The Cost of Pie Charts

Harri Siirtola
TAUCHI Research Center
Faculty of Information Technology and Communication Sciences (ITC)
Tampere University, Finland
harri.siirtola@tuni.fi

Abstract

Visualization of proportions is a very common need and there are many techniques for it. Pie chart is popular among practitioners and general audience, but many prominent experts advice against using it.

This paper reports an experiment where the pie chart is compared to stacked bar charts with a baseline condition of table. The aim is study if the performance differences justify the bad reputation of pie charts.

In the experiment the participants were requested to list out the elements of a visualization in the decreasing order of their size. The task time and the answer were recorded.

The results show that the pie chart is slower and less accurate than the stacked bar chart, especially when the difference between the elements is small, but the participant find it slightly more pleasant to use. The participants also perceive the stacked bar chart as the most effective visualization.

Keywords— **visualization of proportions; pie charts; stacked bar charts**

1 Introduction

Pie chart is probably the most common method to visualize proportions, or to how the size of the some part relates to size of other parts and the whole. They can be found everywhere – for example, in newspapers, scientific articles, and television newscasts. Besides being popular, the pie chart is also controversial. Practitioners and the general public favour them, and the experts advice not to use them, mainly because there are more efficient representations available.

The pie chart is criticized to be inefficient since the data encoded can be read from multiple sources (angle, length of arc, area of segment). It is known that (linear) length is easier to estimate than angle, and the estimation of area is even worse [2]. This is probably the fundamental reason why so many prominent experts advice against using them ('The only thing worse than a pie chart is several of them' [13, p. 178], 'Save pies for the dessert' [3], 'Death

to pie charts' [9]). On the other hand, there are also prominent defenders of pie chart ('Why Tufte is Flat-Out Wrong about Pie Charts' [4], In Defense of Pie Charts' [5]).

This paper compares three common methods to visualize parts of some whole, namely, stacked bar chart, pie chart, and a table of numbers (as a baseline). There are other methods for the same task, such as Treemap [11], Voronoi diagram [1], rose diagram [8], and methods based on the small multiples [14], but they are not so common.

One typical misunderstanding in the pie chart vs. bar chart controversy is the comparison of pie charts to regular bar charts. The point is that a regular, side-by-side bar chart is missing the 'whole', the visual sum of all values – we need to use a *stacked bar chart* to make a fair comparison.

In the earlier studies, the choice of task has proven to be a problem. It is not easy to choose a task that does not favor one of the techniques. In this experiment, we concentrate on the magnitude comparisons by graphics. We ask the participants to simply name the parts of the visualization in descending order of their proportion. This is an elementary task when the proportions of a whole are visualized.

One might question what is the ecological validity of simply listing the elements in descending order. After all, we could present the parts in sorted order, or include the exact proportion in the label. However, often the parts of some whole have a natural order which can't be violated (e.g. political stance, or anything expressed in ordinal scale).

Another important aspect is how to define 'performance' when the proportions are perceived from a visualization. In this study, the performance is the combination of task time, error rate, and subjective satisfaction. The research question in this study is to find out what is the performance difference between pie chart, stacked bar chart, and a table of numbers.

2 Method
2.1 Participants

Twenty-eight students volunteered to participate the experiment. Twenty-three participants were recruited in May

2014 from the introductory course in *Interactive Technology* who had to write a short report about their participation as a partial fulfillment of the course (the author of this paper has no other connection to this course). After the summer break in August, five additional participants were recruited among the students available in our campus to complete the counter-balancing. Nationalities of the participants were Finn (21), Estonian (3), Japanese (2), and Turkish (2). All participants received a free-pass movie ticket for their time (a standard practise in our laboratories).

The participants were screened to have either normal or corrected-to-normal vision. Genders were balanced (14/14), and the average age was 27.5 (median 25, min: 20, max: 52, sd: 7.7). Exposure to pie charts was investigated by asking how often they read newspapers (sometimes: 18, every day: 10) or watch the evening news on TV (never: 7, sometimes: 15, every day: 6). Familiarity to data graphics in general was queried by asking the number of statistical courses taken (none: 14, one: 5, two: 4, more: 3). The fluency in touch-typing was also queried (none: 7, adequate: 9, ok: 8, good: 4).

2.2 Apparatus

The experimental setup was a Java program run on an iMac computer. The program displayed a stimulus, recorded the participant's input, and the time spent on a task. Figure 1 shows the user interface. The input is a string of letters from the beginning of the alphabet, and the program accepts only those letters that appear in the stimulus image. The answer button (on the bottom right, Figure 1) activates when all the letters appearing in the stimulus have been inserted into the answer field. The program requires another button press before proceeding to the next task, thus allowing the participant to rest between tasks if so desired.
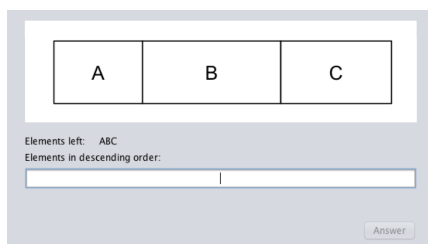

Figure 1: User interface of the *GraphTest* application.

The stimuli was created programmatically to have the desired difference between the part sizes. The visualizations were as simple as possible – no colors, no scales, plain data labels, and all pixels are data-bearing. The numbers in table condition were not percentages, but the same

random data that was used to generate the graphic stimuli, shown with two decimal places.
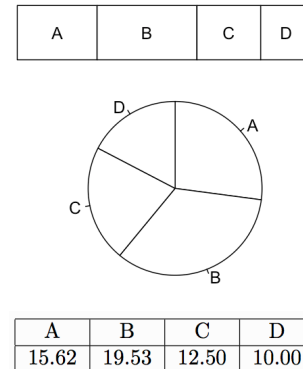

A	B	C	D
15.62	19.53	12.50	10.00

Figure 2: Three variants of stimuli (from top to bottom): stacked bar chart, pie chart, and table.

2.3 Procedure

Participants first completed a questionnaire for demographics and background information, filled in consent form, and were informed that they may withdraw from the experiment at any time, or even decline afterwards the use of their data.

Next, each participant was given the same instructions (read aloud from a paper), and familiarized with the four visualization types. The functionality of the *GraphTest* application was explained, and the participants did eight training tasks with it. The participants were encouraged to ask if there was anything unclear in the *GraphTest* application, the visualizations, or the test itself.

The test task was to list the elements in the visualization in a descending order of their size. E.g. in Figure 2 the correct answer is *BACD*. Participants were instructed to perform the tasks as quickly as possible, but not at the expense of the correct answer. The participants did not compete against each other, but they could ask about their overall placement afterwards if desired, with respect to time and accuracy (many of them were interested about this).

Each participant did the same set of tasks with all four visualizations. Both the task order and the order of visualization types were counterbalanced with Latin square designs. The tasks were created by varying the number of elements in the visualization (between 4 and 7) and the minimum difference between elements (from 6% to 18%, in 4% increments). There were 4 (percentage difference) $\times$ 4 (element counts) $\times$ 3 (visualization types) = 48 tasks altogether.

After the test tasks the participants filled in a questionnaire about their subjective opinions on the four visualizations and their performance, and were interviewed for ad-

ditional comments.

The *GraphTest* application recorded the participants' answers and computed the string distance to the correct answer. The distance metric is the number of adjacent letter swaps required to transform the answer to the correct one. Figure 3 shows all the possible adjacent letter swaps for the string *ABCD*. If the participant responds *BCAD* and the correct answer is *ABCD*, then the distance is 2.
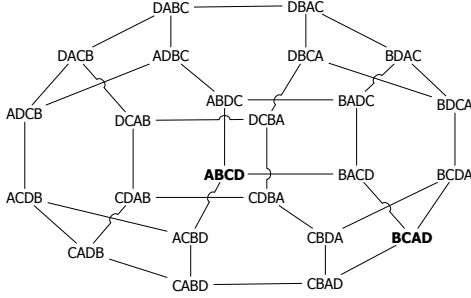

Figure 3: Distance metric: the number of adjacent letter swaps to create the correct answer.

3   Results

Type	Time	Distance	Errors
Bar	9542	0.16	58
Pie	11420	0.40	130
Table	12471	0.04	18
Mean	11145	0.23	206

Table 1: Grand mean of task time (ms), mean distance from the correct answer as adjacent item swaps , and the number of errors.

Linear mixed effects modeling was used to describe the effects of visualization type and data properties on task time and correctness. The visualization methods (*Type*) were stacked bar chart, pie chart, and a numerical table. The data was characterized by the number of data items (*Count*) and the minimum difference between them (*Percentage*). *Type*, *Count*, and *Percentage* were fixed effects, and participant was a random effect. There was a separate slope and intercept for each visualization type.

The task time was modeled as *TimePerElement*, i.e. the task time was divided by the number of elements. Values higher than $3 \times SD$ (4508ms) were rejected as outliers (25 observations, 1.86%, these were mainly cases where the participant had problems typing the answer). In addition, the task time was log-transformed to meet the normality assumption. The detailed results follow.

3.1   Time

The grand mean for task completion time was 11.1 seconds (Table 1). The main effects of *Type* ($F_{2,26.01} = 36.35, p < .001$), *Count* ($F_{3,1190.56} = 148.11, p < .001$) and *Percentage* ($F_{3,1190.56} = 76.49, p < .001$) were statistically significant. There were also significant interactions: Type:Count ($F_{6,3,1190.56} = 5.71, p < .001$), Type:Percentage ($F_{6,3,1190.54} = 16.93, p < .001$), and Count:Percentage ($F_{9,3,1190.55} = 4.55, p < .001$). Figure 4 shows the interaction plots. All interactions between stacked bar chart and pie chart were ordinal.

We can use the estimated marginal means to determine the source of these differences. Table 3 shows the details of pairwise comparisons, and Table 2 is a summary.

	4	5	6	7
6%	Bar < Pie	Bar < Pie	Bar < Pie Bar < Table Pie > Table	–
10%	Bar < Pie Bar < Table	Bar < Pie Pie > Table	Bar < Pie Bar < Table	Bar < Table
14%	Bar < Table	Bar < Table Pie < Table	Bar < Table Pie < Table	Bar < Table Pie < Table
18%	Bar < Pie	Bar < Table Pie < Table	Bar < Table	Bar < Table Pie < Table

Table 2: Summary of the differences in task time between Bar Chart, Pie Chart, and Table per the number of elements and difference in percentage. 'Bar < Pie' denotes that Bar Chart was faster with statistically significant difference.

3.2   Errors

In the following, the answers are first considered as number of errors (answer is either right or wrong), and then as a distance from the correct solution.

We can define *index of difficulty (IOD)* for a task as function of the number of elements in visualization and the percentual difference between those elements as $IOD = log(5 \times Count/Percentage)$ (the constand 5 is just a scaling factor to return positive values). This is akin to the predictive model of human movement often applied in human-computer science [7]. Figure 5 shows the relationship between IOD and the number of errors. Unsurprisingly, the task with highest number of elements and smallest difference between them is the most difficult one, and vice versa. Figure 6 shows the same dependence per visualization type. With easy tasks (low IOD) there is no difference, i.e. it doesn't matter which visualization to use, but with hard tasks there is a clear difference.

The grand mean for distance was 0.203 swaps (206 answers out of 1344 were incorrect). A generalized linear
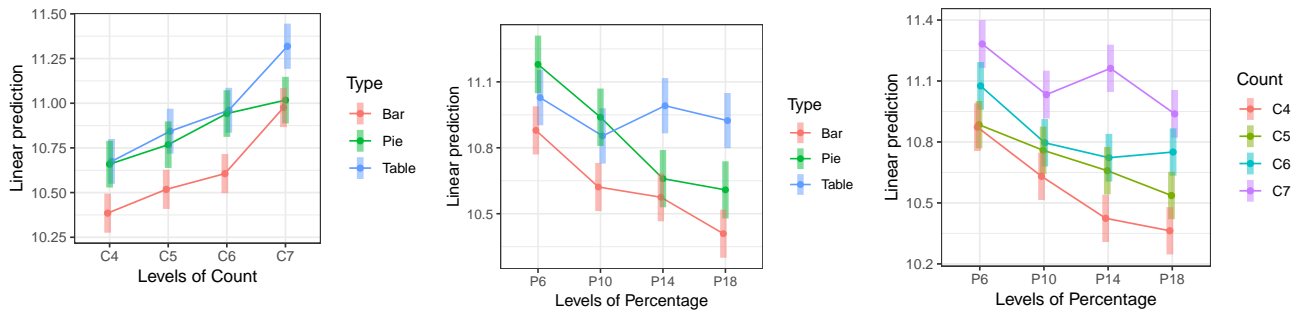
Figure 4: Interaction plots for *Type × Count*, *Type × Percentage*, and *Count × Percentage*.

%	Count	contrast	estimate	SE	df	t.ratio	p.value	
6	4	Bar - Pie	-0.2881	0.1024	402	-2.815	0.0142	*
		Bar - Table	-0.0728	0.0952	611	-0.764	0.7246	
		Pie - Table	0.2153	0.1044	333	2.063	0.0991	
	5	Bar - Pie	-0.3245	0.0998	374	-3.251	0.0036	*
		Bar - Table	-0.1694	0.0961	623	-1.764	0.1826	
		Pie - Table	0.1551	0.1027	316	1.510	0.2873	
	6	Bar - Pie	-0.5222	0.1014	392	-5.151	<.0001	*
		Bar - Table	-0.2307	0.0969	634	-2.381	0.0460	*
		Pie - Table	0.2915	0.1035	324	2.818	0.0141	*
	7	Bar - Pie	-0.0664	0.1022	401	-0.649	0.7928	
		Bar - Table	-0.1232	0.0969	634	-1.272	0.4115	
		Pie - Table	-0.0568	0.1043	332	-0.545	0.8490	
10	4	Bar - Pie	-0.4124	0.1006	383	-4.099	0.0002	*
		Bar - Table	-0.2462	0.0952	611	-2.585	0.0268	*
		Pie - Table	0.1662	0.1027	316	1.618	0.2393	
	5	Bar - Pie	-0.3978	0.1006	383	-3.954	0.0003	*
		Bar - Table	-0.1105	0.0952	611	-1.161	0.4770	
		Pie - Table	0.2873	0.1027	316	2.798	0.0151	*
	6	Bar - Pie	-0.3809	0.1006	383	-3.785	0.0005	*
		Bar - Table	-0.2797	0.0952	611	-2.936	0.0088	*
		Pie - Table	0.1012	0.1027	316	0.986	0.5862	
	7	Bar - Pie	-0.0789	0.0998	374	-0.790	0.7090	
		Bar - Table	-0.2934	0.0989	660	-2.967	0.0093	*
		Pie - Table	-0.2145	0.1053	342	-2.037	0.1049	
14	4	Bar - Pie	-0.1780	0.0998	374	-1.783	0.1764	
		Bar - Table	-0.3942	0.0952	611	-4.139	0.0001	*
		Pie - Table	-0.2162	0.1019	309	-2.121	0.0872	
	5	Bar - Pie	-0.1218	0.0998	374	-1.221	0.4416	
		Bar - Table	-0.4923	0.0961	623	-5.125	<.0001	*
		Pie - Table	-0.3705	0.1027	316	-3.608	0.0010	*
	6	Bar - Pie	-0.2223	0.1006	383	-2.210	0.0707	
		Bar - Table	-0.5215	0.0969	634	-5.380	<.0001	*
		Pie - Table	-0.2992	0.1043	331	-2.870	0.0121	*
	7	Bar - Pie	0.1839	0.0998	374	1.842	0.1574	
		Bar - Table	-0.2580	0.0952	611	-2.709	0.0190	*
		Pie - Table	-0.4418	0.1019	309	-4.335	0.0001	*
18	4	Bar - Pie	-0.2171	0.0998	374	-2.175	0.0767	
		Bar - Table	-0.4446	0.0952	611	-4.668	<.0001	*
		Pie - Table	-0.2275	0.1019	309	-2.232	0.0674	
	5	Bar - Pie	-0.1564	0.0998	374	-1.567	0.2611	
		Bar - Table	-0.5299	0.0952	611	-5.564	<.0001	*
		Pie - Table	-0.3735	0.1019	309	-3.665	0.0009	*
	6	Bar - Pie	-0.2197	0.0998	374	-2.201	0.0723	
		Bar - Table	-0.3863	0.0952	611	-4.056	0.0002	*
		Pie - Table	-0.1666	0.1019	309	-1.634	0.2325	
	7	Bar - Pie	-0.2047	0.0998	374	-2.050	0.1017	
		Bar - Table	-0.6975	0.0979	647	-7.125	<.0001	*
		Pie - Table	-0.4928	0.1044	333	-4.720	<.0001	*

Table 3: Pairwise comparisons with mvt correction, by using multivariate t distribution with the same covariance structure as the estimates to determine the adjustment. The estimate is in logarithmic scale.
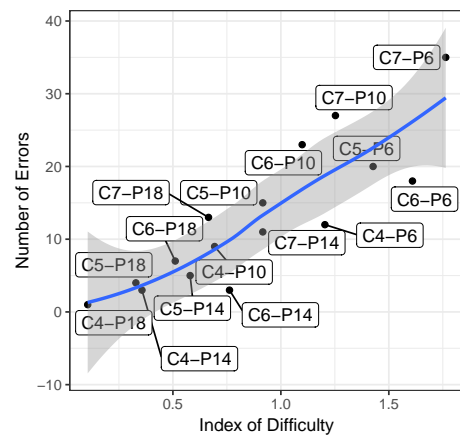

Figure 5: Index of difficulty vs. number of errors made. The blue line is a loess regression model, and the gray area is the 95% confidence interval for the model. Label *C7-P6* denotes a task with element count of seven and percentage differenc of six.
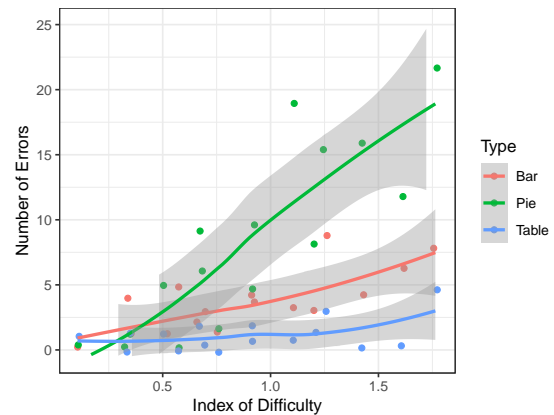

Figure 6: Index of difficulty vs. number of errors per graph type, similarly as in Figure 5.

Type	Percentage	C4	C5	C6	C7
Bar	P6	3	4	6	8
Bar	P10	3	4	3	9
Bar	P14	1	5	1	4
Bar	P18	–	4	1	2
Pie	P6	8	16	12	22
Pie	P10	6	10	19	15
Pie	P14	1	–	2	5
Pie	P18	–	–	5	9
Table	P6	1	–	–	5
Table	P10	–	1	1	3
Table	P14	1	–	–	2
Table	P18	1	–	1	2

Table 4: The number of errors in each condition.
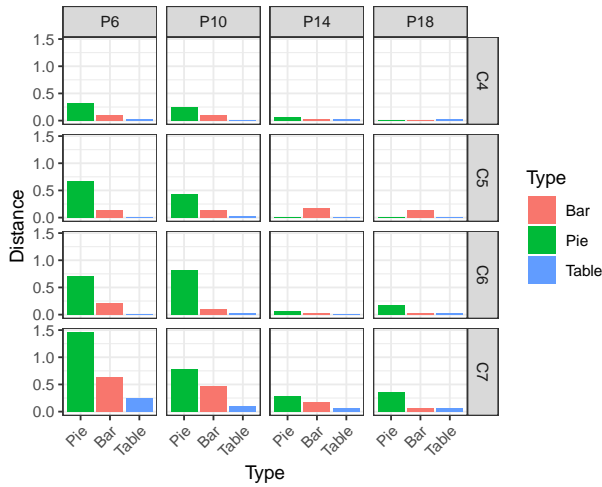
Figure 7: The mean distance from correct solution as swaps of adjacent items.

model with Poisson distribution was fitted. The main effects of *Type* ($F_2 = 83.66, p < .001$), *Count* ($F_3 = 42.23, p < .001$) and *Percentage* ($F_3 = 31.04, p < .001$) were statistically significant (error estimate based on Pearson residuals). The pairwise interactions were not significant, but the three-way interaction was ($F_{18} = 2.00, p < .01$). This was caused by the very small and noisy distances at higher percentage levels (P14 and P18), i.e. when the difference was clear (see Table 4).

3.3 Opinions

There was a short post-test questionnaire for participants to rank the visualizations according to pleasantness of use (1st, 2nd, 3rd), and to estimate which one was the fastest to grasp, the most effective. For the pleasantness of use, there was a small preference of pie chart over bar chart, and the rating for the baseline was half of that for
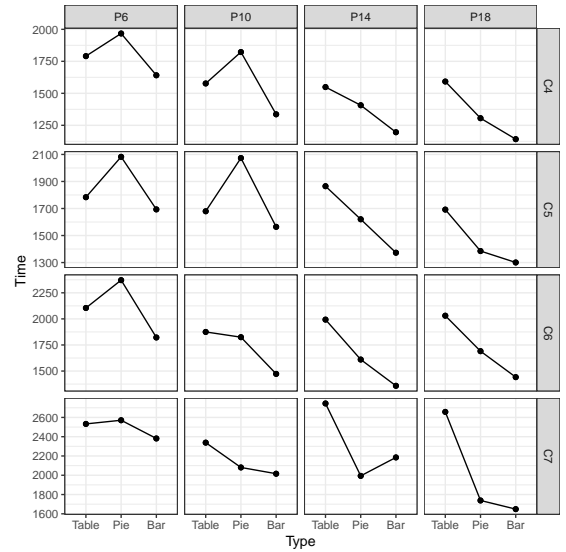

Figure 8: The median task time per element in each condition.

graphical presentations. For effectiveness, the bar chart got twice as many votes as pie chart – even the baseline got a slightly higher ranking.

4 Discussion

The results from this experiment are not straightforward. In task time, there are complex pairwise interactions, but these are caused by the baseline condition – the interactions between Bar and Pie are ordinal (Figure 4).

Table 2 is a summary of the Table 3, showing only the statistically significant differences. Out of the sixteen conditions, there is not a single case where the pie chart would be faster than the bar chart. The pie chart beats the baseline 5 out of 16 cases, but is slower in two cases. On the other hand, the bar chart is significantly faster than the pie chart in 7 out of 16 cases, and beats the baseline in 11 out of 16 cases.

Perhaps the most revealing view to the task times is to look at the median task times in each condition (Figure 8). When the difference between elements is clear (14% or more), the bar and pie chart clearly outperform the textual representation. The bar chart outperforms the pie chart, except that there is no difference with 7 elements. With a small difference (10% or less), the pie chart is slower than the bar chart or the baseline, again excluding the 7 element case.

Table 4 summarises the number of errors, i.e., when the distance from the correct solutions was not zero. As we can see from Figure 6, the number of errors climbs more steeply with the pie and bar charts, the pie chart being steepest.

Figure 7 shows the mean distance from the correct solution. The bar chart is more accurate than the pie chart in all but two cases, and both graphical methods are less accurate than the table.

User opinions from the questionnaire data reveal quite compellingly the root of the pie chart controversy: users like the pie chart although they understand that the bar chart is more effective.

The pie charts in this study were used according to the prevailing 'gold standard': no more than seven segments, no cut-outs, and no 3D. Findings of this study suggest one addition: the pie chart should not be used if the data items have small differences and the magnitude comparisons are important – as they often are.

5 Conclusion

The results support the claim that the pie chart is slower and less accurate than the stacked bar chart. Overall, looking at the data in Table 1, the pie chart was almost 20% slower in this experiment, and over twice as inaccurate as the bar chart. That is not a trivial cost. On the other hand, the stacked bar chart was about 13% faster than the baseline, a table of numbers, but there was a ten-fold difference in accuracy. Clearly, the table should be seriously considered over graphical presentations if the data items have small differences.

Acknowledgments

References

[1] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.

[2] W. S. Cleveland, "Research in statistical graphics," English, *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 419–423, 1987, ISSN: 01621459. [Online]. Available: `http://www.jstor.org/stable/2289443`.

[3] S. Few, *Save the pies for dessert*, Visual Business Intelligence Newsletter, 2007. [Online]. Available: `http://www.perceptualedge.com/articles/08-21-07.pdf` (visited on 03/22/2019).

[4] B. Gabrielle. (2013). Why Tufte is flat-out wrong about pie charts, [Online]. Available: `http://speakingppt.com/2013/03/18/why-tufte-is-flat-out-wrong-about-pie-charts/` (visited on 03/22/2019).

[5] R. Kosara. (2011). In defense of pie charts, [Online]. Available: `http://eagereyes.org/criticism/in-defense-of-pie-charts` (visited on 03/22/2019).

[6] R. Lenth, *Emmeans: Estimated marginal means, aka least-squares means*, R package version 1.3.3, 2019. [Online]. Available: `https://CRAN.R-project.org/package=emmeans`.

[7] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *HUMAN-COMPUTER INTERACTION*, vol. 7, pp. 91–139, 1992.

[8] W. Nemec, "The shape of the rose," *Sedimentary Geology*, vol. 59, no. 1-2, pp. 149–152, 1988.

[9] C. Nussbaumer. (2011). Death to pie charts, [Online]. Available: `http://www.storytellingwithdata.com/2011/07/death-to-pie-charts.html` (visited on 03/21/2019).

[10] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: `http://www.R-project.org/`.

[11] B. Shneiderman and M. Wattenberg, "Ordered treemap layouts," in *infovis*, IEEE, 2001, p. 73.

[12] H. Singmann, B. Bolker, J. Westfall, and F. Aust, *Afex: Analysis of factorial experiments*, R package version 0.23-0, 2019. [Online]. Available: `https://CRAN.R-project.org/package=afex`.

[13] E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1983, 197 p.

[14] ——, *Visual Explanations : Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press, 1997, 156 p.

[15] H. Wickham, *Tidyverse: Easily install and load the 'tidyverse'*, R package version 1.2.1, 2017. [Online]. Available: `https://CRAN.R-project.org/package=tidyverse`.