Reporting and Visualizing Fitts' Law: Dataset, Tools and Methodologies

Alvin Jude
Ericsson Research
San Jose, CA, USA
alvinjude@acm.org

Darren Guinness
University of Colorado, Boulder
Boulder, CO, USA
darren.guinness@colorado.edu

G. Michael Poor
Baylor University
Waco, TX, USA
Michael_Poor@baylor.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892364

Abstract

In this paper we compare methods of reporting and visualizing Fitts regressions. We show that reporting this metric using mean movement time per user over accuracy-adjusted Index of Difficulty (IDe) produces more descriptive visualization. This method displays variance, which is more useful in understanding the interfaces, than an aggregated means-of-means approach using Index of Difficulty. We demonstrate that there is little difference in slope and intercept between the two methods, but has the potential to uncover wider goodness-of-fit coefficients which could allow for better comparison across experiments. We propose the use of quantile regression to report central tendencies as a trend, rather than box plots. The tools released with this paper can be used with any pointing device evaluation done with the FittsStudy program. The dataset released with this paper contains almost 25,000 samples, which can be used in future research for reporting or visualizing Fitts regressions.

Author Keywords

Fitts; Fitts's Law; Gestural Interaction;

ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces -- *Evaluation/methodology*;

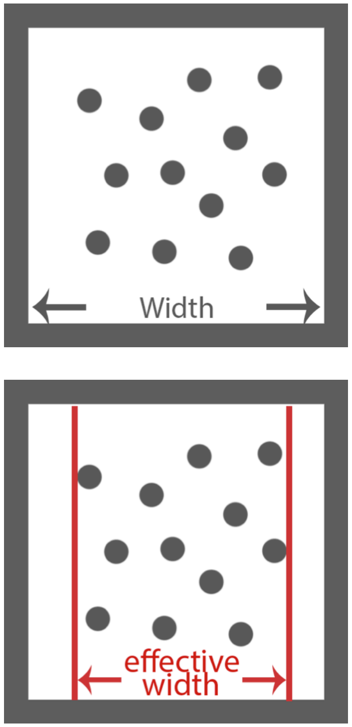

Figure 1 Both boxes in this image represent a target used for device evaluation, while the dots inside the target represent users end-points. The above image shows the actual width while the lower shows the intuition of the effective width (W_e). W_e requires all users end-points per task and is therefore measured post-hoc.

Introduction

The evaluation of pointing devices is largely built on Fitts's Law. It has evolved greatly in the 35 years since it was introduced in this domain. Much has been done since to allow reproducible research, including, an ISO standard for pointing device evaluation (ISO 9241-9). In general, pointing device evaluations consider 2 features of the device: throughput and accuracy. We will ignore the notion of accuracy (or errors) in this paper and solely focus on throughput.

To put it simply, Fitts's law states there is a linear relationship between the "Index of Diffculty" (ID) of a target, and its' movement time. The ID of a target is based on the target's width (W), and the distance or amplitude (A) between the target and the current position of the pointing device (or cursor). A basic method to evaluate pointing devices is by calculating it's average throughput, and comparing it against another device, interaction, or benchmark. This approach is generally referred to as the means of means approach and is fairly simple way of assessing the performance of a device. An improvement over ID is the Effective Index of Difficulty (IDe), based on the notion of *effective width* (W_e), instead of just width. The difference illustrated in figure 1 is based on the intuition that users who constantly hit a smaller area of the target are effectively performing a more difficult task. W_e can only be measured per user task, and is therefore calculated at the end of each task. Fitts regression could be done based on user means per condition based on IDe or means of means per ID.

Fitts regression could be used to check if a pointing device is compliant with Fitts's Law. It can also be used in designing interface and interaction styles to be used with the pointing device, for example, a larger slope means targets with higher ID will require higher movement time. A quick review of existing literature showed that researchers do often use means of means over ID for the regression. In this paper, we show that this method condenses data, which obfuscates potentially meaningful evidence. We show that regressing with user means over IDe, allows the ability to trend central tendencies with quantile regression, which we propose over boxplots. This also illustrates if the data points used to fit the Fitts regression line is (or is not) equal variance. The difference between these is illustrated in figure 2.

Related Work

Pointing device evaluations generally report 3 different metrics: throughput, accuracy [9], and regression. The throughput of a device is calculated generally by overall means-of-means per experiment [11]. Soukoreff & MacKenzie who advocated for the aforementioned method also showed that pointing device evaluation performed in accordance to ISO 9241-9 guidelines produces results that are consistent and comparable across studies. They suggested that a pointing device is considered Fitts compliant if the intercept is between 400 to -200 ms. They discussed the different types of regressions that can be performed, including all data points, means per condition, and means of means. It was indicated that a means per user approach could account for within-group variability, this was especially interesting to us as it considers human factors. These regression methods were performed and visualized in figure 3.

Goldberg et al [2] reported Root Mean Squared Error (RMSE) in their Fitts' models, and showed a worse fit as

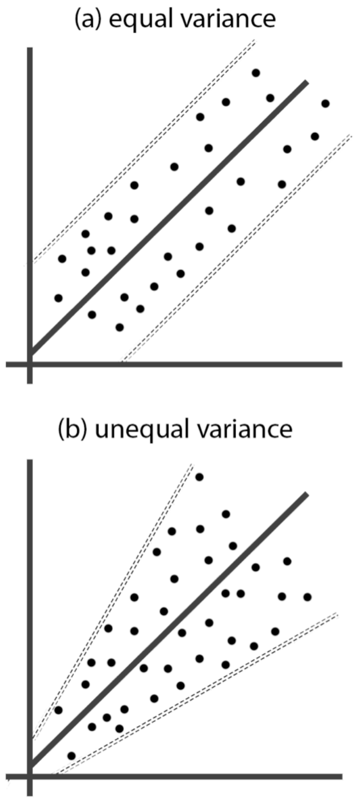

Figure 2 This cartoon example demonstrates the difference between plots with equal and unequal variance. Fitts regression plots that use MT per ID could imply to a reader that the graph is equal variance when it may not be.

the ID increases. While not explicitly stated in the paper, this shows that, pointing devices are not of equal variance, at least in their case. They used ID as a categorical variable in their paper whereas we use IDe, and as a continuous variable on the x-axis. We also do not calculate RMSE per ID, but the overall RMSE of the model. There is debate over the use of RMSE vs Mean Absolute Error (MAE) in statistics; in general RMSE is considered more suited when errors are in a Gaussian distribution, whereas MAE is more suited when errors are a uniform distribution [1]. Both RMSE and MAE are reported in this paper, as reporting a combination of both was found to be useful [1]. A cursory investigation of prior [4,2] and our own data showed that errors in pointing devices are closer to Gaussian than uniform.

Edward Tufte stated that “making an evidence presentation is a moral act as well as an intellectual activity.”[12] He has also argued against designs that obfuscate rather than reveal, for example over-simplifying graphics without considering multivateness [14]. We believe this principle is applicable to Fitts regression, as we could be eliminating necessary information such as heterogeneity of regression. It has been argued that HCI researchers should use better statistical methods, and to rethink tools and techniques, replacing them with those that would better serve the community [7]. One such method is identifying errors or residuals. Plotting of residuals vs the independent variables have been recommended to detect outliers, assess homogeneity of variance, to determine the power of prediction, and if a transformation is required (eg into a quadratic term)[8]. In our paper we show how plotting residuals can be useful to Fitts regression, we propose a method of plotting and reporting variance with quantile regression.

Methodology

This analysis uses a subset of the data from a previous experiment on pointing evaluation with gestural interaction. This subset does not significantly differ in the main effects, and therefore does not affect the results, lessons, or recommendations here. We direct readers to our previous work for further details about this experiment [3] and interaction [6].

In the aforementioned experiment, we compared 3 models of gestural interaction. The hyperplanar method was shown to be better in some ways and the resulting dataset was therefore used in this paper. The original experiment was done with the FittsStudy software [13] which employed the ISO 9241-9 ring-of-circles task. Table 1 shows the 4 target amplitudes and 3 widths used, resulting in 12 conditions but 10 distinct ID due to repeated IDs. In this paper, we only consider data from 2 widths (64 and 96) for simplicity and to avoid repetitions giving 8 unique IDs. Each task consisted of 3 test trials and 20 actual trials. With 8 conditions and 15 users, this gave us 20×15=300 trials per ID, and 300×8=2400 trials overall.

The original experiment was repeated over 3 days at 2 rounds per day. Therefore the dataset released contains 3×2×2400=14400 actual trials. However we only use data from the second round of the third day in this analysis for simplicity. There is expected to be a difference between rounds in gestural interaction [5], but this was this not found in our dataset on day 3.

Width	Amp.	ID
64	256	2.322
64	512	3.170
64	1024	4.087
64	1408	4.524
96	256	1.874
96	512	2.663
96	1024	3.544
96	1408	3.970
128	256	1.585
128	512	2.322
128	1024	3.170
128	1408	3.585

Table 1 Widths (W), Amplitudes (A) and Index of Difficulties (ID) used in the original experiment. Data from width=128 was dropped in this analysis to eliminate repeated IDs. IDs were calculated with the Shannon Formulation [10] as $\log_2(A/W+1)$.

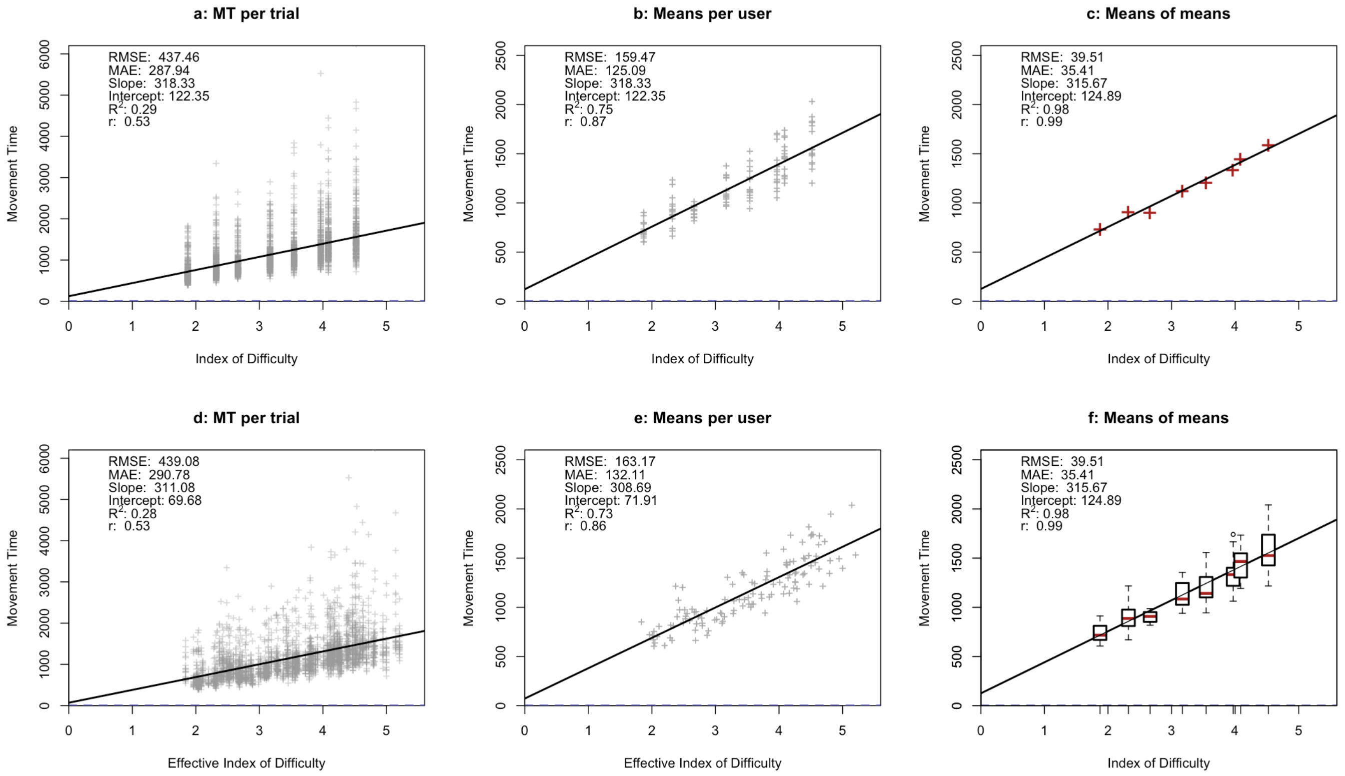

Figure 3 Plot (a) shows MT per trial for all trials by ID. Plot (d) shows the same MT but the x-axis uses IDe instead. Note that plots (a) and (d) are on a different scale on the y-axis from the rest. Plot (b) uses means per user by ID, plot (e) uses the same MT but over IDe. Plot (c) shows means of MT per ID, which is a common method used in research. Plot (f) augments plot (c) with boxplots per ID to illustrate central tendencies and to contextualize errors or residuals. Note how both axes always start at 0 which we believe better contextualizes the IDs used and reduces the possibility of misinterpretation, especially when comparing across plots.

Interpreting Plots

Figure 3a plots MT against ID for all 2400 trials, and is essentially the most basic form of visualization that can be formed from the data. Figure 3d on the other hand uses IDe, which smoothens the graph along the x-axes and visibly shows a better trend. A visual inspection of both 3a and 3d showed us that variance increased as the ID increased, implying unequal variance.

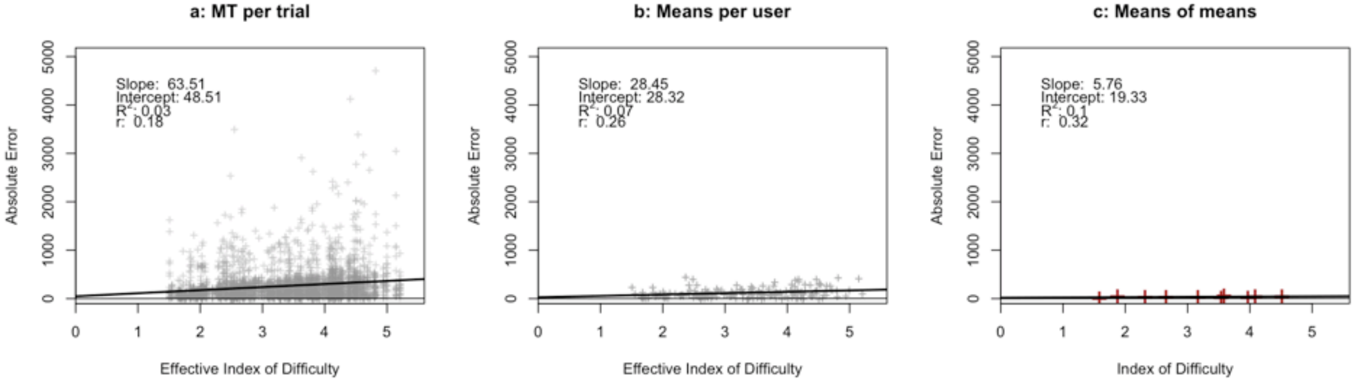

Figure 5 Absolute values of residuals for plots from figure 3. Figure (a) is based on 3d, (b) is based on 3e, and c is based on 3c. (a) and (b) shows that variance is dependent on ID, but this fact is not apparent when the means of means per ID method (c) is used.

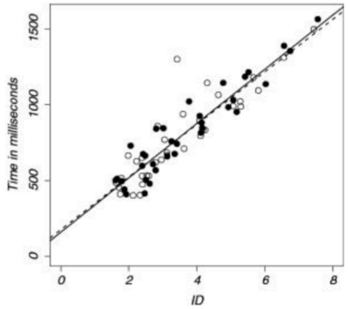

Figure 4 A comparison of 2 mice from [4], image used with author's permission. It was difficult for us to interpret if this graph is equal variance. We believe that an addition of quantile regression lines or independently plotting the absolute errors by ID could help make this apparent.

Figure 3b and 3e both use means per person per condition, but 3b uses ID, while 3e uses IDe. There is a subtle difference in stating that 3b uses means per person per condition and not IDe: a task is made up of one specific ring-of-circles or A×W condition, but two different tasks could have the same ID. We noticed lower variance in 3b and 3e than 3a and 3d, but still a visible – albeit less obvious – increase in variance as the ID increases. Plot 3e is the form we recommend for use in reporting and visualizing Fitts regression, as we believe this plot should be more indicative of variance than the means of means approach, and fewer outliers than the per-trial method. This approach effectively reduces the plot's y-limit from about 6000 in 3a and 3d to about 2500. There is also a better fit (*r* and R^2), which is an expected outcome due to the use of means.

Figures 3c shows the overall means per ID, a popular method used in Fitts studies. Figure 3f is a variation of 3c which augments the graph which boxplots. These boxplots could use means per user per condition and is therefore semantically similar to 3b. There is a near-perfect fit of the regression line, which is in part due to the fewer data points used in the regressions. Specifically there are as many points as there are tasks. This plot therefore does not factor in variance between users. Conceptually, we argue that the plot in 3c reports the performance of a device, but ignores the variance in the users that use them. We believe this approach ignores the users themselves, possibly making the results less generalizable to the population.

Errors, Residuals, and Quantile Regression

The term error is often used interchangeably with residuals. We define errors as the overall error of the model measured with RMSE or MAE, while residuals are the difference between one data point and its corresponding prediction. It is quite apparent from figure 3 that there is high variance in the per-trial plots (3a, 3d), while the user means plots in (3b, 3e) have visibly lower variance. Apart from the visual evidence, we can enumerate error based on the RMSE and MAE in each plots. We noted that the RMSE is 50% higher than MAE in the per trial plots, indicating very large but infrequent error. In comparison the user means plots' RMSE are about 30% higher than MAE.

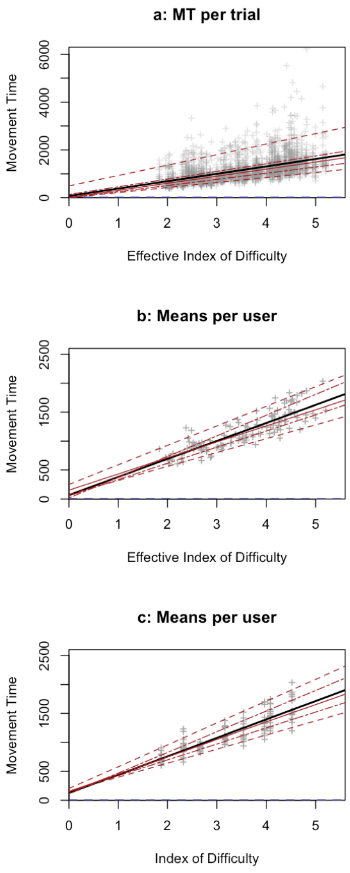

Figure 6 Images from figure 3 are augmented with quantile regression lines in red. Figure (a) is based on 3d, (b) on 3e and (c) on 3b. The quantiles used in each graph are 0.95, 0.75, 0.5, 0,25 and 0.05 respectively.

In the per-trial plots, the RMSE and MAE can be useful to intuit variance per trial. We can plug this value into figure 3d's intercept and slope to estimate the variance obtained in each trial. For example, suppose we want to know the range per trial when the IDe is 4, we can expect each trial to be between ((276.01×4 + 169.04) + 423.21)ms to ((276.01×4 + 169.04)-423.21)ms. We can similarly plug in the values to means per user plots to estimate the user's means by ID. But this cannot be used to infer meaningful information with the means of means per ID graphs.

The above method works well if the plots are equal variance; figure 5 however shows that our data is not. To address this, we propose the use of quantile regressions lines to augment the Ordinary-Least-Squares method generally used in Fitts regression. These graphs are shown in figure 6 and we believe exhibit the following benefits (1) it allows readers to visually infer that the data is of unequal variance, (2) it shows central tendencies as a trend rather than categorically, thus eliminating the need for boxplots, (3) it allows researchers to better estimate variance when the slope and intercept of each quantile is reported, (4) it can be used with ID or IDe, and (5) it is easier to define and identify outliers.

Conclusions and Future Works

In this paper, we showed that there are more benefits to reporting and visualizing Fitts regression based on means per user and IDe such as that in figure 3e. We believe this will provide better metrics that can be compared across experiments such as goodness of fit. Studies that use means of means per ID report goodness-of-fit (R^2) and Pearson's r very close to 1, implying that this interaction has a near perfect fit. This is true if we are solely considering the device performance in itself, but we argue that HCI research should be more interested in how users use the device. To that end, the movement times per trial (figures 3a and 3d) or means per user over IDe (figures 3b and 3e) should be a bigger concern.

We also demonstrated that quantile regressions as in figure 6b can be used to add more context to Fitts regression. Although not done here, we recommend reporting the slope and interception of each quantile regression line to allow for better understanding of the device's trends and errors as the ID increases.

The tools used for visualization in this paper can be used with the FittsStudy program to explore the results of a pointing device experiment in the future. We intend to first use these tools along with our dataset and other publically available Fitts datasets to check if our claims regarding quantile regression holds. It was said that IDs between 2-8 bits suffice for Fitts evaluations [11]. It would be interesting to visualize this, especially to see the variance at higher IDs. We noticed that the use of an alpha channel was useful in graphs with dense plots (e.g. figures 3d and 5a), as transparency allowed denser regions to be more apparent. We intend to demonstrate this with examples and possibly explore a more scientific method of defining alpha values. We believe more work needs to be done to identify the error distribution of our Fitts trials, which will in turn allow us to identify which error metric to use: MAE or RMSE. We believe the errors are in fact from a gamma distribution, but this requires statistical validation. We also intend to find out if the error, residuals, and unequal variance affects design considerations of pointing devices.

References

1. Tianfeng Chai, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." *Geoscientific Model Development* 7.3 (2014): 1247-1250.
2. Ken Goldberg, Siamak Faridani, and Ron Alterovitz. "Two Large Open-Access Datasets for Fitts' Law of Human Motion and a Succinct Derivation of the Square-Root Variant." Human-Machine Systems, IEEE Transactions on 45.1 (2015): 62-73. APA
3. Darren Guinness, Alvin Jude, G. Michael Poor, and Ashley Dover. 2015. Models for Rested Touchless Gestural Interaction. In *Proceedings of the 3rd ACM Symposium on Spatial User Interaction* (SUI '15). ACM, New York, NY, USA, 34-43. DOI=http://dx.doi.org/10.1145/2788940.2788948
4. Poika Isokoski, and Roope Raisamo. "Speed-accuracy measures in a population of six mice." *Proc. APCHI2002: 5th Asia Pacific Conference on Computer Human Interaction*. 2002.
5. Alvin Jude, G. Michael Poor, and Darren Guinness. 2014. An evaluation of touchless hand gestural interaction for pointing tasks with preferred and non-preferred hands. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (NordiCHI '14). ACM, New York, NY, USA, 668-676. DOI=http://dx.doi.org/10.1145/2639189.2641207
6. Alvin Jude, G. Michael Poor, and Darren Guinness. 2014. Personal space: user defined gesture space for GUI interaction. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*(CHI EA '14). ACM, New York, NY, USA, 1615-1620. DOI=http://dx.doi.org/10.1145/2559206.2581242
7. Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 1105-1114. DOI=http://dx.doi.org/10.1145/2207676.2208557
8. Wayne A. Larsen, and Susan J. McCleary. "The Use of Partial Residual Plots in Regression Analysis". *Technometrics* 14.3 (1972): 781–790.
9. I. Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. 2001. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '01). ACM, New York, NY, USA, 9-16. DOI=http://dx.doi.org/10.1145/365024.365028
10. I. Scott MacKenzie, "A Note on the Validity of the Shannon Formulation for Fitts' Index of Difficulty." *Open Journal of Applied Sciences* 3.06 (2013): 360. DOI=http://dx.doi.org/10.4236/ojapps.2013.36046
11. R. William Soukoreff and I. Scott MacKenzie. 2004. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *Int. J. Hum.-Comput. Stud.* 61, 6 (December 2004), 751-789. DOI=http://dx.doi.org/10.1016/j.ijhcs.2004.09.001
12. Edward R. Tufte, "Beautiful evidence." *New York* (2006).
13. Jacob O. Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 1639-1648. DOI=http://dx.doi.org/10.1145/1978942.1979181
14. Mark Zachry, and Charlotte Thralls. "An Interview with Edward R. Tufte."*Technical Communication Quarterly* 13.4 (2004).