# Interpreting Visualizations of Historical Variability for Estimating Future Events

Wayne Chi Wei GIANG and Birsen DONMEZ
*Human Factors and Applied Statistics Laboratory, Mechanical and Industrial Engineering, University of Toronto*

**ABSTRACT**

Estimations of variables like travel time are often important for scheduling and logistics decisions. When these estimations are made under high time pressure, visualizations of historical data can be used to help produce more accurate estimates and decisions. In this preliminary study, we examine four visualizations that represent increasing amounts of information about the dispersion and shape of the historical data and examine how these visualizations are used to produce estimates of task completion times. In particular, we are interested in whether the level of variability information provided causes the decision makers to systematically adjust their estimate away from the central tendency of the historical data. We found that participants were more confident and tended to deviate from the central tendency more when they had information about the range and shape of historical data compared to when they only had a point estimate or a point estimate and standard deviation.

**KEYWORDS**

*Planning and prediction; transportation; estimation; visualizations.*

## INTRODUCTION

Humans are often required to make decisions in situations that are characterized by large degrees of complexity and uncertainty that cannot be deterministically modelled. These situations often arise due to incomplete or uncertain information about the current or future state of the world, and thus decision makers must estimate and predict the variables that are critical for their decisions. Furthermore, the presence of time pressure increases the difficulty of these decisions and may lead to the use of heuristics and biases, which may not always be appropriate (Payne, Bettman, & Johnson, 1993; Tversky & Kahneman, 1974). Decision support systems providing historical data trends is one method to support evidence-based decisions that can help mitigate some of these difficulties.

An example scenario where decision makers are required to make short-term, time-critical, evidence-based decisions is medical dispatch. For example, medical dispatchers at Ornge, the medical transport system in Ontario, Canada, make their dispatch decisions by estimating time to definitive care, i.e., how long it takes to transfer patients between hospitals. However, in our previous work (Giang, Donmez, Fatahi, Ahghari, & MacDonald, 2014), we had identified that Ornge dispatcher estimates of time to definitive care tended to be shorter than actual times. Transfers are impacted by a variety of factors, such as traffic and patient condition. These factors are often hard for dispatchers to account for, and can cause transfers to deviate from normally expected times.

In response to these findings, a decision support tool was developed to provide the dispatchers with estimates based on historical transfer information that had less error than the dispatchers' own estimates. The tool provided point-estimates of transfer times calculated based on descriptive statistics and linear models. However, the point-estimates only communicated information on central tendency and fail to provide insight about the historical variability and uncertainty associated with these estimates. Ornge's dispatchers are expert decision makers who often have additional contextual information (e.g., knowledge about the crews involved or the weather) that they use to modify their own transfer time estimates. Visualizations of the variability of historical data may allow disaptchers to use this contexutal information in a way that is tied to historical data.

However, there has been little work done on how visualizations of historical data are interpreted for decision making. Uncertainty visualizations for dynamic decision making scenarios have typically dealt with providing classification information about objects instead of display information about the variability of continuous variables. For example, Neyedli, Hollands, and Jamieson (2011) developed and tested visualizations that showed the reliability of a system which detects friendly or enemy targets. Bisantz et al. (2011) developed visualizations of the uncertainty associated with object classifications in a missile detection game. In both these examples, the data that is being visualized is a classification of an object, the uncertainty or reliability information is a measurement of the likelihood of belonging to a category, and decision makers must use this information to make a judgment of the true identity of the object. However in applications such as medical dispatch, a critical part of logistics and coordination is the estimation of a specific time as opposed to a judgment of which category (e.g., late or not late) the time estimation belongs. Tasks such as scheduling ambulance arrival times, booking helipads, and letting staff at the receiving facility know when they should expect to receive a patient all benefit from having more accurate time estimates.

Full contact information for first author: Wayne Chi Wei Giang, wayne.giang@mail.utoronto.ca, Department of Mechanical and Industrial Engineering, Univerisity of Toronto, 5 King's College Road, Toronto, ON, M5S 3G8, Canada.

Presenting uncertainty information about continuous variables has shown mixed results in terms of performance and usage that appear to be highly tied to the method of presentation. Nadav-Greenberg and Joslyn (2009) examined verbal, numeric, and graphical representations of uncertainty information about nighttime temperature lows in a road-salting decision task. Participants were asked to predict the expected nighttime low, while making a decision about whether to salt the road if the temperature was expected to drop below freezing. They found that the uncertainty information helped participants make better decisions about road salting, and that the estimates of the nighttime lows were impacted by the type of information given (e.g., full range, probability of freezing). However, they only explored one graphical representation of uncertainty (for full range only) which they found to be not as effective in improving salting decisions in comparison to numerical representations. Similarly, Scown, Bartlett, and McCarley (2014) found that non-expert decision makers often did not use error bars when making two-point comparisons about product review scores. The benefits of visual representations of uncertainty information are often harder to study because individuals tend to construct different internal models of underlying probability distributions that are influenced by the graphical elements of the visualization (Tak, Toet, & van Erp, 2014). Thus, there may be factors that influence how visualizations of variability information are interpreted by decision makers, and these effects might be tied to the amount of variability information provided.

The goal of this preliminary study is to examine how the amount of variability information influences the way decision makers interpret visualizations of historical data. We examined four visualizations that represent increasing amounts of information about the dispersion and shape of the historical data: central tendency only, mean and standard deviation, boxplot, and violin plot. In particular, we are interested in examining whether the level of variability information provided will cause the decision makers to systematically adjust their estimate away from the central tendency (i.e., median, mode, or mean) of the historical data.

## METHODS

### Participants and Apparatus

We recruited 22 participants from the local community and the undergraduate and graduate population at the University of Toronto. Participants were selected using a screening questionnaire for completion of at least one probability or statistics course during their post-secondary education. Furthermore, all participants had normal or corrected-to-normal vision and normal colour perception.

Of the 22 participants, 13 were male and 9 were female. Participant ages ranged between 19 and 30 with a mean (M) of 24.5 years and a standard deviation (SD) of 3.0 years. Participants also reported taking an average of 1.8 probability or statistics courses during their post-secondary education (SD = 0.8), with 10 participants having taken these courses in graduate school and 12 participants at the undergraduate level.

The experiment was conducted in a quiet office environment. Participants were seated in front of a 24-inch monitor that displayed the experimental tasks. Participants responded to the tasks using a keyboard and mouse. The experimental software was created using the open-source PsychoPy framework (Pierce, 2007).

### Experimental Scenario and Task

An experimental scenario was created where the participants would not have any contextual information to draw from other than the information presented in the visualizations. Participants took on the role of a mission commander responsible to overseeing a number of scientific space rovers exploring a planet. The role of the mission commander was to monitor the amount of time required for a rover to complete a scientific task (e.g., collecting or analysing samples) in order to determine whether the rover would be able to stay on schedule with their upcoming tasks. Participants were told that each rover had their own set of historical data that represented the task completion times for that rover in the past, so that every rover should be treated independently.

Eight datasets were generated to serve as the historical data for the rovers. Each of these datasets was formed by sampling 50 data points from normal distributions with 4 different means and 2 levels of standard deviation for each mean. The four means used were 33, 56, 70, and 79, and the two levels of standard deviation were 10% of the mean and 30% of the mean. In addition, a "true" task time was also sampled from the distribution which represented the correct task completion time. The 8 datasets were presented using each of the 4 visualizations and replicated 4 times for a total of 128 trials per participant. Each of these trials represented a new rover that the participants had to monitor.

Participants were responsible for two tasks, similar to those used by Nadav-Greenberg and Joslyn (2009). In the first task, i.e., the estimation task, participants were required to estimate how long they thought it would take for the rover to complete its current scientific task. The participants selected a value using a slider scale that was superimposed on the uncertainty visualization, as shown in Figure 1. Estimates were restricted to integer values. In the second task, i.e., the judgment task, participants were asked to make a judgment about whether they felt that the rover was going to be able to complete its task by a specific cut-off time, also shown in Figure 1. Participants were also asked to rate their confidence in these two tasks on a scale between 1 and 100. During the experiment, participants completed trials containing both the estimation task and the judgment task, with the order of task presentation counterbalanced.

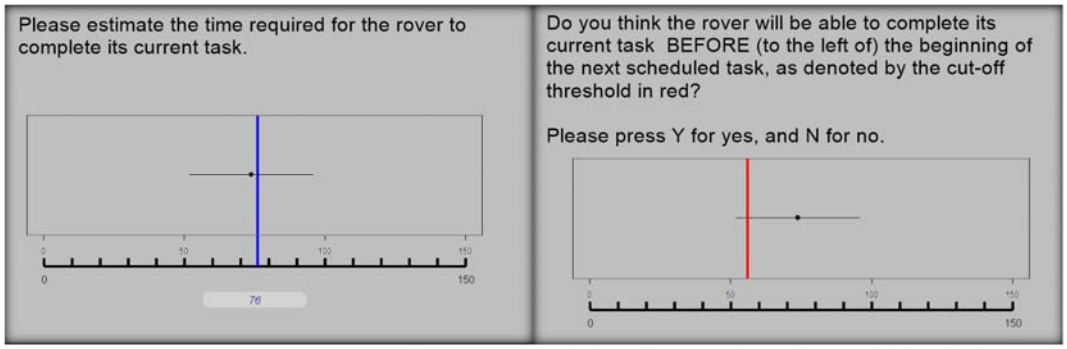


Figure 1: The estimation task with a slider bar (left) and the judgment task with the cut-off time (right).

## Experimental Design

The primary independent variable of interest was the type of visualization of the historical data, a within subject variable. Four visualization conditions (Figure 2) were designed for this experiment with increasing amounts of information about the dispersion and shape of the distribution of historical observations. In a baseline, central tendency only condition, only the median of the historical sample was displayed. The mean & standard deviation visualization condition showed both a measure of central tendency and a measure of dispersion but did not provide information about the shape of the historical sample. The boxplot visualization provided a measure of central tendency (median) as well as two measures of dispersion (interquartile range and range). The boxplot visualization also provided information about the skewness of the historical sample; the kurtosis of the historical sample could be inferred from the relative lengths of the box and whiskers. Finally, the violin plot visualization provided an indicator for the median, the interquartile range and range, as well as an estimate of the distribution of the historical sample (i.e., a kernel density estimate) which provides information of both skewness and kurtosis. All visualizations were generated using R, with the violin plots created using the 'vioplot' package. The x-axis in these visualizations represented minutes (task completion time), while the y-axis in the violin plot represented an estimate of the probability density of the historical sample.

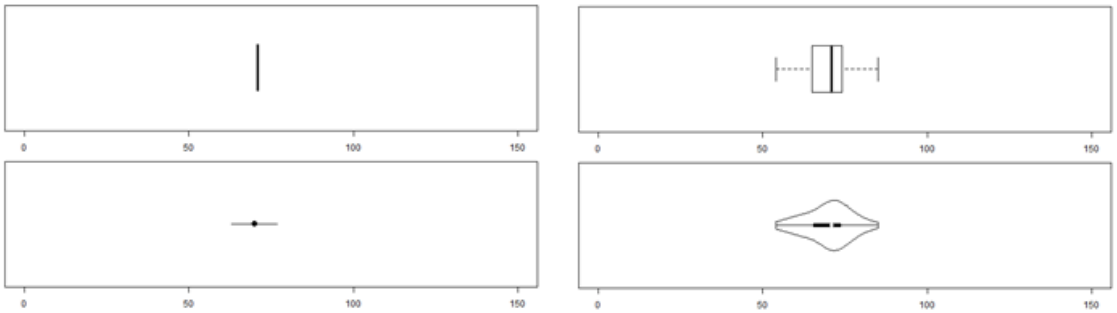

Figure 2: The visualizations of historical data: a) median only, b) mean & standard deviation, c) boxplot, and d) violin plot.

## Experimental Procedure

Before beginning the experiment, participants were given a review of the concepts of central tendency and dispersion, and an introduction to the visualizations used in the experiment. Participants were then provided with a short practice session of 4 trials for each of the visualization conditions. They were also told that there would be a $5 performance bonus during the experimental trials, although everyone was given this performance bonus.

In the first half of the experiment, participants were not given any feedback on their task performance. In the second half, participants were provided with feedback about the "true" task duration for that rover, and feedback about the correctness of their judgments after each trial. The experiment took approximately 90 minutes to complete, and participants were paid $20, which included the performance bonus.

## Data Processing

The analysis for this paper focuses on the no feedback trials. There were three dependent variables of interest with respect to the estimation of task times: 1) the distance between the participant's estimate and the central tendency expressed in number of standard deviations of the historical data, 2) the participant's confidence in the estimate rated between 1 and 100, and 3) the number of times the participant's estimate was not a central tendency point. The first dependent variable was calculated by using the closest measure of central tendency (i.e., mode, median, or mean) to the participant's estimate since each visualization type had a different prominently displayed central tendency measure, and for the violin plot, there were multiple features that the participant could use as a central tendency point. Since participants were restricted to responding in integer values, a correction was applied for the

third dependent variable: a response was counted to be on a central tendency measure as long as it fell within 0.5 units below or above it. Seven outlying trials, which had participant estimates greater than 2 standard deviations from the closest central tendency were removed. For all three dependent variables, values were averaged for each participant up to the visualization level.

## RESULTS

A linear mixed model was fitted to the distance to central tendency data (dependent variable 1) as a function of visualization type (median, mean & standard deviation, boxplot, and violin plot). Participant was used as a random factor and a square-root transformation was employed to ensure normality of residuals. Visualization type was found to be a significant factor, $F(3, 63) = 3.33$, $p = .03$. Post-hoc analyses using Tukey contrasts found that the violin plot ($\Delta = 0.074$, $p = .02$) had significantly larger distances to central tendency compared to the median only visualization, while the boxplot ($\Delta = 0.063$, $p = .06$) had only marginally significantly larger distances to central tendency than the median only visualization.

A second linear mixed model was fitted to the confidence data as a function of visualization type, with participant as a random factor. Visualization type was found to be significant, $F(3, 63) = 17.6$, $p < .0001$. Post-hoc comparisons using Tukey contrasts revealed that the median only visualization resulted in significantly lower confidence ratings than the mean & standard deviation ($\Delta = 5.7$, $p < .001$), the boxplot ($\Delta = 7.1$, $p < .001$), and the violin plot ($\Delta = 9.3$, $p < .001$) visualizations. In addition, the mean & standard deviation visualization also resulted in lower confidence ratings than the violin plot, $\Delta = 3.5$, $p = .04$.

Finally, a Poisson regression model was fitted to the third dependent variable (number of times the participant's estimate was not a central tendency point) with visualization type as a predictor variable. The number of trials completed by each participant per visualization (typically 16) was used as an offset variable, and participant was treated as a random factor. Again, visualization type was found to be a significant factor, $\chi^2(3) = 33.3$. $p < .001$. The median data values for this dependent variable for the four visualizations (i.e., median only, mean & standard deviation, boxplot, and violin plot) were 4.5, 5.5, 10, and 11, respectively. Post-hoc analysis using Tukey contrasts found that participants had a greater number of deviations from the central tendency with the boxplot ($p < .001$) and the violin plot ($p < .001$) visualizations when compared to the median only visualization. Similarly, the boxplot ($p = .003$) and the violin plot ($p < .001$) visualizations also resulted in a greater number of deviations than the the mean & standard deviation visualization.

## DISCUSSION AND CONCLUSION

These results suggest that the type of visualization used has an impact on how individuals make estimations through historical data. Participants did deviate away from the central tendency, and deviations were both more likely to occur and also to have larger magnitudes for the visualizations that provided more variability information (i.e., boxplot and violin plot). Participants' ratings of confidence were also higher for the boxplot and the violin plot visualizations. The central tendency points are the statistically optimal responses (since the scenarios were sampled from normal distributions), yet it appeared that the shape and range information led our participants to feel safe about estimating away from central tendency. It is also possible that participants had a harder time picking out a single measure of central tendency from the graphs containing more information, leading to the larger deviations from the central tendency measures that we calculated. Overall, this preliminary study suggests that participants do try to use variability information in adjusting their estimates of future events using historical data, and that further study of how contextual information might also influence their estimates is required.

## ACKNOWLEDGMENTS

We thank the Natural Sciences and Engineering Research Council and Ornge for their funding and support.

## REFERENCES

Bisantz, A. M., Cao, D., Jenkins, M., Pennathur, P. R., Farry, M., Roth, E., Potter, S. S., Pfautz, J. (2011). Comparing uncertainty visualizations for a dynamic decision-making task. *Journal of Cognitive Engineering and Decision Making*, *5*(3), 277–293.

Giang, W. C. W., Donmez, B., Fatahi, A., Ahghari, M., & MacDonald, R. D. (2014). Supporting Air Versus Ground Vehicle Decisions for Interfacility Medical Transport Using Historical Data. *IEEE Transactions on Human-Machine Systems*, *44*(1), 55–65.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making*, *3*(3), 209–227.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond Identity: Incorporating System Reliability Information Into an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 338–355.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. New York, NY: Cambridge University Press.

Scown, H., Bartlett, M., & McCarley, J. S. (2014). Statistically Lay Decision Makers Ignore Error Bars in Two-Point Comparisons. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1746–1750.

Tak, S., Toet, A., & Erp, J. Van. (2014). The Perception of Visual UncertaintyRepresentation by Non-Experts. *IEEE Transactions on Visualization and Computer Graphics*, *20*(6), 935–943.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–31.