Raising the Bars: Evaluating Treemaps vs. Wrapped Bars for Dense Visualization of Sorted Numeric Data

Mehmet Adil Yalçın *
Keshif, LLC

Niklas Elmqvist †
University of Maryland, College Park

Benjamin B. Bederson §
University of Maryland, College Park

Abstract

A standard (single-column) bar chart can effectively visualize a sorted list of numeric records. However, the chart height limits the number of visible records. To show more records, the bars could be made thinner (which could hinder identifying records individually), and scrolling requires interaction to see the overview. Treemaps have been used in practice in non-hierarchical settings for dense visualization of numeric data. Alternatively, we consider wrapped bars, a multi-column bar chart that uses length instead of area to encode numeric values. We compare treemaps and wrapped bars based on their design characteristics, and graphical perception performance for comparison, ranking, and overview tasks using crowdsourced experiments. Our analysis found that wrapped bars perceptually outperform treemaps in all three tasks for dense visualization of non-hierarchical, sorted numeric data.

Keywords: Visualization; Graphical Perception; Treemaps; Evaluation.

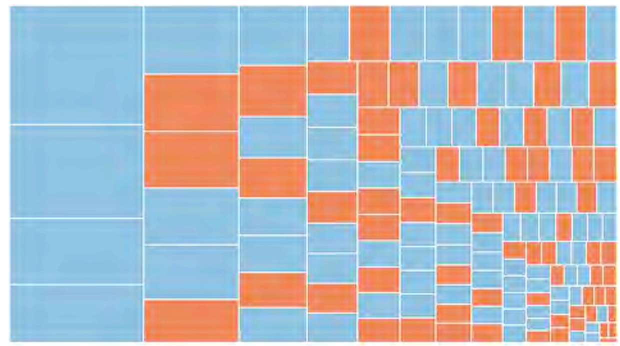
Index Terms: H.5.2. Information Interfaces: User Interfaces.

1 Introduction

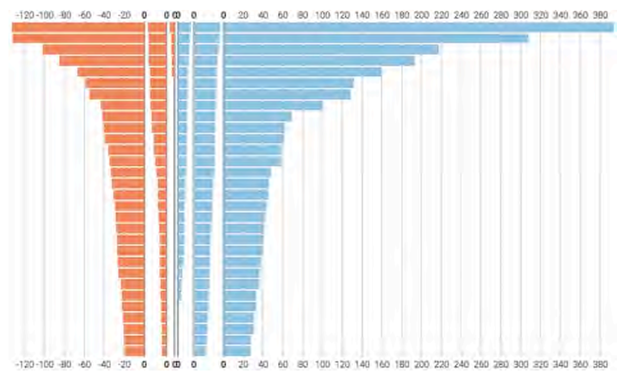
Lists of numeric measurements for specific items—such as country populations, smartphone prices, or university acceptance rates—are ubiquitous. Visualization can amplify people's ability to comprehend data [5], and the standard (single-column, sorted) bar chart does so with perceptual effectiveness and simplicity. However, it can only show a few dozen records given common constrained screen sizes. How can we visualize *more* records—such as 150 countries, 75 smartphones, or 300 universities—in a single chart while maintaining perceptual accuracy for data comprehension? Among potential solutions, (i) larger screen spaces for charting may not be available, (ii) interaction, such as scrolling or focus+context, are not supported in ubiquitous print and image media, and (iii) aggregation of underlying data prevents observing records individually. In addition, there currently exists no detailed evaluation of alternative visualizations and their graphical perception performance targeting this data setting and context.

This paper focuses on *treemaps* [23] and *wrapped bars* [15], which are both dense data visualizations for sorted numeric data that support overviews of all records and comparisons between records. (Figure 1). We consider the *treemap design* because of its common use [1], [32], [33] for presenting large numbers of records without hierarchical structure, although the technique was originally designed for visualizing hierarchical data structures [23]. Visualization tools such as Tableau [29] also include treemaps as a suggested plot for a numeric attribute [25], which leads to its adaptation in various dashboards [8]. We also consider *wrapped bars,* which uses multiple columns for dense visualization of larger datasets. It was (as far as we know) first introduced by Stephen Few [15]. We contribute a detailed analysis of the two techniques, and discuss the use of color and bi-directional axis for visualizing negative values and grouped records, as well as showing record labels.

We report the graphical perception performance of the two techniques through crowdsourced human experiments, comparing them on three complimentary tasks: comparison, ranking, and overview. Our results suggest that wrapped bars outperform treemaps in perceptual accuracy for all three tasks. We also discuss the effects of data density on perception performance.



(a) Treemap



(b) Wrapped Bars

Figure 1. Two visualization techniques showing 150 records. (Top) Treemap, a space-filling design, shows the magnitude by the block size, and the sign by block color. (Bottom) Wrapped bars are multi-column bars, and can organize +/- numbers across two sides. What are the design characteristics of treemaps and wrapped bars for flat data? Which chart design can improve perception for comparison, ranking, and overview under varying data conditions?

* adil@keshif.me † elm@umd.edu § bederson@umd.edu

2 Related Work

Increasing data density is among Tufte's visualization guidelines [31]. Another goal of effective visualization design is graphical perception accuracy, requiring a careful design process, and evaluation of alternative designs.

Fekete et al. [13] demonstrated the use of treemaps to visualize up to a million records on large screens. Under such settings, many records occupy just a few pixels, and the visualization primarily supports perceiving overviews of record groups, and comparison of larger records. In our study, we aim for high legibility of every value in the chart, thus avoiding large data scales in a limited chart area. Kong et al. [24] compared the perceptual performance of treemaps to *single-column* bar charts *in a hierarchical setting* with up to 8,000 records at the deepest branches in a 600x400 pixel chart size. They reported, "As data density increases, treemaps become faster than bar charts while exhibiting equivalent accuracy." This effect may be due to the tiny size of *single-column* bars at dense displays that makes them harder to observe, which could be mitigated by using *multiple columns*. Their study did not consider the use of treemaps in a non-hierarchical setting, or data overview and ranking tasks. Therefore, our study is different than existing studies because of its motivation, data types, and inclusion of visual overview and ranking tasks.

Among the techniques for dense information visualization, *horizon charts*  [14] display time-series data in a compact chart height using a refined filled line chart. They divide the numeric data axis into equal sized bands, and collapse the bands while adjusting the color darkness per band. The chart height is reduced in the order of the number of bands, keeping the overall trend visible. Heer et al. [20] studied perception of horizon charts and identified the effect of banding and chart height on estimation accuracy and speed. Javed et al. [22] discussed alternatives to visualizing multiple time series in a limited area, including braided charts , and assessed perceptual performance with lab experiments. Fuchs et al. [16] evaluated alternative glyph designs for time series data in small multiple settings, where each glyph represents dense temporal data.

Evaluating the graphical perception of visualization design has a long history in the field of statistical graphics. The comparison task used by Cleveland and McGill in 1984 [6] has become an established method to assess graphical perception. Talbot et al. [30] extended their results on bar chart perception to better understand the reasons for performance differences across aligned and nonaligned bars, and the effects of separation and distracting bars. Perceptual studies have been extended to a crowdsourced methodology by Heer et al. [19]. Their results were aligned with results in lab settings, albeit with more variance. They found that uncontrollable display size and viewing distance across crowdsourced participants can be balanced by recruiting more participants from a wide online population than traditional lab settings with few participants. Crowdsourcing has been used to evaluate graphical perception experiments in other studies as well [3], [28].

3 Dense Visualization Design for Sorted Numbers

We consider the following objectives in the visualization design space for sorted numeric data:

O1: Each record is *perceptually distinguishable*. All records must fit within the chart, and must be presented with their own visual glyph. This ensures that all records can be observed and compared visually.

O2: An *overview* of all records is visible without interaction. This objective fits the use of visualization in static media, such as in print and in social media image previews. While interaction can be used to reveal multiple perspectives and views over time, it is beyond the scope of graphical perception studies. The direct perceptual response to a visualization is critical, as any interaction is likely to slow down perception.

O3: The records are visually sorted by value. This improves the visual structure, and simplifies assessing min/max, variance, and rankings. Without such order, the visual representation of data would be weaker in revealing data distribution characteristics.

O4: The records do not overlap visually. We expect that avoiding overlaps will reduce clutter, increase the ability to identify patterns, and reduce amount of hidden and lost information [11].

Table 1 presents a summary treemaps and wrapped bars, which both meet these design objectives. *Treemaps* are a commonly used chart type that uses a space-filling technique, making use of all the chart pixels to encode the data by area. *Wrapped bars* increase the number of visible records by utilizing a multiple column chart. The two chart designs target similar chart sizes and aspect ratios. These techniques can handle more than a standard bar chart even though the chart area still bounds the number of perceptually distinguishable records.

To motivate our objectives and their implications, let us also consider alternative techniques that do not meet the objectives. **(i)** Aggregated visualizations [12], such as histograms, violate O1 as they do not show each record individually. **(ii)** Standard bar charts can be extended beyond the visible area with scrolling. This fails to show a complete overview (O2), and requires interaction to observe different sections of data. **(iii)** Standard bar charts can show more data using shorter bars, however this makes individual records harder to observe (O1). **(iv)** A space-filling design could encode numeric data by color on fixed block size, instead of by area. However, the number of colors that can be effectively compared is fairly limited [26]. **(v)** Circular encodings, such as packed bubble charts, are not strong for perceptual comparison, and use screen space ineffectively.

Alternative contexts, such as systems for visual analytics or interactive reporting, may have different objectives that would benefit from the use of interaction, such as scrolling or focus+context views [17]. In such cases, visualization designs that do not meet all of our stated objectives may be preferable.

	Treemaps	Wrapped Bars
Visual Encoding	Space-filling rectangular area	Length
Baseline and Grid-	Not available	One baseline per column. Supports gridlines
Block order by value	↓ & → (Not guaranteed)	Columns first ↓ , then Rows →
Filled pixels	All	Partial - Depends on the distribution and variance of data
Adding columns	Not available	Shrinks bar width ↔
(-) Negative Values	Another visual variable (color) is required.	Bi-directional length encoding can visualize negative values.
Grouping Records	Color-coding blocks per group	Color-coding bars per group
Label Display	Must be within blocks	Can be within or next to blocks (using more ↔ space)
Other properties	Reveals and emphasizes part-of relations.	Columns can be separated by additional gaps.

Table 1. Summary of treemaps and wrapped bars visualization techniques that satisfy the design objectives of this paper.

3.1 Treemap Technique

Treemaps are a space-filling visualization technique where each data record is visualized using a rectangular block, and the rectangular area encodes the data value. Treemaps were originally designed to visualize hierarchical data groupings [23] using a nested block layout. Treemaps are also commonly used in practice to display *non-hierarchical* data in order to scale to more records than possible with a standard bar chart.

An advantage of the space-filling design of treemaps is that all pixels are used to visualize data. Treemap algorithms commonly aim to generate a layout with the largest block in the top-left corner, the smallest on the bottom-right corner, and blocks ordered along one direction in decreasing size first. Yet, the optimized layout does not guarantee a sorted order, thus relaxing the objective O3. The area encoding used by treemaps has been shown to be perceptually less effective for comparison task compared to linear encodings of length and position on a shared baseline [6], [19]. Other perception experiments report that rectangles with lower aspect ratios improve perceptual accuracy and extreme aspect ratios should be avoided [19], [24]. Thus, a squarified treemap layout [4], which aims to avoid elongated rectangles, is commonly preferred, and is used as the layout in this study.

3.2 Wrapped Bars Technique

Wrapped bars [15] use multiple columns of aligned bars, which can effectively show more records than a single-column bar chart. Where new bars would extend vertically beyond the chart area, they are wrapped to start a new column, similar to the two-column text layout of this paper. The bars are comparable across the columns since the length encoding has the same unit scale in all columns. The column width decreases as the size of the records in the column decreases. The columns may be separated with a horizontal ↔ gap to emphasize separation, thus improving readability.

Given a fixed chart area and bar height, adding more records may result in additional columns. To make space for new columns, existing bars must shrink horizontally ⇆, in turn decreasing data resolution and perceptual accuracy. Increasing bar height ↕ for a fixed record count may have the same effect, i.e. as bars get taller, they get narrower (Figure 2). Thus, the column layout influences the aspect ratio of bars.

3.3 Grouping, Negative Values, Bi-directional Axis

Grouping data is an important design consideration. Figure 2 shows sample data that represents two groups. Treemaps can group multiple records spatially to represent the distribution of group totals, while both design approaches can use color to show the group association of each record, and emphasize the overall distribution of values.

Next, we consider how to represent negative values (Figure 1). In treemaps, block area is implicitly positive and cannot be used to encode negative values. The sign of the values is therefore encoded by color. In wrapped bars, the baseline can be moved towards the middle of the chart, and the bars can then be extended in both directions (←,→) to encode the sign. Using color can emphasize the column of the sign flip. In summary, treemaps can show sign preferably using color encoding, while wrapped bars can group +/- values along two sides of the baseline.

When the records are sorted by non-numeric criteria (such as alphabetically), there are various impacts on the visualizations. In treemaps, the layout algorithm can be based on alternative metrics, and the chart space would still be used efficiently. Wrapped bars, on the other hand, have a problem in that the columns would not get narrower since the small records are not grouped, and thus horizontal space would not be used effectively.

Lastly, we also consider displaying record labels (Figure 4) As the visual layout of treemaps strictly follow the data distribution, labels must be placed within the blocks, and smaller values offer smaller label space. Wrapped bars are more flexible. Labels can be placed within or next to bars. They may also be shown for all columns or for a selected column [15]. Alternatively, record labels can be displayed as tooltips on mouse-over to individual records in interactive applications of the both techniques.

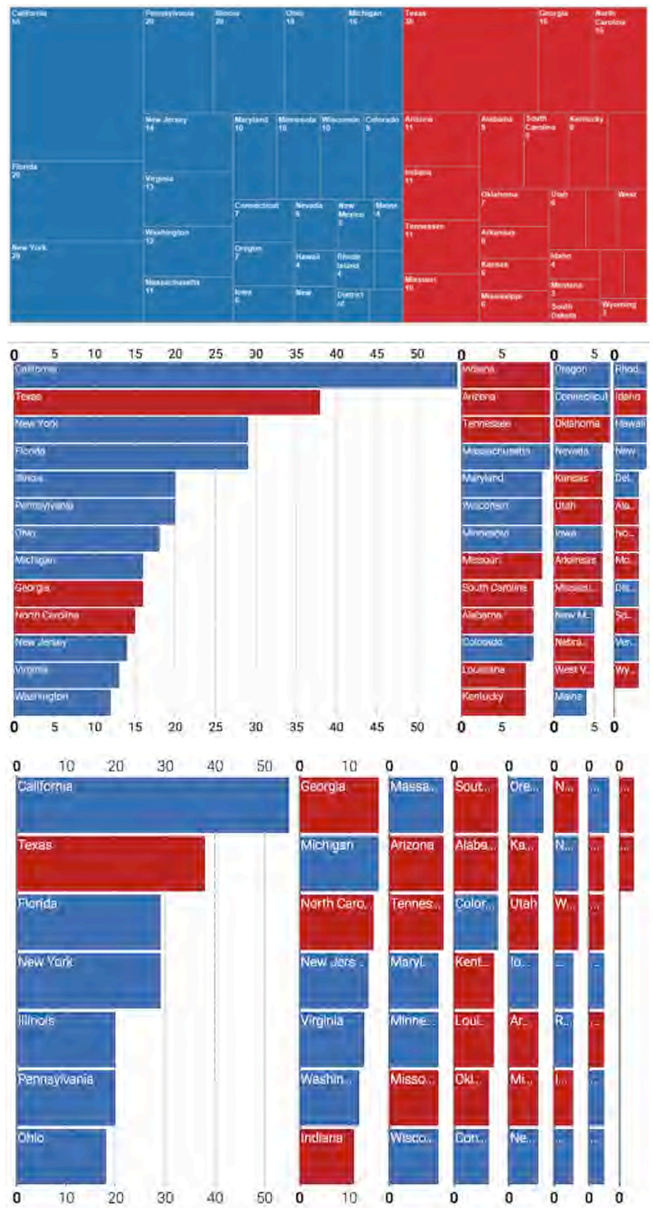


Figure 2. Electoral vote results for the 50 states in the U.S. 2012 presidential elections. Each state has a number of electoral votes (block size) and a winning party (grouped as Democrat or Republican). (Top) In treemaps, records can be grouped by winning party (from [27]). The distribution across two parties is emphasized. (Middle & Bottom) Wrapped bars can order states by electoral vote and show groups by color. Among the states with higher votes (leftmost column), Democrats are more frequent. The bottom chart has thicker and shorter bars, and more columns compared to the middle chart. Notice that these features interact with each other.

4 Graphical Perception Experiments Setup

To evaluate the graphical perception performance of the two visualization techniques, we designed online crowdsourced experiments for three task types under varying data densities. We first describe the three tasks and the shared settings and procedures in conducting these experiments. We follow with the detailed description and results for each task.

4.1 Tasks

To cover a wide range of perceptual characteristics of the alternative designs, we chose three graphical perception tasks (Figure 3) such that the answer would be *(i)* data-driven (i.e. changing data would predictably influence the answer), *(ii)* can be given within a few seconds following a quick impression in a casual use, *(iii)* based on a single chart. The tasks were designed to apply fairly to all chart designs. We present a summary of the three tasks below.

Comparison of two records: Two records are highlighted. The participant determines which is larger and by how much. Comparison is the basis of visualization. However, this task focuses on two marks, and does not require reading the whole chart. This task is thus insufficient for assessing the perception of data distribution.

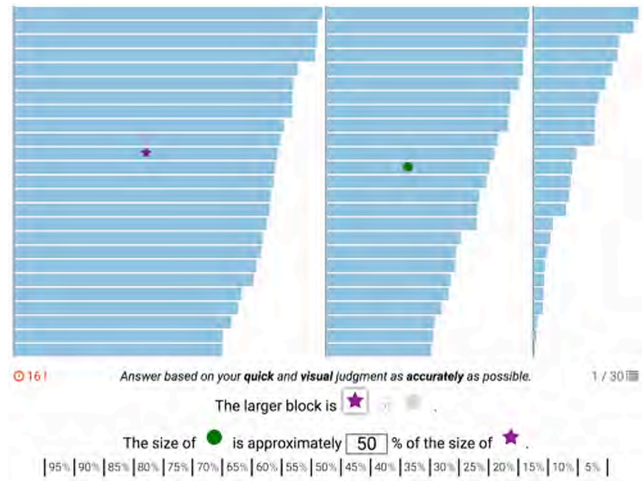
Ranking of a record: The participant determines the rank of a highlighted record among all records. Ranking is a common task, such as finding the rank of a country or a university within an ordered list. This task requires observing the complete data distribution in relation to the focal record. While the rank of each record can be displayed by default (increasing chart ink) or on interaction (with a tooltip), graphical perception allows a quick assessment of the record ranks. When the data is visually sorted, the position of the record among all records suggests its rank. Thus, sorted visualizations avoid tedious size comparison across all records for ranking, and ranking becomes independent of the distribution characteristics.

Overview of all records: The participant is asked to assess whether a given statement on data distribution matches the displayed data. This task is solely based on interpretation of the overview of data. No individual records are highlighted, and the data is generated with specific targeted distribution characteristics. Our rationale is that understanding the overall distribution of data, without anchoring to a set of selected marks, is also an integral part of visual data comprehension.

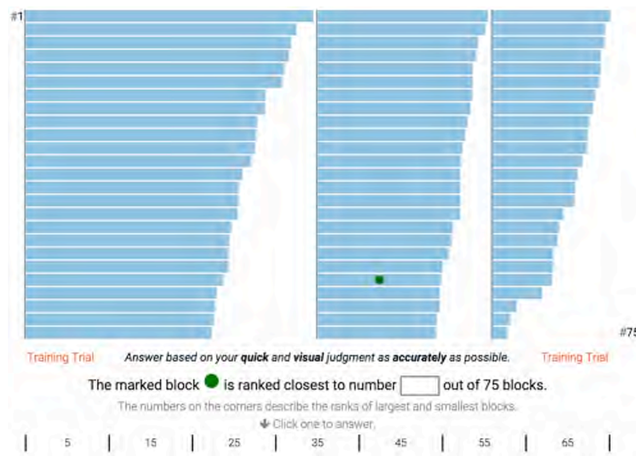
Among other overview tasks, finding min/max is trivial in sorted data. While mechanical computation of *average* and *variance* is easy, such numeric characteristics are not naturally perceptible given many (50+) records, and can be easily annotated on the chart if necessary. We also avoided tasks that would require interaction within the chart to answer, such as clicking on a block that may best present the mean or the median. The measurements could include selection (motor-skill) errors that may negatively influence the measurements. As we aimed to assess how well the visualization by itself can communicate the data, we did not use the line-up protocol [21], which presents multiple charts with a presumable outlier for hypothesis testing. Charts are commonly shown in isolation to illustrate a single set of measurements, rather than with multiple alternatives that may serve as anchors to understand distribution differences. Overview tasks can also require comparing characteristics across data groups within a single chart, such as the moving average over time series [7], or differing glyphs per category in scatterplots [18]. We avoided such tasks since they require a design change, either using color or bi-directional multi-columns, which are not applicable consistently across all chart types in a similar fashion.

4.2 Experimental Design and Chart Parameters

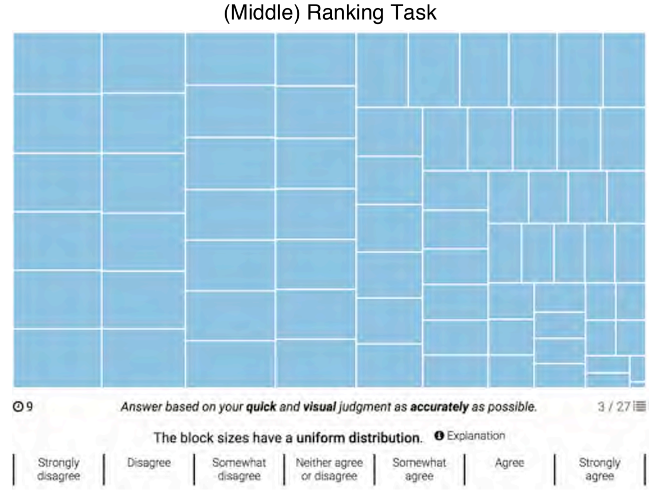
Each participant answered multiple questions (trials) for a *fixed* graphical perception task on a *fixed* chart type with variations in chart density (record count): 75, 150 or 300 records. All charts had 800×450px size (16×9 aspect ratio), a comfortable size for medium to large personal display devices. Figure 4 shows 300 records within the selected chart size, which create a dense setting for casual visualizations; doubling the scale would decrease the



(Top) Comparison Task



(Middle) Ranking Task



(Bottom) OverviewTask

Figure 3. The graphical perception tasks of our experiments.

readability of individual records. Treemaps were generated using the squarified layout of d3.js [2] with default parameters (v3.5.5) and 2px border between blocks. For wrapped bars, gridlines were hidden except the baselines (to maintain consistency with treemaps which do not have guidelines). Bar heights were fixed to 16px, resulting in 25 bars per column. The number of columns was therefore dependent on the record count (3, 6 and 12 columns respectively for 75, 150 and 300 records). The experiments used a 2-pixel ↕ gap between rows, and a 5-pixel ↔ gap, in order to sufficiently separate the bars visually, similar to the gaps and borders used in treemaps. Charts did not include textual record labels.

4.3 Participants

In our experiments, each chart type and task combination (2×3) was answered by 20 participants, resulting in 120 crowdsourced participants across the 3 tasks. We ensured that a participant could not participate in multiple experiments, and thus a participant responded to only a single chart type and a single task type.

We recruited participants using Amazon Mechanical Turk. Our experiments were composed of quick, casual tasks that did not require significant training and could be completed remotely by participants. The qualification requirements were set to historical performance of at least 90% approval rate and at least 1,000 HITs completed to make sure that workers demonstrated good performance in crowdsourced tasks before. We geographically limited participation to the U.S. We rejected participation from mobile devices and screen sizes with less than 1280×800 pixel resolution to ensure their screens could display the tasks fully. We did not collect demographics, since that would increase the burden on participants without substantial analysis benefits. Our results are based on the diversified Amazon Turk worker pool across gender, age and education levels [27]. Thus, across experiment conditions, participant demographics are not controlled, and are expected to be randomized. We awarded the participants with a targeted $8/hour rate, based on expected task durations.

4.4 Training and Other Procedures

Our experiments included multiple approaches to train the participants and to collect high quality data. All experiments included *training trials* using simpler versions of the task to ensure that the participants were able to understand the task. The participants could only proceed when they answered the training trial questions correctly. They were allowed to repeat trials until they answered correctly. In experiment trials, participants were not allowed to change their answers. To help participants stay focused while repeatedly answering the same task under different density conditions, we presented a training trial after ⅓ and ⅔ of experiment trials. As with the initial training, participants needed to answer these trials correctly to proceed, and they could repeat their attempts until finding the correct answer.

We also prepared animated training sequences to explain the chart designs by animated transitions from standard (single-column) bar charts. In this sequence, the participant first saw 75 records in a single-column overflowing chart, with an animated scroll showing all the records. Then, on a button click, the single-column chart was transitioned to the chart type of the experiment with animation. The participant observed three data distributions and transitions, and could replay the sequences. The animated sequences were shown as the first step into the study.

When the participant selected an answer, the answer and response time were recorded, and the study progressed with a new trial. The marked block(s), if the task required, were visible until the task was answered. We displayed a time ticker next to the task. At 10 seconds, the ticker changed to display **10!** (note the exclamation point) to alert the participants of the passing time.

After running the experiments, we confirmed that the analyzed data correctly represented the experimental settings, with the correct number of trials and variations per participant, and the number of participants per trial group.

5 Task 1: Comparison

For the comparison task, the participant observed a chart with two highlighted blocks, and estimated what percentage the smaller block is of the larger block. We highlighted the selected records with colored marks (●,★) placed in the middle of the record blocks. We first asked, "The larger block is A or B?" with random A-B order, where A and B represent the visual marks. After selecting an answer (e.g., B), we then asked, "The size of A is approximately [__] % of the size of B." with A/B order based on the previous answer. The answer options were multiples of 5%, ordered from 95% to 5% under the question. Our design aimed to assist participants in focusing on their judgment at commonly expressed perception granularity (5x%) as reported in previous studies [24], [30]. Each participant answered 30 trials in randomized order on a single chart type with 10 conditions on percentage difference, and 3 conditions on chart density (record count).

Sixty uniformly distributed random data configurations were generated, as a combination of 10 true percentages with 3 record counts for each setting (75, 150, 300). We selected 10 true percentages at non-regular points in relation to 5% intervals (8%, 17%, 23%, 38%, 47%, 53%, 62%, 77%, 83%, 92%), such that the accuracy of an answer could be measured within 1%. The larger value was picked randomly among the top 25% of the sorted data. The smaller value was computed using the true percentage, and it replaced the smallest value. The same data configurations were used across the chart types. We used five training trials with 75 records and (10%, 30%, 50%, 70%, 90%) for true-percentages and answer options.

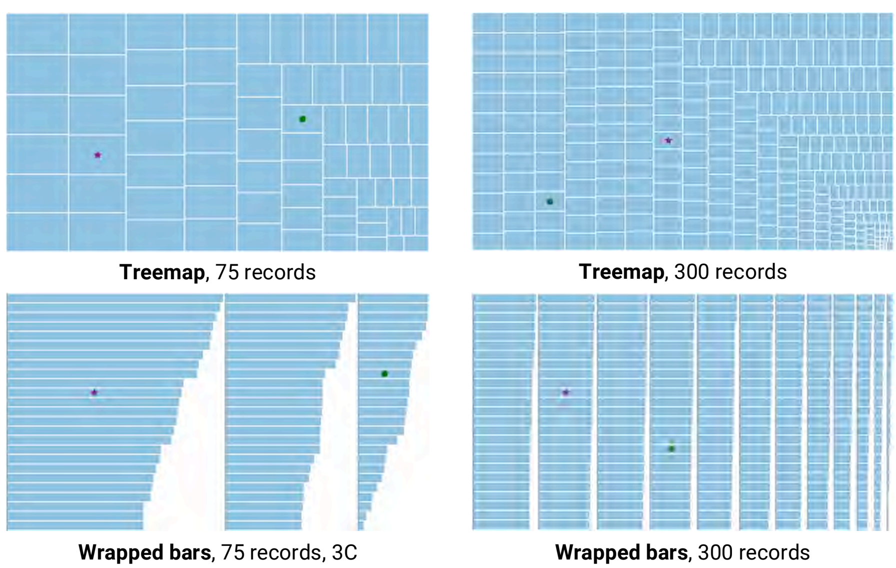


Figure 4. Sample charts from the comparison task.

Figure 5. Results on accuracy (% error) in comparison and ranking tasks across chart types and data densities. Top two rows show the percentiles of observed errors of all densities in 10% increments. ▌shows the median. ▲ shows the mean of values within 10-90 percentile. The six rows below show bootstrapped error means per chart type and record count. ● shows the bootstrapped mean. The bars show 95% confidence intervals. Each row includes 20 participants. Top two rows have 600 responses each, and rows below have 200 responses each.

Results and Discussion

To analyze the perceptual performance in comparison, we measured the error as the absolute difference between the response percentage and the true percentage difference of marked blocks. Figure 5 shows the resulting error measurements, grouped by chart type and data density. To analyze the effect of data density using 75, 150 and 300 records, we use the bootstrapped group error mean with 95% confidence intervals [10]. Bootstrapping produces statistical estimates based on resampling the observations with replacement. It has been advanced in psychology [9] to address the shortcomings of significance testing and p-values, and we adopt it here for similar reasons.

Our participants estimated relative sizes of records more accurately using wrapped bars compared to treemaps, both for overall and for different densities. Increasing data density did not have a substantial effect on treemap accuracy. However, at the highest density level (300 records), wrapped bars performed marginally worse, and its performance gap from treemaps is narrower. Overall, smaller blocks in denser charts are expected to decrease accuracy performance. The response times was similar, with treemaps performing marginally slower (Figure 6).

We also applied standard parametric statistical tests to responses in data density setting with mixed linear two-way, factorial model with interaction using the subject as random effect. Results confirm the significant effect of chart type on error ($F(1, 1156) = 6.19, p = 0.013$).

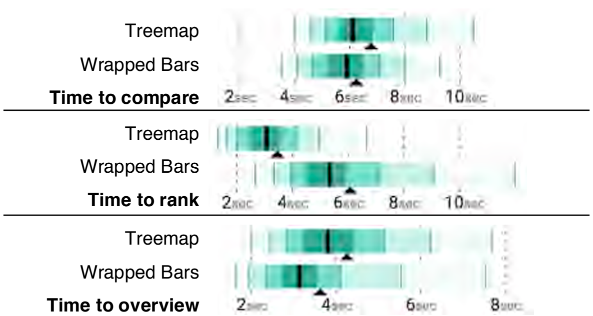


Figure 6. Response time overviews. The charts show the percentiles of observed errors of all densities in 10% increments. ▌shows the median. ▲ shows the mean of values within 10-90 percentile.

6 Task 2: Ranking

For the ranking task, the participants observed a chart with a block marked with ● placed in the middle of visible portion of the block. We asked, "The marked block ● is ranked closest to number [__] out of N blocks", where N is the number of blocks. The marked blocks were generated using 10 percent-based rankings (8%, 17%, 23%, 38%, 47%, 53%, 62%, 77%, 83%, 92%), rounded to an integer. For example, a 23% ranked record across 150 records has rank 35. We presented 14 options, evenly spaced across all records and in absolute ranks since it is a natural form of interpreting ranks given a variety of scale. Each participant answered 30 trials in randomized order on a single chart type. Across two trial groups, 40 participants answered 1,200 rankings. The data was generated using random normal distribution with μ:2 and σ:0.8, with absolute values. We showed index labels for the first and last ranked records on the chart corners to help reading the chart structure. We used seven training trials with 75 records and (5, 15, 25, 35, 45, 55, 65) options for true-ranks and answers.

6.1 Results and Discussions

We measured the accuracy of a ranking response as a percent difference from true absolute rank normalized by the number of blocks (max rank). Figure 5 shows the resulting error measurements, grouped by chart type and density. To analyze the effect of data density using 75, 150 and 300 records, we used bootstrapping for group mean to generate 95% confidence intervals.

Our participants estimated the rank of highlighted records more accurately using wrapped bars compared to treemaps, both overall and per different chart densities. Similar to the comparison task, accuracy of treemaps did not significantly change in different data densities, while the performance of wrapped bars showed a decrease in highest density setting, closing the performance gap between charts marginally. Although treemaps showed lower accuracy performance, the response time with treemaps were significantly faster than with wrapped bars (Figure 6). We believe that participants spent more time reading the column structure of wrapped bars to have higher accuracy in the ranking task. Therefore, the responses can be explained as a tradeoff across time and accuracy across the two techniques.

We also applied standard parametric statistical tests to responses in data density setting with mixed linear two-way, factorial model with interaction using the subject as random effect. The results show significant effect of chart type ($F(1, 1156) = 19$, $p < 0.0001$).

7 Task 3: Distribution Overview

For the overview task, the participant stated their agreement to a data distribution statement given a chart, on a 7-point Likert scale as shown in Figure 3. The chart and the question of a trial were selected among three distribution characteristics, resulting in nine permutations. Each trial group was based on three conditions on chart density, and each participant answered 27 experiment trials in randomized order. We generated 10 groups of random data distributions for 27 trials. Each data group was answered by two participants, totaling to 20 participants answering 540 trials.

The three data distribution characteristics of this experiment (with explanations presented on "ℹ Explanation" mouse-over) were: (i) Uniform distribution, i.e. "There are blocks of all possible sizes". (ii) Skewed distribution, i.e. "There are a few blocks that are substantially larger than all the rest". (iii) Normal distribution, i.e. "There are more medium-sized blocks than small and large blocks." In the animated training sequence, we presented one sequence for each data distribution with a description of the data distribution characteristic. After the sequence, three training trials were shown with "agree/disagree" options. The experiment advanced when the response matched the data distribution.

7.1 Results and Discussions

We identified each response as *true*, *false*, or *no decision* based on agreement with the correct response to the data distribution statement, and converted the scale from agreement to response correctness. For example, a "*strongly agree*" response to a *uniform* statement for a *uniform* data distribution is "*strongly true*", and "*somewhat disagree*" response to a *normal* statement for a *skewed* distribution is "*somewhat false*". Table 1 presents an aggregated visual summary of the responses across correctness, confidence, and different chart types under various chart densities.

Treemaps had a higher percentage of false answers compared to wrapped bars. Specifically, treemaps had 46% false responses, while wrapped bars had 30%, with 540 responses in total for each. Regarding the confidence level of the responses, wrapped bars had higher ratio of "strongly" confident (false or true) responses in most settings. Analysis of the response time for the overview task (Figure 6) shows that treemaps were also slower than wrapped bars.

We also performed a standard statistical analysis based on a generalized linear mixed model for the binary outcome (with no-decision responses considered false). We detected significant effect of the chart type ($F(1, 38) = 16.71$, $p = 0.0002$), using Tukey HSD post-hoc analysis.

The accuracy effect across chart type vs. distribution characteristic is shown in Table 2. The charts show similar performance under normal distribution, however treemaps performed significantly worse for skewed distribution (46% vs. 74% of true responses), and marginally worse for uniform distribution.

		Str False	False	Sw. False	False (All)	No Dec.	True (All)	Sw. True	True	Str True	# of Resp.'s
All		7	20	11	38	3	59	10	28	21	1080
Treemap		7	25	14	46	3	51	12	27	13	540
Wrapped Bars		8	14	8	30	2	68	9	29	29	540
Density 75 Records	TM	2	22	17	42	2	57	16	30	12	180
	WB	6	17	11	33	2	65	9	28	28	180
Density 150 Records	TM	7	29	14	50	2	48	11	26	11	180
	WB	7	11	8	26	3	71	10	33	28	180
Density 300 Records	TM	11	26	9	46	5	49	8	26	16	180
	WB	11	15	6	32	2	67	9	27	11	180
		%	%	%	%	%	%	%	%	%	#

Table 1. Response accuracy of the overview task. Accuracy values are shown in percentage and color-coded, with darker color showing larger value (True: green. False: red. No-decision: yellow). For example, of the 540 responses given for treemaps, 46% were false, while only 30% were false of the responses to the wrapped bars.

	Uniform	Normal	Skewed	# of Resp.'s
Total	56%	58%	65%	1080
Treemap	51%	58%	46%	540
Wrapped Bar	62%	57%	74%	540

Table 2. Accuracy (ratio of true responses) across data distribution and chart types, based on the data density setting. Values are color coded from red to green, with the white midpoint at 61%, the accuracy considering all 1620 responses.

8 Summary of Experiment Results and Implications

Overall, our results and analysis show that wrapped bars yield higher perceptual performance compared to treemaps in a non-hierarchical setting. Its performance is likely due to its clean, easy to interpret, non-overlapping design, inherited from standard bar charts and extended with a natural reading order across different columns. Its design can be further extended by using color and bi-directional encoding, and its flexibility to show labels in various forms. Therefore, we believe that wrapped bars should be considered over treemaps as an effective and flexible design to present dense, sorted, non-hierarchical numeric data.

Treemaps performed worse or equal to wrapped bars in all our task types. Their lower performance for comparison is predictable since treemaps rely on area assessment instead of length assessment. Its lower performance for ranking reflects its relaxed ordering/layout strategy. Results from the overview task show that treemaps do not outperform wrapped bars there either. Overall, our results suggest that treemaps are not a preferable design when records do not have an explicit hierarchy. The use of treemaps outside the context it was designed for, i.e. presenting hierarchical data through grouping records, does not lead to highly accurate graphical data perception, even though it can utilize all pixels in a given chart area.

9 Limitations and Future Work

Our experiments focused on basic chart designs without labels, legends, or axis. Displaying labels may impact chart readability. We did not evaluate designs with color or bi-directional axes, or display axis labels or gridlines in wrapped bars, to maintain comparability with treemaps. Including guidelines is likely to improve accuracy for wrapped bars, further strengthening its advantages.

We reported perception results from data densities of up to 300 records in an 800×450 pixel chart area, with randomly generated uniform, normal, and skewed distributions, while targeting a casual use by general audiences. If these settings are adjusted and experienced data analysts are considered, the data density may be increased further in future studies. Our findings may not extrapolate to higher data densities, smaller (mobile) or larger displays. Increasing data densities on highly skewed data may amplify the strength of treemaps with its non-overlapping, space-filling design, and emphasis of part-of-whole relations. In addition, the wrapped bars of our experiments had all their columns and rows full. In practice, it is likely that the final column has fewer bars/records than supported, such as the right-most columns in Figure 2. The perceptual influence of such an imbalance in wrapped bars for novices can be further studied, especially for ranking and overview. The previous study by Talbot et. al [30] on the effects of bar separation and alignment for comparison task can inform implications of our study, and future studies can focus on more specific variations, such as whitespace, for wrapped bars.

Lastly, our analysis is based on crowdsourced experiments that have limited training opportunities and cannot control for multiple experimental characteristics. Our participants may not have been familiar with either chart design, and had limited experience performing the tasks similar to our study. They also probably had varying experimental conditions including differences in screen size and quality and physical layouts. Future studies may extend our results and analysis with variations in data density (record count), data distributions, experiment setup, and participant backgrounds.

10 Conclusion

In this paper, we discussed and evaluated two alternative chart designs for dense visualizations of numeric data. Specifically, we compared wrapped bars with non-hierarchical treemaps. We analyzed the design characteristics of the techniques under various use cases and settings. We evaluated perceptual characteristics of the two charts using crowdsourced graphical perception experiments based on comparison, ranking, and overview tasks. Our results suggest that treemaps perform less accurately for visual comparison, ranking and overview tasks, when compared to wrapped bars technique for non-hierarchical numeric datasets.

Acknowledgments

The authors wish to thank Sriram Karthik Badam for his help, and the study participants for their time. This work was partially supported by the National Socio-Environmental Synthesis Center (SESYNC) through a grant from the U.S. National Science Foundation (# DBI- 1052875) to the University of Maryland. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Bloomberg Visual Data, "Which companies are biggest?" [Online]. Available: http://www.bloomberg.com/visual-data/industries/q/biggest-companies. [Accessed: 22-Sep-2015].

[2] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[3] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete, "A Principled Way of Assessing Visualization Literacy," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1963–1972, Dec. 2014.

[4] M. Bruls, K. Huizing, and J. J. van Wijk, "Squarified Treemaps," in *Data Visualization 2000*, D. ir W. C. de Leeuw and ir R. van Liere, Eds. Springer Vienna, 2000, pp. 33–42.

[5] S. K. Card, J. D. Mackinlay, and B. Shneiderman, "Readings in Information Visualization: Using Vision to Think," S. K. Card, J. D. Mackinlay, and B. Shneiderman, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 579–581.

[6] W. S. Cleveland and R. McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, Sep. 1984.

[7] M. Correll, D. Albers, S. Franconeri, and M. Gleicher, "Comparing Averages in Time Series Data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 1095–1104.

[8] A. Crahen, "Human Trafficking #MakeoverMonday." [Online]. Available: http://tabsoft.co/1NK9GoY. [Accessed: 28-Apr-2016].

[9] G. Cumming, "The New Statistics Why and How," *Psychological Science*, vol. 25, no. 1, pp. 7–29, Jan. 2014.

[10] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, pp. 189–212, 1996.

[11] G. Ellis and A. Dix, "A Taxonomy of Clutter Reduction for Information Visualisation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1216–1223, Nov. 2007.

[12] N. Elmqvist and J.-D. Fekete, "Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 439–454, May 2010.

[13] J.-D. Fekete and C. Plaisant, "Interactive information visualization of a million items," in *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002*, 2002, pp. 117–124.

[14] S. Few, "Time on the Horizon," *Visual Business Intelligence Newsletter*, 2008.

[15] S. Few, "Wrapping Graphs to Extend Their Limits," *Visual Business Intelligence Newsletter*, 2013.

[16] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, "Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2013, pp. 3237–3246.

[17] G. W. Furnas, "A Fisheye Follow-up: Further Reflections on Focus + Context," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2006, pp. 999–1008.

[18] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri, "Perception of Average Value in Multiclass Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2316–2325, Dec. 2013.

[19] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 203–212.

[20] J. Heer, N. Kong, and M. Agrawala, "Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2009, pp. 1303–1312.

[21] H. Hofmann, L. Follett, M. Majumder, and D. Cook, "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2441–2448, Dec. 2012.

[22] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical Perception of Multiple Time Series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 927–934, Nov. 2010.

[23] B. Johnson and B. Shneiderman, "Tree-maps: a space-filling approach to the visualization of hierarchical information structures," in *IEEE Conference on Visualization, 1991. Visualization '91, Proceedings*, 1991, pp. 284–291.

[24] N. Kong, J. Heer, and M. Agrawala, "Perceptual Guidelines for Creating Rectangular Treemaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 990–998, Nov. 2010.

[25] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show Me: Automatic Presentation for Visual Analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1137–1144, Nov. 2007.

[26] S. N. Pattanaik, J. A. Ferwerda, M. D. Fairchild, and D. P. Greenberg, "A Multiscale Model of Adaptation and Spatial Vision for Realistic Image Display," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1998, pp. 287–298.

[27] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 2863–2872.

[28] D. Skau, L. Harrison, and R. Kosara, "An Evaluation of the Impact of Visual Embellishments in Bar Charts," *Computer Graphics Forum*, vol. 34, no. 3, pp. 221–230, Jun. 2015.

[29] C. Stolte and P. Hanrahan, "Polaris: A System for Query, Analysis and Visualization of Multi-Dimensional Relational Databases," in *IEEE Symposium on Information Vizualization*, Washington, DC, USA, 2000, p. 5–.

[30] J. Talbot, V. Setlur, and A. Anand, "Four Experiments on the Perception of Bar Charts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2152–2160, Dec. 2014.

[31] E. Tufte, *The Visual Display of Quantitative Information*, 2nd edition. Cheshire, Conn: Graphics Pr, 1983.

[32] J. Underwood, "Data Visualization Best Practices 2013," 2013. [Online]. Available: http://www.slideshare.net/idigdata/data-visualization-best-practices-2013. [Accessed: 23-Apr-2016].

[33] "USPTO PatentsView." [Online]. Available: https://www.cssipdata.org/patentsview/#visual. [Accessed: 22-Sep-2015].