertainty

Concerns about overburdening the user have prompted some science communication experts to develop simplified, qualitative descriptions of uncertainty (e.g., using phrases like “very likely”, “likely”, “unlikely”, etc to describe how likely an outcome is [1, 10]). However, different people may interpret the same qualitative expression of uncertainty differently [33], and interpretations may also change depending on context [35].

More recent research has demonstrated that non-experts can understand and benefit from more expressive representations of uncertainty in the transit domain. One study compared a point estimate versus a gradient plot to communicate remaining range for an electric vehicle [19]. The gradient plot reduced driver anxiety about range as they completed a driving task. Another study presented people with gradient plots depicting train journeys and connections, including alternate connections [37]. Compared to other journey planning tools, users were better able to understand delays and their effects on the trip.

In the work closest to our study, Kay *et al.* [20] developed and evaluated visualizations depicting uncertainty in bus arrival time predictions. The visualizations were designed to be glanceable; in other words, to allow for quick in-the-moment decisions to be made. Inspired by prior work on the benefits of frequency framings for improving statistical reasoning from the fields of cognitive psychology [9] and visualization [8, 12], Kay *et al.* [20] developed a discrete outcome adaption of a PDF called a quantile dotplot. They found that dotplots had a 1.15x reduction in variance of people’s estimates of probability intervals compared to other plots tested (PDFs and stripe plots[6]). However, they did not evaluate the effect of these displays on decision-making (only on uncertainty extraction), and did not evaluate simpler encodings of uncertainty like intervals or textual uncertainty. It is possible that simple text descriptions of uncertainty, accompanied with numeric estimates to reduce ambiguity, might lead to better outcomes for some users. We extend Kay *et al.* [20] by evaluating a range of representations, including no-uncertainty displays, text and interval displays, discrete outcome displays, and continuous visualizations; and use incentives to evaluate decision quality.

While a few prior studies on uncertainty visualization designed realistic tasks in which subjects make decisions (e.g., [5, 23, 36]), relatively few evaluations of uncertainty visualizations take steps to incentivize subject decisions by simulating rewards and consequences after a decision is made (a few exceptions being [11, 18]). Without these incentives, subject decisions may not resemble the decisions that people outside of a controlled experiment would make. Many real-world decisions involving uncertainty include penalties not just for an “incorrect” decision, but also for precautionary actions. For example, taking an umbrella to work incurs the “cost” of

having to carry the umbrella all day and remembering to take it home. Additionally, in many real-world decision settings in which people make similar decisions on a regular basis, people have the ability to adapt their decision-making strategy based on the outcome of prior decisions. Someone using a mobile weather application, for example, may make changes to how they use uncertainty information based on whether they were caught without an umbrella in the rain in a prior situation. Experimental paradigms that do not provide concrete feedback on outcomes prevent subjects from such calibration.

Psychology researchers interested in how individuals make decisions from uncertainty in weather forecasts have applied a method that uses financial incentives (awarded through a utility function, as used in experimental economics [4]) to motivate subjects [26]. When applied to a given decision-making situation, a properly informed utility function enables evaluation of a decision based on the merits of its outcome relative to other decisions and their outcomes. Such an economic framework, in which subjects are informed of monetary rewards and penalties associated with different decision outcomes, also makes it possible to define an unambiguously “correct” decision in any given context in the form of the decision that maximizes payoffs. Joslyn and colleagues use this paradigm to evaluate how uncertainty affects the decisions that can be made around weather forecast [17, 24, 18]. We adapt a similar financial incentive framework to explore the effects of uncertainty displays on decisions around real time bus arrival data.

INTERFACE AND UNCERTAINTY REPRESENTATIONS

We adapted the interface of OneBusAway (OBA), a real-time bus arrival time application, for use in our study. We modified the interface to incorporate predictive distributions based on Kay *et al.* [20], represented using uncertainty displays drawn from the literature.

Adapting OneBusAway

We selected OneBusAway as a model for our experimental interface because of its proven effectiveness [7] and widespread use in the real world. OneBusAway is a open-source realtime transit application used in six US cities or regions and two other international cities. To use OBA, an individual opens the application and selects a nearby transit stop. The interface then displays predicted departures for that stop, including recent departures (Figure 1.A).

With the goal of customizing a design to maintain the general look and feel of OBA, we followed an iterative design process. We improved our designs to better aid realtime decision-making by integrating feedback gathered from guerrilla usability tests [15] conducted with colleagues, along with pilot runs of our experimental platform with more than 80 pilot participants. We used a think-aloud protocol during in-person feedback sessions as well as during in-person pilot runs of our experiment. During these think-alouds, pilot subjects explained their rationale for a decision, often giving insight into how a representation lead them to a certain misunderstanding. We revised each representation to minimize common misconceptions.

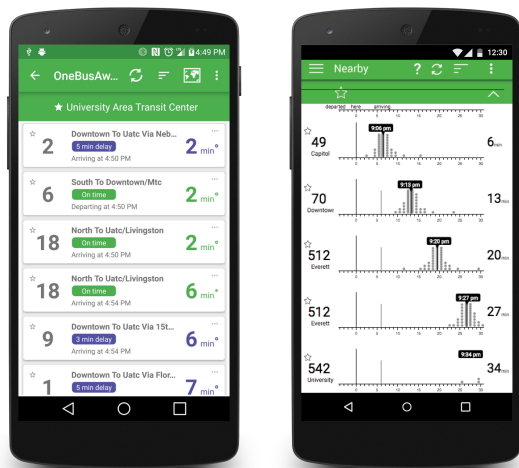


Figure 1. (A, left) Build of OneBusAway mobile application used for the Seattle Metropolitan Area as of September 2017. (B, right) Example of interface with a dot50 condition as it would appear to subjects.

We made the necessary modifications which allow for uncertainty representations to be displayed while preserving both the overall look and feel of the application and information displayed. (Figure 1.B shows how we modified the layout of the current OBA build. Figure 2 provides an example of the two variations tested in our experiment (arrival and no arrival)).

In our observations of how people use OneBusAway, we saw that people often scan the different entries, comparing different buses and deciding which they might take. However, in representations that present bus arrivals spatially on a timeline, this creates a problem: the information for many buses could appear off the screen to the right, and thus be unavailable to the viewer. To address this, we implemented a content-aware scroll animation [14] that shifts the time axis of each bus to the center of the screen upon scrolling. In other words, as the user scrolls to later buses, the time axis shifts to the left, centering predictions that occur further to the right in time on the timeline. In cases where even this automatic scrolling cannot display a prediction, the label for point predictions of later busses “stick” to the right edge of the screen (as seen in the bottom row of Figure 1.B). This ensures all predictions are available to the user in some form, and facilitates comparisons among buses without reducing the horizontal scale of the timeline to unreadable sizes.

From feedback gathered in our pilot experiment, we choose to omit annotations informing a rider how late or early a bus is relative to its scheduled arrival time. In early runs of our pilot experiment, we found pilot subjects were using “late” annotations incorrectly in their decision-making process. Late annotations in OBA inform a rider on how late or early a bus may arrive compared to the original scheduled arrival time of the bus. However, riders cannot accurately determine if a bus will come earlier or later based off the late annotation alone because the factors that make a bus late will differ from situation to situation. Similar to the real-world, the late annotations in our interface did not encode any accurate uncertainty information not already accounted for in the predictive distribution (the predictive distributions, which are based on models

from Kay *et al.*, already account for uncertainty caused by a bus being later than its scheduled time). In our pilots, some participants indicated that they would treat these annotations as some sort of indicator of uncertainty. Thus we removed the late annotations to reduce noise in our measurements caused by some participants misusing this cue.

Bus Arrival Time Predictive Distributions

To develop displays of probabilistic predictions of bus arrival times, we needed a model of probabilistic bus arrival times to ensure our approach was effective on realistic-looking predictions. We adopted the model developed by Kay *et al.* [20] (described in more detail in their supplemental materials). They collected bus arrival data and OBA predictions from the Seattle Metropolitan area, and fit a Box-Cox t regression model [25] to them. From their model, we generated a set of distributions resembling “typical” bus arrival predictions (not overly narrow or overly wide; see supplemental material). Each predictive distribution has a single most probable arrival time (the mode) a minimum of 5 minutes from “now” (0) a maximum of 25 minutes from “now”.

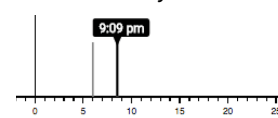
Uncertainty Displays

We developed a set of experimental conditions representing different uncertainty displays that allowed us to ask, at a high level, if uncertainty information improves decision-making about bus arrival times, how expressive should a representation be, and what framing of uncertainty (e.g., discrete outcome, interval, etc.) is most effective. Our no uncertainty condition represents the status quo point estimate depiction used in many current transit systems.

We choose display types based on prior work, described below. Additionally, prior work on communicating uncertainty in OBA found that while bus riders want to see uncertainty information, they also still want to see point estimates [20]—point estimates may aid *glanceability* in the quick decision-making context of realtime transit. Consequently, we designed two versions of each representation: one which displayed a point estimate (most probable arrival time of a bus) and another which did not (Figure 2).

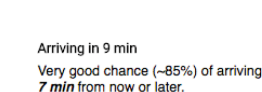
We discuss the rationale for the inclusion of each representation below, presenting each condition in an order of increasing complexity that we believe they present to a real-time transit decision maker.

No-uncertainty



This display type represents the status quo: it is informationally similar to the existing OneBusAway app, except we did not include annotations for how late or early a bus’s expected arrival time is relative to its scheduled arrival time (for the reasons described above).

Textual Predictive Intervals: 60%, 85%, and 99%



Compared to visual representations of probabilistic estimates, natural language representations provide a condensed illustration of a distribution in a possibly easier- (and faster-) to-digest form than

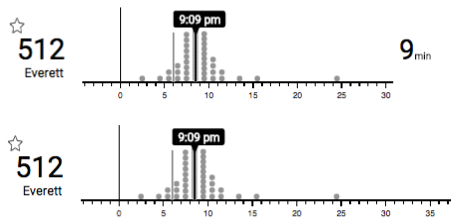
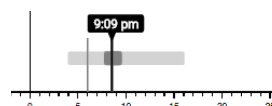


Figure 2. Single row of a Dot50 quantitative prediction with a most probable time to arrival supplied (top) and without (below).

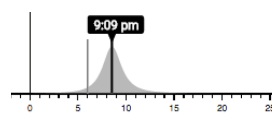
visualizations. Such representations may also be less susceptible to *deterministic construal errors*, that is, misinterpreting uncertainty as representing some other concept [27]. For example, in weather forecasting, misinterpreting the lower end of a predictive interval for a daily high temperature as the daily low for that day. Reducing an entire distribution to a single one-sided prediction interval (here, the probability a bus arrives at a certain time or later) necessarily reduces the applicability of the prediction. From person to person and scenario to scenario, a different probability level may represent the more optimal decision or more appropriate risk tolerance threshold. Thus, we included three textual displays to test the sensitivity of people’s decisions to the interval used: a 60%, 85%, and 99% predictive interval. For each textual representation, we choose not to include the timeline included in visualizations, instead expressing the most probable arrival time in natural language in order to match the form of the representation.

Interval Plot



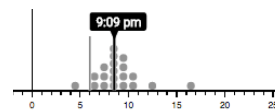
Interval plots are perhaps the most common uncertainty visualization. We tested interval plots to see if the familiarity of the visualization allowed people to easily understand the representation and make decisions accurately and efficiently. In the use case of real-time transit, a predictive interval allows a viewer to understand a range of plausible arrival times for a bus and the chance it will arrive during this range. Interval plots provide a quick understanding of the shape of a distribution without, perhaps, giving too much extra information. Our representations plotted the 50% and 95% quantile (equi-tailed) predictive intervals from the most probable time of arrival for a bus.

Probability Density Plots



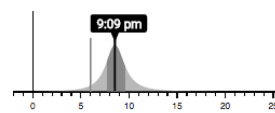
Probability density function (PDF) plots use an area encoding for probability that allows people to understand the shape of a distribution at a glance, and to estimate intervals by ratios of areas if desired, though such estimates are not the most accurate [13, 20]. Although other plots, like quantile dotplots, outperform PDFs for estimating probability intervals [20], participants in Kay *et al.* found PDFs more aesthetically pleasing and some specifically requested PDFs in interviews [20]. PDFs are also a common uncertainty visualization used in a wide variety of contexts. Following guidelines described in Kay *et al.* [20], our display normalized the height of probability density visualizations to be at most the height of each prediction’s row, so that very narrow distributions could be viewed without being clipped.

Quantile Dotplots



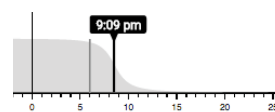
Kay *et al.* [20] introduced quantile dotplots, which are discrete analogs to the common probability density plot based on Wilkinson dotplots [34]. They found that quantile dotplots allowed people to more precisely extract probability intervals than other common uncertainty visualizations, such as stripe plots and PDFs, and speculated that this was because quantile dotplots allow people to understand the shape of the distribution while also enabling them to estimate probability intervals through counting. When the interval of interest is near the edge of a distribution, they speculated that counts are processed quickly through *subitizing* (i.e., quickly recognizing small numbers of items). In previous work, people’s performance degraded to that of PDFs when dotplots with higher numbers of dots were used on a space-constrained screen [20], lending evidence to this speculation. Our experiment tested low density quantile dotplots, displaying 20 or 50 dots. We refer to quantile dotplots with 20 dots as *dot20* and quantile dotplots with 50 dots as *dot50*.

Probability Density and Interval Plot



We include a hybrid PDF-interval plot to provide a condition that allows a user to selectively attend to the level of detail appropriate to their needs for a particular situation. PDF-interval plots combine the affordances of both probability density functions and interval plots. Our experiment tested a PDF-interval density function hybrid taking the shape of the density function and marking the central 50% interval within the shape.

Complementary Cumulative Distribution Plot



Outside of the transit domain, prior work in uncertainty visualization found CDFs to be effective for conveying some probabilities to the public, but could be confusing, particularly if the target attribute to be estimated was a mean [13]. Prior work in the bus domain has not tested the feasibility of cumulative distribution functions, but found through interviews that multiple bus riders indicated that there may be value in using CDFs to communicate uncertainty [20]. We tested both the CDF and the complementary CDF (CCDF) during our iterative design process. We selected the CCDF, since it better corresponds to people’s primary question: if I arrive at the bus stop at a certain time, how much of a chance to do I still have of catching the bus? If I delay, how does that affect my chances of catching the bus?

ONLINE EXPERIMENT

To evaluate how well our uncertainty displays support realistic decision-making from predicted bus arrival times, we conducted an online experiment, with the goal of answering three high-level questions in a mobile bus arrival time prediction context:

- How do decisions made from uncertainty information compare to those made without uncertainty information?
- How do different uncertainty representations (e.g., discrete versus continuous representations of probability, displays of varying expressiveness, text, intervals) compare for decision-making?

- How do different uncertainty representations compare for supporting effective decisions over time?

Experimental Design and Procedure

We used a between-subjects repeated measures design. Subjects in our experiment complete multiple trials, each of which represents a single decision about when to arrive at the bus stop given a realistic scenario (described below). The scenarios describe rewards for outcomes like catching the bus or spending time on a worthwhile activity rather than waiting for the bus, and penalties for wasting time at a bus stop or missing the bus (thus missing an important event). After each trial decision, the subject is presented with an outcome (e.g., “you arrived at the bus stop after 5 minutes; the bus arrived 1 minute after you and you caught the bus”) and informed of the rewards or penalties for their decision. The experimental procedure is depicted in Figure 3. Uncertainty display type and scenario were varied between subjects using random assignment.

To provide realistic bus arrival data for the quantitative predictions used, we generated a set of probability distributions from the Box-Cox t model described above, presenting them randomly in each trial. In our version of the OBA interface, the distribution for the trial would appear at the top of the list of possible arriving buses and was highlighted as the route in question for the trial. Before settling on the final setup for our experiment, we piloted several different experimental setups. In particular, we looked at the effect of trial length on decision making, finding that 30-40 trials appeared to allow subjects time to learn how to use the display while not requiring more trials than necessary².

During the experiment, we collected quantitative data including: the time it took to complete each trial, when the subject decided to arrive at the bus stop, and the reward given for each trial. In addition, every 9 experimental trials, subjects gave a rationale on why they decided to arrive at a bus stop at the specified time. We used the rationales mentioned previously to gauge how and why each display effectively or ineffectively aided its user at making decisions. Upon finishing the experimental phase of our study, subjects filled out surveys gathering demographic data and testing their understanding of their assigned display type.

Bus Arrival Scenarios

We constructed scenarios with the goal of creating realistic decision contexts that bus riders would find familiar. Each scenario describes a situation where subjects make a decision of when they will arrive at the bus stop in order to maximize the number of “coins” they receive (coins are translated into a bonus in their compensation for participating). Subjects gain coins for every minute they are able to continue an activity that is valuable to them (e.g., watching TV at home) before going to the bus stop, and gain a bonus for arriving at their intended destination early. Subjects incur a coin penalty for time spent waiting at a stop for a bus to arrive. In 40 successive trials on the same scenario, subjects see an arrival prediction and then specify in how many minutes from the present moment they would choose to arrive at the stop to catch the bus.

²Pilot experiment analysis can be found in our supplementary materials in “pilot_exploration.pdf”

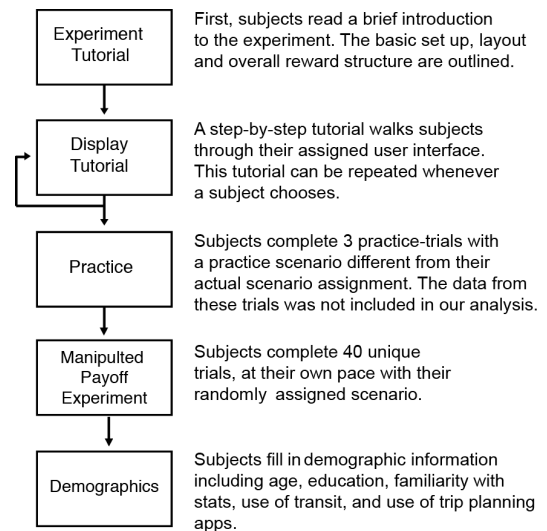


Figure 3. The general flow each subject progressed through when completing our experiment.

Each trial presents a slightly different predictive distribution. Abbreviated versions of the scenarios are:

- 1. Brunch With Friends** It is Sunday morning and you are watching your favorite television show before leaving to catch a bus to meet friends for brunch. You earn 8 coins for every minute spent watching TV at home and 14 coins for every minute spent at brunch. You lose 14 coins for every minute spent waiting at the bus stop.
- 2. Sunday Festival** It is Sunday evening and you are enjoying a festival before catching a bus to return home to rest before the busy work week. You earn 14 every minute spent at the festival and 14 coins for every minute spent at home. You lose 14 coins for every minute spent waiting at the bus stop.
- 3. Sunday Museum** It is a Sunday evening and you are at a new museum exhibit. It’s raining and you must catch a bus to return home to rest before a busy work week. You earn 8 coins for each minute remaining at the museum and 17 coins for each minute spent at home. You lose 17 coins for each minute spent waiting at the bus stop for the bus.

Rewarding Decisions: Utility functions

To improve the likelihood that subjects in our experiments engaged in realistic decision-making, we adapted a paradigm commonly used to study decision-making in behavioral economics and related fields [4, 29, 10]. In this paradigm, subjects make decisions and are financially rewarded or penalized (as they would be in real decision contexts) according to a utility function. For further realism, we derived the utility function that determines the payoffs and penalties by soliciting valuations of the costs and penalties of various outcomes from bus riders.

To assign relative coin values to designated behaviors within each scenario mentioned previously, a preliminary evaluation survey asked 10 regular bus riders to provide their monetary assessments of how much they would have to be compensated to be increasingly late for an event. In this survey, we used

similar scenarios as described above and asked subjects how much they would need to be paid to be a certain number of minutes late for the event. We found that responses were roughly linear: that is, as a subject arrived increasingly late (or missed the event altogether), their monetary assessment increased in proportion to the time. Using a linear regression, we created normalized values for how an "average" person valued their time in the survey.³

Decision Metric: expected/optimal payoff

We derived two measures from the results of our experiment to assess decision-making quality on a per-trial basis. The first measure, *expected payoff*, is the average payoff a subject could expect to receive based on their response on a given trial. This is the expected value of their payoff: the average of the payoff for all possible outcomes given the choice a participant makes, weighted by the probability of each outcome. Put another way, this is the average payoff a participant would get if they always made that choice under the same circumstances. We use this measure in place of the payoff a participant actually receives because actual payoff is affected by the random draw from the particular distribution for each trial. Using the expected value avoids measurement error caused by this random draw, allowing us to better assess how well a person's choices would do on average.

The second measure we used was *optimal payoff*: the *expected payoff* under an optimal strategy. This was derived by finding the maximum expected payoff possible for a trial over the space of possible response times in the experiment: 0 minutes to 30 minutes.

We used the ratio of these measures as a metric of decision quality: *expected/optimal payoff*. An *expected/optimal payoff* that is closer to 1 corresponds to a better decision, or one that is more rational [28] in terms of the payoffs for that scenario.

It is worth noting that the expected payoff in our experiment was not a monotonic function of the time chosen—arriving as early as possible was not always the correct decision (due to penalties for waiting at the stop). As chosen arrival time increases, eventually the chance of missing the bus becomes so high that expected payoff drops again. For realism, we included a backup bus in each scenario guaranteed to arrive after any time the user could have chosen, so expected payoff does move upwards again if the user arrives so far after the first bus that they may be close to catching the bus after it without waiting long (though never reaches the expected payoff of arriving before the first bus, since catching the backup bus guarantees the user to be late to their destination).

Pilot Results, Model, and Pre-registration

To calibrate our experimental parameters and to determine an analysis plan, we piloted our experiment on Mechanical Turk with over 80 pilot participants. We conducted an exploratory analysis on those results, and found that *expected/optimal payoff* tended to bunch towards 1 with distributions resembling beta distributions. Therefore, we adopted a beta regression model, which is suitable for outcomes on a scale from 0 to 1

³See "valuation_analysis.pdf" in our supplementary material.

[30]. Because some values of *expected/optimal payoff* may be exactly 1 (when the participant makes the optimal choice), and beta regression requires outcomes on (0, 1) exclusive, we use a correction for values equal to 1 [3]: we replace those values with $\text{expected payoff} / (\text{optimal expected payoff} + 1)$.

After Kay *et al.* 2016 [22], we use a mixed-effects Bayesian regression model with weakly-informed priors. We include *display type*, *trial*, and their interaction as fixed effects: this allows different visualization types to have different learning curves. We include a random intercept for *participant*, as well as a random slope of *trial* for each *participant*: this allows different participants to have different learning rates. We also include a random intercept for each *scenario*, assuming that different scenarios have different baseline difficulties (we did not find evidence for different learning rates by scenario in our pilot data, so we did not also include a random slope by *trial* here). Finally, we also allowed the precision parameter of the beta distribution, ϕ , to vary according to fixed effects of *display type*, *trial*, and their interaction: in other words, not only could average performance improve over time, but the variance of performance could also improve over time. We developed this model through exploratory analysis of pilot data. We also used pilot data to determine experimental parameters like number of trials and base pay. We pre-registered our Bayesian regression model and priors using AsPredicted (<https://aspredicted.org/iv7jb.pdf>) before collecting and analyzing our final dataset.

Our experiment was devised in such a way to test not only the effect of each condition but also the effect of including a point arrival time next to probabilistic distributions; see Figure 2 top (with point *arrival*) and bottom (*no arrival*). However, in our pilot results, we found that the two variants had similar performance and similar learning curves. Therefore, our pre-registered (and final) analysis pools the *arrival* and *no-arrival* variants of each display. We still collected data on these two variants (and include them in our supplemental data) in order to fuel exploratory analysis for future studies.

Participants and Compensation

After multiple rounds of piloting, we deployed our experiment on Amazon Mechanical Turk. Our experiment included only Master Turkers (i.e., 1,000 HITs completed with a 99.9% approval rate) and reside in the United States. Turkers could complete our experiment once. Each participant was compensated a base rate of \$1.25 for completing a HIT as well as a bonus compensation proportional to the coins they had earned throughout their experimental trials, which ranged from \$0.96 - 2.4 per participant depending on the scenario's conversion rate. Our final analysis included participants who completed at least 31 out of 40 trials, as our pilot analysis showed learning curves began to level off around 30 trials.

RESULTS

The dataset we used for analysis includes the work of 408 individuals⁴. 385 of those subjects submitted complete demographic information at the end of the experiment. Participants

⁴See "final_analysis.md|Rmd|html" in our supplementary material for the complete final analysis, and "data/final_trials.csv" for the

completed the experiment in an average of 10 minutes. 45% of participants were women and the median age was 35 years.

Most Conditions Exhibit Learning

Figure 4 shows the results of our model of expected payoffs as a proportion of the optimal payoff (*expected/optimal*). Our model allows us to make marginal posterior predictions: predictions for how a random subject in a random scenario will perform (Figure 4.1). For example, looking at the posterior predictive intervals (PPIs), the model predicts that by the last trial in *dot50*, about 50% of decisions will be above 95% of optimal (dark blue band), about 80% of decisions will be above 90% of optimal (lighter blue band), and more than 95% of decisions will be above 80% of optimal (lightest blue band). Most other conditions exhibit learning like this, although none result in as consistent decisions as *dot50* does. In some conditions, such as *text85* and no-uncertainty, the learning curve is quite flat, and many decisions in the final trials are still likely to be of lower quality (lower than 70% of optimal in interval, *text85*, and none, judging by the 95% PPIs).

While these posterior predictions can be more straightforward to interpret, it is important to look at the uncertainty in the model that drives these predictions. Figures 4.2 and 4.3 do this, showing quantile credible intervals around the conditional mean and standard deviation for each trial. In contrast to the marginal posterior predictions, which average over the random effects for participant and scenario, these estimates are conditional: they show the mean and standard deviation of *expected/optimal* for a "typical" participant and "typical" scenario. These estimates show that learning improves both the average performance and the standard deviation of performance: people's decisions are more consistently rational, particularly in *dot50*, *cdf*, and *dot20* conditions, with mean *expected/optimal* reaching around 95% of optimal (Figure 4.2) and standard deviation of *expected/optimal* reaching around 4 percentage points (Figure 4.3). Textual uncertainty conditions vary: *text99* and *text60* reach nearly similar performance, while *text85* exhibits very little learning, and appears more comparable to the no uncertainty condition, where neither average performance nor consistency improves much.

Dotplots, CDFs Converge on Better, Consistent Decisions

It is difficult from Figures 4.2 and 4.3 to directly compare performance in the final trial across conditions. To more directly compare performance, we also look at estimated differences in the conditional means (Figure 5.1) and standard deviations (Figure 5.2) in the last trial. These differences are shown against control (no uncertainty) and *dot50* (the best-performing condition).⁵

By the final trial, *dot50*, *cdf*, and *dot20* out-perform most other conditions in both mean and standard deviation by a few percentage points (Figure 5). While this may seem a small effect, combining a higher mean with lower variance can result in much

more consistent decisions. *Dot50*, for example, has a standard deviation in the last trial that is around 4 percentage points or more less than *pdf*, *pdfinterval*, *text85*, and *control* (last panel of Figure 5), combined with a higher mean, this difference drives the fact that the spread of predicted decisions in the last trial is on the order of 20 percentage points less: compare the extent of the 95% PPIs in Figure 4.1 between those conditions.

Dot20 did not have as clear a difference in performance against the best-performing text conditions as *dot50* did, but still yielded more consistent decisions: e.g., average decision quality in *dot20* is similar to *text99* (mean difference = -0.1 percentage points, 95% CI [-2.2,2.0]), but *text99* had higher variance by about 1.3 percentage points (95% CI [0.2,2.6]). In other words, *dot20* likely supports better decisions for the worst-performing users compared to *text99*.

DISCUSSION & LIMITATIONS

Our work provides evidence that predictions of bus arrival times with uncertainty can be a more effective form of quantitative prediction than point estimates alone. Displays with uncertainty generally performed comparably to or better than our control condition (no uncertainty) in terms of mean decision quality and variance in decision quality as measured by *expected/optimal payoff*. Further, our results found that visualizations that lead to more accurate estimates—CDFs [2, 13] and low-density quantile dotplots [20] with 20 or 50 dots—also lead to the best decisions in a realtime transit prediction context. Perhaps due to the advantages of frequency-based representations for reasoning about uncertainty [9, 12], quantile dotplots with 50 outcomes led to better decisions than CDFs, if slightly.

Sensitivity of Text Displays to Probability Level

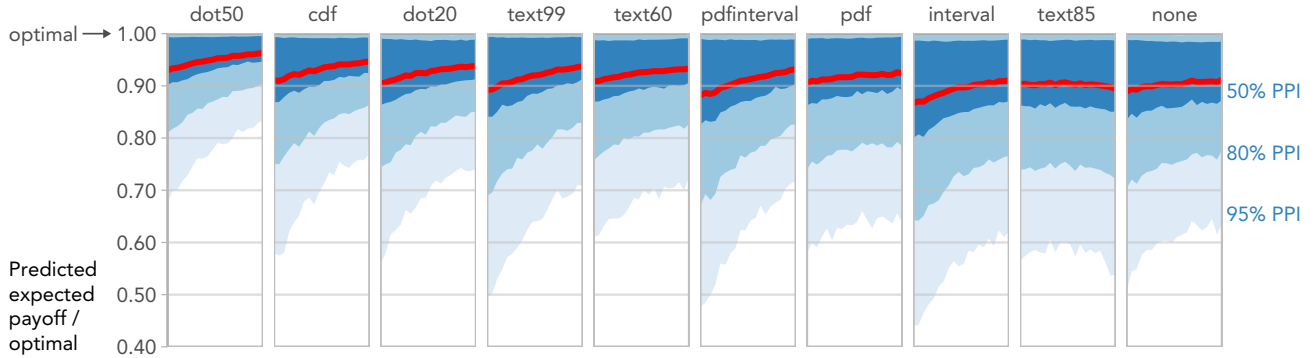
While low density quantile dotplots and CDF plots produced the most consistent and accurate decisions, *some* textual displays approached the same performance: *text99* and *text60* performed nearly as well as *dot20* (the lower-performing of the two dotplots). Meanwhile, decisions produced by *text85* were poor, and exhibited little learning. In other words, text decision quality was inconsistent (compare the text estimates in Figure 5): it appears that the effectiveness of textual displays is sensitive to the probability level displayed.

We believe this sensitivity is related to text displays' expressiveness. When a visualization is more expressive, allowing accurate estimates of various intervals, we expect people are more able to adapt the uncertainty information to new situations as the predictions (and their priorities) change. For example, the optimal choice in some situations may be to arrive at the bus stop at a time with very high probability of catching the bus (say 99%—perhaps in the case of an important meeting). In other cases, where the cost of missing the bus is lower, a reasonable gamble on missing the bus may be worth a shorter expected wait time (say 85%—like in the case of a casual lunch with friends). Because text intervals show only one interval estimate, people cannot adapt their decisions as easily to the costs and benefits of a particular situation. Thus, we have no reason to believe that *text99* or *text60* will always yield better performance than *text85*. This demonstrates a particular difficulty with text displays

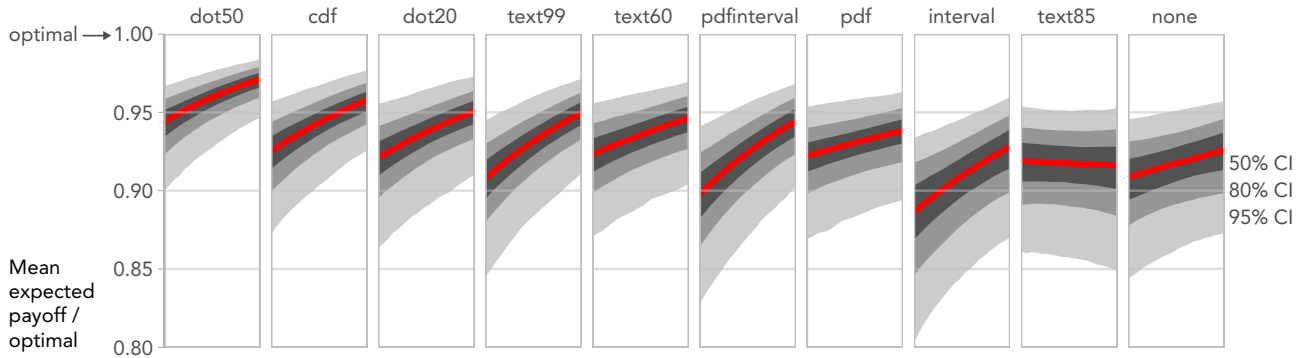
data. Our supplement is also available at <https://github.com/Michael-Fernandes/uncertainty-displays-for-transit> (DOI: 10.5281/zenodo.1136329)

⁵The interested reader is referred to "final_analysis.mdlRmdlhtml" in our supplementary material for a plot of all pairwise comparisons.

1. **Posterior predictive intervals** and predicted **mean** for performance in each condition. These intervals are what we would predict 50%, 80%, or 95% of new observations of performance to fall into. Performance improves with additional trials, especially for dot50, cdf, and dot20. Meanwhile, performance in text depends on the risk threshold, with text85 performing similarly to no uncertainty.



2. Quantile **credible intervals** (Bayesian analog to confidence intervals) and posterior median of the **mean** performance in each condition. These intervals show the uncertainty in the location of the red line in chart #1, above.



3. Quantile **credible intervals** and posterior median of the **standard deviation** of performance in each condition. Not only does mean performance improve, but variance in performance also improves over time: people get better on average and more consistent; dot50, the best-performing condition, achieves an SD in performance likely less than 4 percentage points by the final trial.

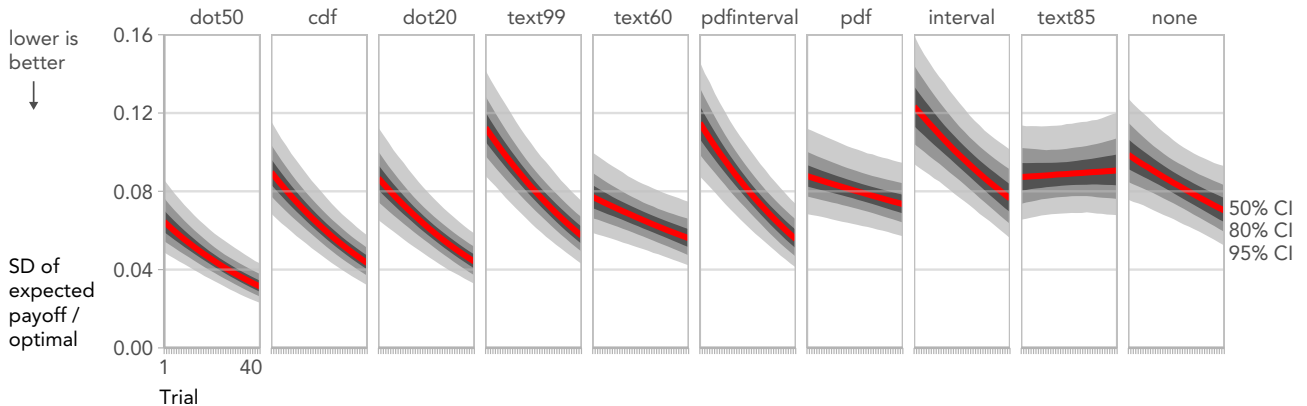


Figure 4. Fit lines from the model of the ratio of participants' expected payoffs to the expected payoff under an optimal strategy. Performance on dotplots and CDFs starts relatively high and becomes even better by the final trial: in these conditions, mean performance on the last trial is around 95% of optimal, and is very consistent: e.g. in *dot50*, more than 95% of predicted decisions in the last trial are more than 80% of optimal (look at the 95% PPI in subchart 1). Other conditions showed less consistency, and/or flatter learning curves.

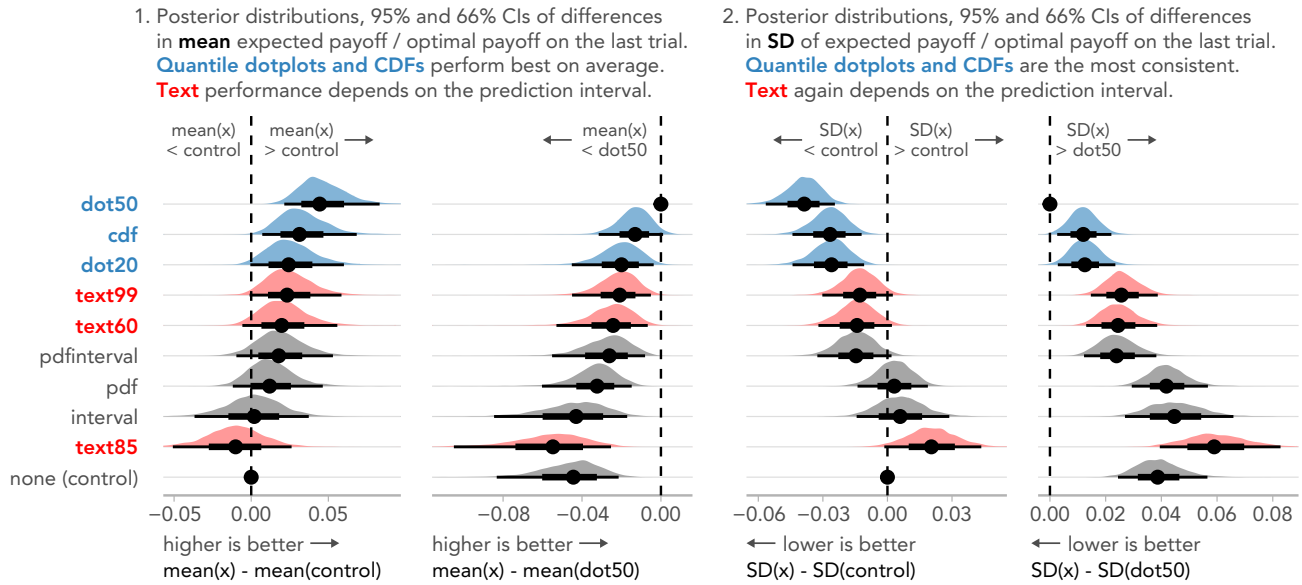


Figure 5. Posterior distributions of differences in mean and standard deviation of expected payoff / optimal payoff in the last trial. For both mean and standard deviation, we first show differences between each condition and control (no uncertainty), then show differences between each condition and the best-performing condition, *dot50*.

such as the ones tested: they may be simple but not sufficiently flexible to be adapted to new contexts.

Based on previous research on how people interpret probabilities close to 100% [38], people may also interpret 85% and 99% intervals as communicating nearly the same thing ('high chance'), possibly leading to overconfidence in *text85*. Meanwhile, the 60% interval was narrow (usually 1-3min left of the mode), giving less sense of the distribution shape but representing low certainty; perhaps users were more conservative to compensate for the low probability it represented.

Application to Other Domains

As previously mentioned, past work has shown that CDFs [13] and low density quantile dotplots [20] allow for accurate probability interval estimation. Our work expands on these findings about estimation to show that uncertainty representations that allow for more *accurate* probability estimation also produce higher quality (more accurate and consistent) *decisions*. We believe that this grounding in perceptual work on *estimation* (which is reasonable to expect to generalize across contexts) suggests that our findings about *decision quality* may also generalize to other contexts in which probability interval estimation is important and no single canonical interval is known to represent the best decision.

Limitations & Future Work

We designed the decisions that subjects made in our experiment to be similar to those made by bus riders when deciding to leave for a bus using a realtime transit prediction application. We encouraged subjects to make their decisions quickly, and rewarded better decisions using payoffs that were informed by real-world incentives. Our experiment used the same utility function for all participants. In the real world, each rider's utility function will be personal and change according to each

situation, though participants should still be motivated in our experiment to maximize their profit under the utility function given to them.

In the real world, when making decisions on-the-go, bus riders will have little time to reflect on the effectiveness of their decisions. Thus, although we based our scenarios on real world bus-catching situations, there is a need to study the effects of uncertainty displays on longitudinal decision-making in the wild. For example, using displays with well-expressed and well-calibrated uncertainty, instead of just point estimates, may make people feel complicit in bad decisions when they experience consequences of their decision like missing the bus [20]. Future real-world deployments could examine the effects of uncertainty displays on blame and decision satisfaction, in addition to decision quality.

CONCLUSION

In this paper, we demonstrate that including uncertainty displays in realtime transit decision-making can produce higher quality decisions. Using a method adapted from experimental economics, we successfully tracked serial decision making to evaluate what type of uncertainty displays produced the highest quality decisions. We found that cumulative distribution function (CDF) plots and low-density quantile dotplots produce more accurate and consistent decisions compared to other uncertainty visualizations, textual displays, and displays with no uncertainty. We also found that when using uncertainty displays, decision quality can improve over time. The types of displays that we have found to be the best for supporting decision-making in the transit have also been shown to be more accurate at estimating probability intervals more generally, suggesting that our results may generalize to similar situations where the same type of uncertainty information would help inform decision-making.

REFERENCES

1. Jessica S. Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B. Starren. 2006. Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association* 13, 6 (2006), 608–618. DOI: <http://dx.doi.org/10.1197/jamia.M2115>
2. William S. Cleveland and Robert McGill. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554. DOI: <http://dx.doi.org/10.1080/01621459.1984.10478080>
3. Francisco Cribari-Neto and Achim Zeileis. 2009. Beta regression in R. (2009).
4. Charles Holt D. Davis. 1996. Experimental economics : Douglas D. Davis and Charles A. Holt, (Princeton University Press, Princeton, 1993), pp xi+571. *Journal of Economic Behavior & Organization* 30, 3 (September 1996), 411–416. <https://ideas.repec.org/a/eee/jeborg/v30y1996i3p411-416.html>
5. Beverley J. Evans. 1997. Dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences* 23, 4 (1997), 409–422. DOI: [http://dx.doi.org/https://doi.org/10.1016/S0098-3004\(97\)00011-3](http://dx.doi.org/https://doi.org/10.1016/S0098-3004(97)00011-3) Exploratory Cartographic Visualisation.
6. Eric D. Feigelson and G. Jogesh Babu. 1992. *Statistical Challenges in Modern Astronomy*. Springer-Verlag, 155–157 pages. <http://www.springer.com/us/book/9781461392927>
7. Brian Ferris, Kari Watkins, and Alan Borning. 2010. OneBusAway: Results from Providing Real-time Arrival Information for Public Transit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1807–1816. DOI: <http://dx.doi.org/10.1145/1753326.1753597>
8. Rocio Garcia-Retamero, Mirta Galesic, and Gerd Gigerenzer. 2010. Do Icon Arrays Help Reduce Denominator Neglect? *Medical Decision Making* 30, 6 (2010), 672–684. DOI: <http://dx.doi.org/10.1177/0272989X10369000> PMID: 20484088.
9. Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102 (1995), 684–704.
10. Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural Language Generation enhances human decision-making with uncertain information. *CoRR* abs/1606.03254 (2016). <http://arxiv.org/abs/1606.03254>
11. Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. 2015. *Investigating Representation Alternatives for Communicating Uncertainty to Non-experts*. Springer International Publishing, Cham, 256–263. DOI: http://dx.doi.org/10.1007/978-3-319-22723-8_21
12. Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering. *PLOS ONE* 10, 11 (2015). <http://idl.cs.washington.edu/papers/hops>
13. Harald Ibrekk and M. Granger Morgan. 1987. Graphical Communication of Uncertain Quantities to Nontechnical People. *Risk Analysis* 7, 4 (1987), 519–529. DOI: <http://dx.doi.org/10.1111/j.1539-6924.1987.tb00488.x>
14. Edward W. Ishak and Steven K. Feiner. 2006. Content-aware Scrolling. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST '06)*. ACM, New York, NY, USA, 155–158. DOI: <http://dx.doi.org/10.1145/1166253.1166277>
15. James J. Chudley and Jesmond J. Allen. 2012. *Smashing UX Design: Foundations for Designing Online User Experiences*. Vol. 1st edition. Wiley.
16. Susan Joslyn and Jared LeClerc. 2013. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science* 22, 4 (2013), 308–315. DOI: <http://dx.doi.org/10.1177/0963721413481473>
17. Susan Joslyn and Sonia Savelli. 2010. Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorological Applications* 17, 2 (2010), 180–195. DOI: <http://dx.doi.org/10.1002/met.190>
18. S. L. Joslyn and J. E. LeClerc. 2012. Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error. *Journal of Experimental Psychology: Applied* 18, 1 (2012), 126–140.
19. Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. 2015. Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2201–2210. DOI: <http://dx.doi.org/10.1145/2702123.2702479>
20. 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)*. ACM, New York, NY, USA, 5092–5103. DOI: <http://dx.doi.org/10.1145/2858036.2858558>
21. Matthew Kay, Dan Morris, mc schraefel, and Julie A. Kientz. 2013. There's No Such Thing As Gaining a Pound: Reconsidering the Bathroom Scale User Interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 401–410. DOI: <http://dx.doi.org/10.1145/2493432.2493456>

22. Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4521–4532. DOI: <http://dx.doi.org/10.1145/2858036.2858465>
23. Michael Leitner and Barbara P. Buttenfield. 2000. Guidelines for the Display of Attribute Certainty. *Cartography and Geographic Information Science*, 27, 1 (January 2000), 3–14.
24. Limor Nadav-Greenberg and Susan L. Joslyn. 2009. Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making* 3, 3 (2009), 209–227. DOI: <http://dx.doi.org/10.1518/155534309X474460>
25. Robert A Rigby and D Mikis Stasinopoulos. 2006. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling* 6, 3 (2006), 209–229. DOI: <http://dx.doi.org/10.1191/1471082X06st122oa>
26. Mark S. Roulston, Gary E. Bolton, Andrew N. Kleit, and Addison L. Sears-Collins. 2006. A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting* 21, 1 (2 2006), 116–122. DOI: <http://dx.doi.org/10.1175/WAF887.1>
27. Sonia Savelli and Susan Joslyn. 2013. The Advantages of Predictive Interval Forecasts for Non-Expert Users and the Impact of Visualizations. *Applied Cognitive Psychology* 27, 4 (2013), 527–541. DOI: <http://dx.doi.org/10.1002/acp.2932>
28. John Scott. 2000. Understanding Contemporary Society: Theories of the Present. (2000). DOI: <http://dx.doi.org/10.4135/9781446218310>
29. Vlad V. Simianu, Margaret A. Grounds, Susan L. Joslyn, Jared E. LeClerc, Anne P. Ehlers, Nidhi Agrawal, Rafael Alfonso-Cristancho, Abraham D. Flaxman, and David R. Flum. 2016. Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey. *BMC Medical Informatics and Decision Making* 16, 1 (01 Dec 2016), 153. DOI: <http://dx.doi.org/10.1186/s12911-016-0391-3>
30. Michael Smithson and Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods* 11, 1 (mar 2006), 54–71. DOI: <http://dx.doi.org/10.1037/1082-989X.11.1.54>
31. Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin* 76, 2 (1971), 105.
32. Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. <http://www.jstor.org/stable/1738360>
33. T.S. Wallsten, R. Zwick, B. Forsyth, D.V. Budescu, A. Rappaport, and NORTH CAROLINA UNIV AT CHAPEL HILL. 1988. *Measuring the Vague Meanings of Probability Terms*. Defense Technical Information Center. <https://books.google.com.fj/books?id=TE6TNwAACAAJ>
34. Leland Wilkinson. 1999. Dot Plots. *The American Statistician* 53, 3 (August 1999), 276–281.
35. Paul Windschitl and Elke Weber. 1999. The Interpretation of 'Likely' Depends on the Context, But 70Perceived Certainty. 25 (12 1999), 1514–33.
36. Hao-Che Wu, Michael Lindell, and Carla Prater. 2015. Strike probability judgments and protective action recommendations in a dynamic hurricane tracking task. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards* 79, 1 (2015), 355–380. <http://EconPapers.repec.org/RePEc:spr:nathaz:v:79:y:2015:i:1:p:355-380>
37. Marcel Wunderlich, Kathrin Ballweg, Georg Fuchs, and Tatiana von Landesberger. 2017. Visualization of Delay Uncertainty and its Impact on Train Trip Planning: A Design Study. *Computer Graphics Forum* (2017). DOI: <http://dx.doi.org/10.1111/cgf.13190>
38. Hang Zhang and Laurence Maloney. 2012. Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. *Frontiers in Neuroscience* 6 (2012), 1. DOI: <http://dx.doi.org/10.3389/fnins.2012.00001>