The Impact of Distribution and Chart Type on Part-to-Whole Comparisons

Robert Kosara
Tableau Research

Abstract

Pie charts and treemaps are commonly used in business settings to show part-to-whole relationships. In a study, we compare pie charts, treemaps, stacked bars, and two circular charts when answering part-to-whole questions with multiple slices and different distributions of values. We find that the circular charts, including the unusual variations, perform better than the treemap, and that their performance depends on whether participants are asked to judge the largest slice or a smaller one.

1. Introduction

Many real-world charts show part-to-whole relationships: government budgets on different levels, business graphics breaking down sales or profits by region or time, etc. The infamous pie chart is used for this purpose, but there are also others. In particular, treemaps are increasingly used not to show complex or deep hierarchies, but as pie chart alternatives.

A little-studied question is how the data shown impacts how readable a chart is. Might the distribution of values make a difference? How would that manifest in part-to-whole scenarios? Given a part-to-whole question, does it make a difference which slice the question is about, the largest, the smallest, or one in-between?

In addition to the established pie chart, treemap, and bar chart, we decided to include two novel designs in our study that are based on an experimental stimulus found to perform no worse than the pie chart in a recent study [SK16]. This was done to expand the design space and to test if they might be workable alternatives to the pie chart. We describe our study below and present our results.

2. Related Work

The classic study of visual perception in visualization by Cleveland and McGill study [CM84] assessed a number of commonly-used visualization techniques such as bar charts, pie charts, etc. They focused on a small number of encodings, in particular length, position, and angle.

Even charts as simple as the bar chart show complicated behavior depending on the way the different elements are drawn (which may have caused some of the effects Cleveland and McGill observed [TSA14]), can be negatively influenced by the noise in the data being shown itself [ZTLS98].

Task also matters. Simkin and Hastie [SH87] found that pie charts were as accurate as bar charts, and better than stacked bars, in part-to-whole comparisons. Another study came to a similar conclusion when comparing pie, donut, and bar charts to a waffle chart [KZ10]. Hollands and Spence [HS92] also found that pie charts performed better than bar and line charts in their proportion task, which they believed was because of the more direct way of seeing proportion in those charts. Recent work [SK16] also found that error on the unusual charts used to separate visual cues, such as their *area-only chart*, was no different than the basic pie chart.

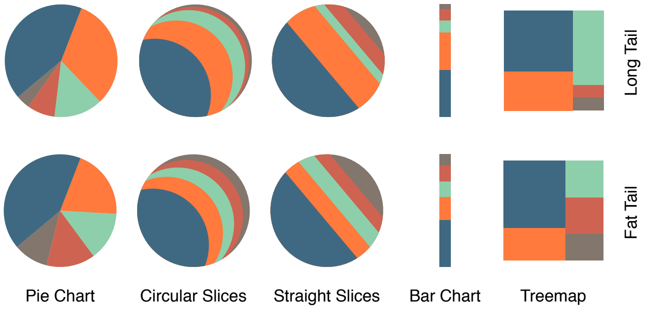

Figure 1: *The charts used in the study, shown for a largest slice percentage of 42% and the two different tail types.*

Studies examining pie charts come to different conclusions depending on the questions asked and the configurations used. Spence [Spe90], for example, found pie charts to perform as well as bar charts in a study comparing only two values to each other. Pie charts also performed worse than some bar chart configurations in Cleveland and McGill's study, but better than some others (in particular stacked bars).

Many papers have compared precision between parts, often including stacked bars and pie charts, but also the more recent treemap [Shn92]. In their comprehensive crowd-sourced study of perceptual tasks, Heer and Bostock compared additional charts like bubble charts and treemaps [HB10]. Kong et al. [KHA10] investigated ways to improve precision when reading treemaps. They found that the aspect ratio has an influence on error, and that squares (like they are common in squarified treemaps [BHvW00]) do worse than some other rectangles.

3. Study Overview

To investigate part-whole charts, we selected the following set of charts for this study (Figure 1), which includes three established chart types and two novel designs:

Baseline Pie Chart. A regular pie chart with five slices.

Circular Slice. This chart represents the value by sliding circles of the same size over a base circle (from a related study [Kos19]).

Straight-Line Circular. A round chart with slices delineated by straight lines. This is based on the *area-only* condition in an earlier study [SK16]. The difference here is that we show more than two values by dividing the circle into several bands.

Stacked Bar Chart. A stacked bar chart.

Treemap. A single-level treemap laid out as a squarified treemap.

All stimuli encode the value as area. In addition, the pie chart adds central angle and arc length, and the bar chart adds length.

In the interest of keeping study complexity low, we chose to show five slices in all stimuli. Instead of varying the number of slices, we asked participants about either the largest slice or the middle slice. In addition, we also decided to test different distributions of those values, which we expected to have an effect on the middle-slice case but not the largest slice. In order to reduce the number of cases to test, we only varied the distribution for the middle-slice question. This leads to the following structure:

Largest Slice. Ask about the largest slice in the chart. The remaining values are distributed in a long tail.

Middle Slice. Ask about the third slice out of five. This case was subdivided into two sub-cases:

Long Tail. After picking the value for the largest slice, this distribution divides the remaining values such that they drop off by a factor of two between consecutive slices. The result is a power-law distribution.

Fat Tail. This distribution divides the remaining values into equal parts and subtracts small numbers in the range $[-3;3]$ from each. The resulting distribution is flat.

In total, this leads to 5 variations × 4 values for the largest slice case, and 5 variations × 4 values × 2 tail types for the middle slice, for a total of $20+40=60$ charts per participant.

4. Hypotheses

This study is based on the common metrics of error and response time. Specifically, we had the following expectations:

- More familiar charts should do better. We expect pie charts, stacked bars, and possibly treemaps to show smaller absolute error and be read faster than the two more unusual charts.
- Choice of slice to read matters. The largest slice should be easier to read than the middle one, leading to less error and faster responses.
- Distribution affects error and reading time. A long-tail distribution should be easier to read and take less time, because the slices are more different in size and the values thus easier to tell apart.

5. Materials

To prepare the study stimuli, we picked the value of the largest slice first, then distributed the rest of the slices according to the tail type. Since the remaining values get too small when the large slice is larger than 80%, and the largest slice cannot be smaller than about 39% in the long-tail case, we restricted the range of values to choose from to $[39;79]$. This range was divided into four bins again and we removed all multiples of 5 and the common value 66. This left us with eight values in each bin.

For each participant and chart variation, we drew numbers randomly from each of the four bins. Values for all charts were sorted, following the common recommendation to draw pie charts with sorted slices, and a similar common practice in squarified treemaps. As an additional step, we perturbed the middle value in each distribution by adding a small amount of randomness (varying within 80% of the difference between the two neighboring slices).

All charts were rotated by a random amount. While this is meaningful for the round charts, it did not appear useful for the bar chart and treemap. These two charts were therefore only rotated in 90° increments.

We chose the colors in the chart from Tableau's *Miller-Stone* palette because we found them easy to differentiate both visually and because their color names clearly differ (which aids distinguishability [BK91]).

All round charts were presented at a diameter of 600 pixels. The treemap was square and sized to 532 pixels on each side so that it would take up the same area as the circles. The bar chart was presented at 600 pixels in length and 60 pixels wide.

The order in which stimuli were presented was randomized by shuffling the parameter sets after they were constructed.

6. Procedure

The study interface was presented in a web browser window. We showed the chart on the left with an input field in the lower right for the estimated value. The question was, *What part of the whole (in percent) does the* 🟦 / 🟩 *part represent?* (with only the respective color shown). When users entered values over 100, they were asked to try again. A progress indicator was shown along the bottom of the screen. After 20 and 40 questions, participants were encouraged to take a short break.

We recruited 81 participants on Amazon's Mechanical Turk. They were given a short introduction that instructed them to read the percentages from the charts shown. 41 participants identified as female, 47 male (42% and 58%, respectively); the majority (41) were in the 30–39 age group, with ages from 18 to 60+ represented. Education was split between high school diplomas and bachelor's degrees, only three each reported a master's degree or other.

Participants completed the study in an average of 9 minutes and 10 seconds. They were paid $2.50, which made for an average hourly rate of $16.39.

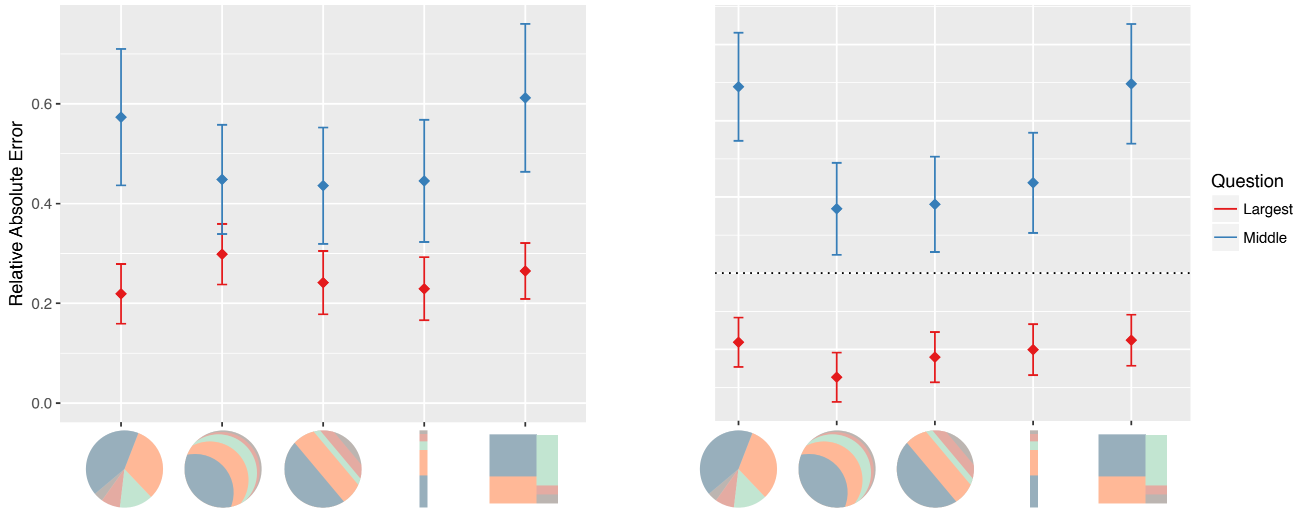

Figure 2: *Absolute and signed error relative to the true percentage by slice asked about and chart type (means and 95% confidence intervals). Relative error is higher for the middle slice than the largest slice question, with the exception of the circular slice chart.*

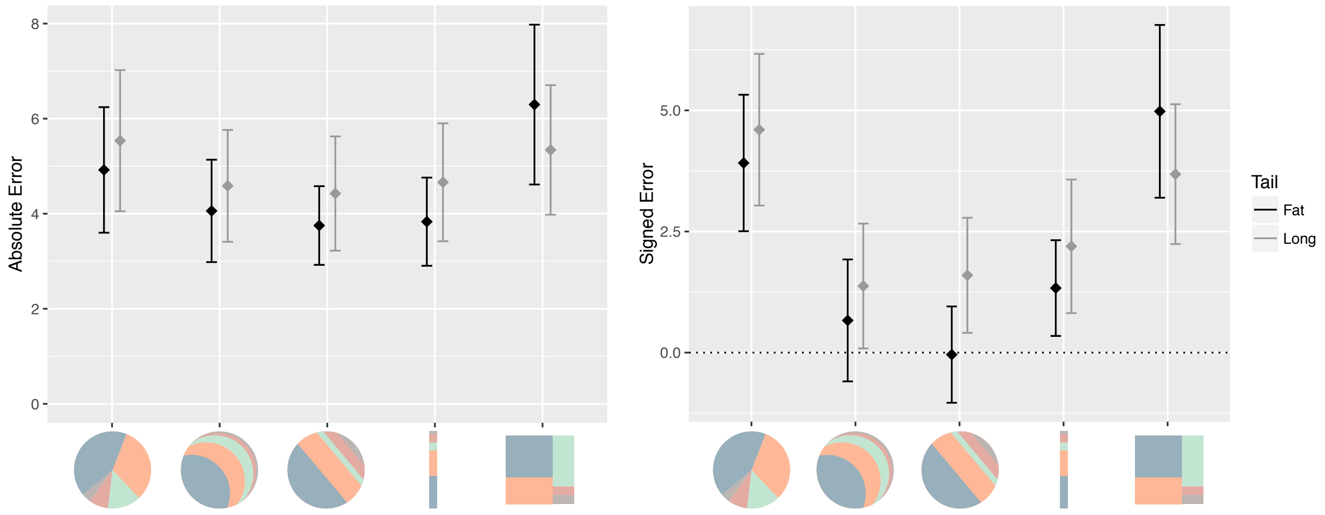

Figure 3: *Both absolute and signed error differ significantly within all variations (means and 95% confidence intervals). The long tail leads to higher absolute error (left) and negative signed error (systematic underestimation, right). The smaller absolute error of the fat tail is less biased.*

7. Results

Despite our efforts to clearly indicate the slice being asked about, we found that roughly 13% of questions were apparently answered for the wrong slice. This was determined by comparing whether the absolute error was smaller for the slice we asked about or the possible alternative (largest or middle, depending on the question). Wrong answers were very evenly distributed across chart types, but they were much more common in a handful of participants who presumably did not pay attention to the question. We therefore decided to remove the data of five participants who appeared to answer more than 20 questions (one third) for the wrong slice. We also found a logging problem in one participant and decided to remove their data as well. We thus ended up with data from 75 participants.

In order to compare between question types, we calculated relative absolute error (absolute error divided by the correct value) and relative signed error (signed error divided by the correct values). This measure directly relates the estimate to the exact true value, unlike log error.

Error by Chart Variation. Absolute and signed error differ between the question types due to the magnitude of the values involved. Absolute error differs significantly between some of the

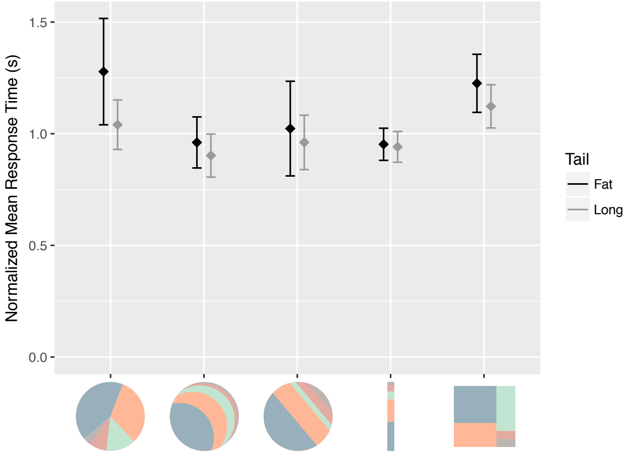

Figure 4: *Response time varies significantly by chart type and tail type (means and 95% confidence intervals).*

chart types when asked about the largest slice, but we found no significant differences in the questions about the middle slice. We performed pairwise, paired t-tests between the pie chart and the four other variations, with a Bonferroni-corrected $\alpha = 0.05/4 = 0.0125$ significance level (for multiple comparisons). The circular slice ($t(75) = -4.76, p < 0.01$) and the treemap ($t(75) = -2.41, p < 0.01$) show a significant difference in the sample means, with the former having lower absolute error, and the latter higher error, than the pie chart.

Error by Slice Asked About. Comparing the slices asked about is challenging because their value ranges were different. As described above, the largest slice had a range of 39% to 79%, with the middle tail ending up with a range of 2% to 26% for the long tail and 5% to 19% for the fat tail condition.

The differences in absolute error are a direct result of the different value ranges. What is notable, however, is the difference in signed error: in the largest slice condition, there was a consistent underestimation of values, whereas the middle slice led to almost universal overestimation (maybe partially due to color [**?**]).

The relative errors shown in Figure 2 are the absolute and signed errors divided by the true percentage. Both differ significantly between the question types (paired t-test, $t(379) = -7.3, p < 0.01$), with the middle slice having generally higher absolute error and leading to slight overestimation (positive signed error), while the larger slices have smaller relative absolute error and lead to slight underestimation.

Error by Tail Type. The questions about the middle segment were asked for two different tail types: fat tail and long tail. Since that narrows the range of values asked about, we report absolute and signed error here, rather than relative error.

We had expected the long tail to fare better, since the slices end up looking more different. We do not find a significant difference in absolute (paired t-test, $t(379) = -0.97559, p = 0.33$) or signed error ($t(379) = -1.4495, p = 0.15$) between the tail types, however (Figure 3).

Response Time. While we expected response times to differ between the different chart types, we did not expect an effect from either the tail type or the slice asked about.

Mean response time (normalized by user) differs significantly depending on which slice we asked about, with the largest slice question ($M = 8.4s$) being faster than the middle slice one ($M = 10.0s$), $t(379) = -2.6215, p < 0.01$. Mean response time also differs by tail type. This was tested only on the middle slice questions, since we did not vary tail type for the largest slice. The long tail leads to faster responses ($M = 9.2s$) than the fat tail ($M = 10.9s$), $t(379) = -3.6263, p < 0.01$.

As expected, response times differ by chart variation (ANOVA, $F(4,755) = 10.8, p \ll 0.01$). Oddly, the pie chart, stacked bars, and treemap perform worse than the circular and straight slice charts (Figure 4). It appears that people spend more time reading the more familiar charts, and are faster with the unusual ones.

8. Discussion

While we expected differences depending on the distribution, we expected those differences in the error not the response time. It is response time that differs, however.

The difference in response time between chart types is confusing. It is unclear why participants consistently spend more time reading the more familiar charts, in particular the pie chart. If the reason is the charts' attractiveness, it may explain their popularity.

Response time differences between question types are notable but not surprising: the larger slice tends to be easier to see and estimate than a smaller one. Which slice of interest to the viewer is not information that usually goes into chart design, however, which might be problematic.

We find no difference in precision between the tail types, but the long tail leads to faster response. This is reassuring, since it is not under the chart designer's control. The difference in response time might be a reflection of the cognitive load being lower when the differences between slices are larger.

The circular slice chart stands out as having lower absolute error than the pie chart and being faster to read as well. We did not expect this novel design to perform better, and would have expected viewers to spend more time answering questions with it, rather than less. Interestingly, the straight-line condition does not do as well.

The poor showing of the treemap for this task is notable: they have higher error and people take longer to read them in all conditions. While many in the visualization community would perhaps want to recommend them over pie charts, we find no evidence that they perform better for a small number of slices.

9. Conclusion

Part-to-whole charts are not widely studied and some of the assumptions about their relative merits appear to be unfounded. We find the treemap to perform worse than the other conditions, in particular the pie chart and the circular slice chart. The latter might be an interesting alternative to the unpopular pie chart, as at least according this study, it performs better than the pie and is read faster.

References

[BHvW00] BRULS M., HUIZING K., VAN WIJK J. J.: Squarified Treemaps. In *Data Visualization Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization*. Eurographics Press, 2000, pp. 33–42. 2

[BK91] BERLIN B., KAY P.: *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1991. 2

[CM84] CLEVELAND W. S., MCGILL R.: Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association 79*, 387 (1984), 531–554. 1

[HB10] HEER J., BOSTOCK M.: Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings CHI* (2010), pp. 203–212. 2

[HS92] HOLLANDS J. G., SPENCE I.: Judgments of Change and Proportion in Graphical Perception. *Human Factors 34*, 3 (1992), 313–334. 1

[KHA10] KONG N., HEER J., AGRAWALA M.: Perceptual Guidelines for Creating Rectangular Treemaps. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 990–998. 2

[Kos19] KOSARA R.: Circular Part-to-Whole Charts Using the Area Visual Cue. In *Short Paper Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)* (2019). 2

[KZ10] KOSARA R., ZIEMKIEWICZ C.: Do Mechanical Turks dream of square pie charts? In *Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)* (2010), ACM Press, pp. 373–382. 1

[SH87] SIMKIN D., HASTIE R.: An Information-Processing Analysis of Graph Perception. 454–465. 1

[Shn92] SHNEIDERMAN B.: Tree Visualization with Tree-Maps: 2-D Space-Filling Approach. *Transactions on Graphics 11*, 1 (1992), 92–99. 1

[SK16] SKAU D., KOSARA R.: Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts. *Computer Graphics Forum 35*, 3 (2016), 121–130. 1, 2

[Spe90] SPENCE I.: Visual Psychophysics of Simple Graphical Elements. *Journal of experimental psychology. Human perception and performance 16*, 4 (1990), 683–692. 1

[TSA14] TALBOT J., SETLUR V., ANAND A.: Four Experiments on the Perception of Bar Charts. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2152–2160. 1

[ZTLS98] ZACKS J., TVERSKY B., LEVY E., SCHIANO D. J.: Reading Bar Graphs: Effects of Extraneous Depth Cues and Graphical Context. *Journal of Experimenal Psychology: Applied 4*, 2 (1998), 119–138. 1