Improving Perception Accuracy in Bar Charts with Internal Contrast and Framing Enhancements

Jose Díaz
ViRVIG Group - UPC
Barcelona, Spain
Email: jdiriberri@cs.upc.edu

Oscar Meruvia-Pastor
Memorial University of Newfoundland
St. John's, Canada
Email: oscar@mun.ca

Pere-Pau Vázquez
ViRVIG Group - UPC
Barcelona, Spain
Email: pere.pau@cs.upc.edu

***Abstract*—Bar charts are among the most commonly used visualization graphs. Their main goal is to communicate quantities that can be visually compared. Since they are easy to produce and interpret, they are found in any situation where quantitative data needs to be conveyed (websites, newspapers, etc.). However, depending on the layout, the perceived values can vary substantially. For instance, previous research has shown that the positioning of bars (e.g. stacked vs separate) may influence the accuracy in bar ratio length estimation. Other works have studied the effects of embellishments on the perception of encoded quantities. However, to the best of the authors' knowledge, the effect of perceptual elements used to reinforce the quantity depicted within the bars, such as contrast and inner lines, has not been studied in depth. In this research we present a study that analyzes the effect of several internal contrast and framing enhancements with respect to the use of basic solid bars. Our results show that the addition of minimal visual elements that are easy to implement with current technology can help users to better recognize the amounts depicted by the bar charts.**

***Keywords*-Information Visualization, Graphical Perception, Evaluation, User Study**

I. Introduction

A chart's ultimate goal is to communicate a certain set of quantities that can be visually compared. Thanks to the popularization of infographics, the arrival of the so-called data-based journalism, and the explosion of visualization tools, it has become easy for designers and writers of all backgrounds to generate them in many different ways. Among them, bar charts are a very common visual communication tool. Data mined from Google reveals that they are the most searched visualization motif (`visualizationuniverse.com`). An online search of bar chart images will issue a wide variety of bar chart representations, going from the simplest to those with all kinds of embellishments, such as shadows, 3D effects, and perspective projection views. This has generated a well-known controversy between visualization specialists and infographics designers. The first, usually lean to minimalistic, data-ink ratio efficient designs, as opposed to color-rich depictions with a variety of decorative elements.

Most of the embellishments used in bar charts tend to increase the error in the perceived quantities [1], [2]. Their defenders claim that they may increase memorability and thus reach a broader audience [3]. However, [4] showed that tick marks present in interactive sliders can introduce bias in survey responses, and found that banded sliders could be used effectively, in terms of speed and accuracy, while maintaining a similar level of response bias than when using undecorated sliders. On the other hand, simple representations have also been analyzed previously, which mainly studied the effects of different bar positions (stacked, adjacent, with distractors ...) on the estimation of length ratios [5], [6]. To the best of our knowledge, there has been no previous analysis on the effect of a moderate use of internal bar encodings (e.g. lines or gradient markers within the bar) in the perception of the absolute values represented by the bars. This work here can be seen as a complement of previous research. Here we evaluate the effects of a moderate use of internal bar encodings, such as the quantized gradients detailed below, to improve the perception of the encoded quantities in bar charts. To do so, we have conducted a study that analyzes the absolute value estimation for bar charts where different forms of internal contrast have been used to reinforce the encoded quantities. We have also analyzed the effect of several elements that have previously appeared in the literature, such as drawing a gridline indicating the maximum value at the top of the chart, bars within boxes at the maximum value (boxed bars), or the negative gridlines in bars proposed by Tufte in his famous book [7], as a way to reduce clutter. Our main contributions are:

- We found improvements in perception accuracy when using quantized gradients and Tufte's internal encodings as opposed to standard solid bars.
- Likewise, we found improvements in perception accuracy when a top gridline or boxed bars were added to the basic chart frame.
- Finally, we produced a set of guidelines to inform the design of bar charts for cases where the goal is to accurately communicate the actual values of the bars.

Through three different experiments, one under laboratory conditions and two more using Amazon Mechanical Turk (sic), we have analyzed the effect of different encodings of bar charts and framing layouts in the estimation of absolute value judgments. As a result, we have found that some designs almost completely eliminate the negative bias that occurs when standard solid bar charts are shown in a basic frame. We have also analyzed other factors such as the relation between accuracy and completion time and the negative bias present in many encodings, where we found that accuracy does not depend on the time spent to do the task.

II. Previous Work

There is a wide amount of research on perception of visualization modalities (see for instance Fuchs *et al.* [8], for a literature review on data glyphs). However, in this article we will mainly concentrate on bar charts.

Since bar charts are such a common visualization technique, they have been the subject of great interest among researchers. Several studies have evaluated different aspects of the most typical representations and different embellishment techniques that have become popular recently. An early and impactful study by Cleveland and McGill [5] concentrated on evaluating the perception of length ratio

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes,creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

between two different bar charts. The independent variables were the *position of bars*, as in adjacent, aligned, stacked, and so on. Their results showed that aligned bars scored significantly better than other strategies. Furthermore, there seemed to be a negative bias when judging the perceived length, especially between percentages 30 and 70. These experiments were later replicated by Heer and Bostock in a crowdsourced experiment using Amazon's Mechanical Turk (AMT) [9]. Their results basically replicate McGill & Cleveland's with the same ranking of accuracy among 5 bar chart configurations. One difference was that they obtained slightly better results in all the cases, perhaps due to the kind of population obtained from the AMT experiment. Further work by Cleveland and McGill sets up a simpler layout to concentrate on the perception of lengths and positions and shows that we are better at estimating positions than lengths [10]. Similar results were independently obtained by Simkin and Hastie [11]. A follow-up study by Talbot *et al.* [6] extends the approach by McGill and Cleveland by analyzing the *effect of distractors* (e.g. other bars appearing in the same chart) on the accuracy of bar ratio measurement. An important result they obtained is that separating bars makes their comparison more difficult.

With the improvement of bandwidth connections and increased affordability, infographics, with all their extra decorations or embellishments, have become quite common. Multiple tools facilitate the creation of charts with a few clicks, such as Datavisu.al (`datavisu.al`), Plot.ly (`plot.ly`), Tableau Public (`public.tableau.com`), Vizydrop.com (`vizydrop.com`) and ZingChart (`www.zingchart.com`) among others. As a consequence, those charts appear in many websites. Some of its advocates argue that embellishments may increase memorability (actually, some visual effects effectively do [3]). They actually may be good to attract attention and entertain, but their utility, from the point of view of quantitative rigor, is questionable. A recent study showed that, for bar charts, most embellishments reduce the accuracy estimating quantities [1], while Zacks et al. [2] had previously shown that some extraneous features, such as 3D volumetric bars, also harm perception.

Even though most of the previous studies on bar charts have concentrated on the aspect of relative ratio comparison, other aspects have also been studied, such as peak detection, or temporal location in time series [12]. We concentrate on absolute value estimation.

For completeness, we also mention other work that analyzed the perception of bar charts from other perspectives. Elzer *et al.* [13] studied the way to sort the bars in order to convey a certain message. Wu *et al.* studied the effect of transition changes in the estimation of bar values for animated visualizations [14]. Correl and Heer [15] and Pandey *et al.* [16] analyzed several ways of manipulating or misguiding through visualization. Wrapped bars is a method that encodes multiple data values through horizontal bars, but grouping sets of them in different columns to take advantage of space. Although they are more accurate than treemaps, the fact that multiple columns are used for the bars, makes inter-column comparison more difficult [17]. Finally, Spence and Lewandowsky analyzed the perception of proportions in charts and pie charts [18].

After analyzing the previous work, we can see that most of the experiments have focused on geometric properties of bar charts, and mostly with ratio comparison, instead of absolute value estimation. Moreover, some experiments show a negative bias, especially notable between percentages 30 and 70 for stack bars, for example. Our objective here is to get further insights on absolute value estimation in bar charts, as well as analyzing some designs where internal contrast is added. We also want to study whether a negative bias is present in the analyzed designs.

III. Materials and Methods

This section describes the designed user study: the research hypotheses and our main goals, the setup of the three experiments conducted and the methodology followed in their execution.

A. Goals and Hypotheses

The main goal of the proposed study is to analyze the effect of internal contrast enhancements in the estimation of absolute values in bar charts. When designing bar charts, several elements such as gridlines are part of the design, and their effect may be influential in the judgment of the bars' values. For this reason, we also considered studying the effects of such elements on the perceived values. Since we want to keep bar charts as simple as possible, we restricted our study to the simplest designs found in the literature and in the web (*e.g.*, by searching for "bar charts" in Google Images), together with some internal contrast enhanced designs that we found either potentially useful, intuitive, or that have been proposed previously in the literature. We ended up with two different aspects of bar chart design that were studied independently:

- The use of different framing elements, such as a top gridline, or a box around the chart (as found in Cleveland & McGill [5] for unaligned bars).
- The use of different internal contrast encodings for bar charts, such as the negative gridlines in Tufte's book [7].

The goals of the study made us wonder whether internal contrast-enhanced encodings are better at communicating quantities and whether the top gridline helps better estimate chart values. These questions lead us to formulate the following two hypotheses:

- **H1:** Internal contrast-enhanced encodings communicate quantities better than standard solid bars.
- **H2:** The line at the top of the chart improves the estimation of the bars' values.

Although we do not consider unaligned bars as in [5], we are interested in finding out whether having the box around each bar helps to better perceive the values, compared to having the top gridline bar. Our hypothesis is that these two framing methods help users estimate values more accurately and thus, both cases will give similar average error. Besides, we hypothesize that both Top gridlines and Boxed bars will result in similar estimation ratios, because the distance to the top horizontal line and the top side of the box would be the same for the encodings presented in the study. Previous experiments [5], under some configurations, have shown that values estimation suffer from a negative bias especially in the range of 30-70%. We also are interested to see if any of the chart designs used in this study avoids the negative bias when determining the values of the bars. These two assumptions (line framing vs. box framing and the biased estimation of values) led us to the following additional hypotheses:

- **H3:** Boxed bars have the same impact as top gridlines as visual aids when estimating the absolute values of the bars.

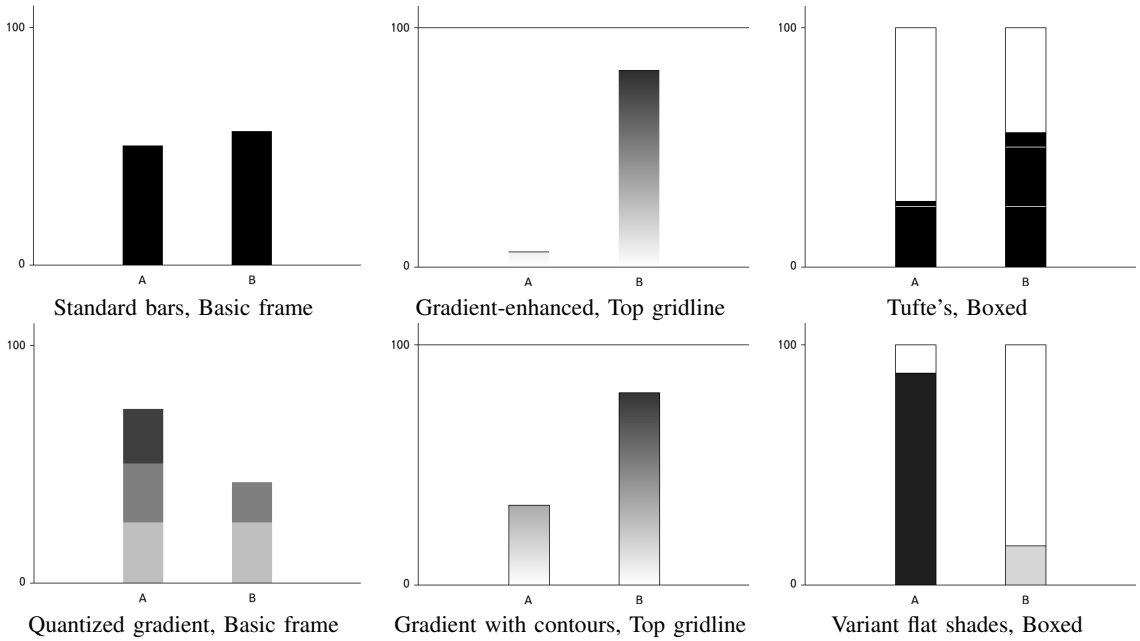


Figure 1. Different combinations of the six bar types and chart framing styles: basic (left), top gridline (center) and boxed (right) were used in our three experiments.

- **H4:** Internal contrast-enhanced encodings prevent the negative bias present when estimating values.

In order to prove the hypotheses stated above, we designed a set of experiments with the objective to analyze the effect of these elements. The data space is very big. The combination of different framings with the number of internal contrast-enhanced designs we considered made the approach unfeasible for a single experiment. Thus, we followed a top down approach with an initial (pilot) experiment in laboratory conditions that helped us discard some of the encodings and analyze the effect of two framing styles. Then, two subsequent crowdsourced experiments using Amazon Mechanical Turk (AMT) helped us analyze in detail the effect of the individual encodings.

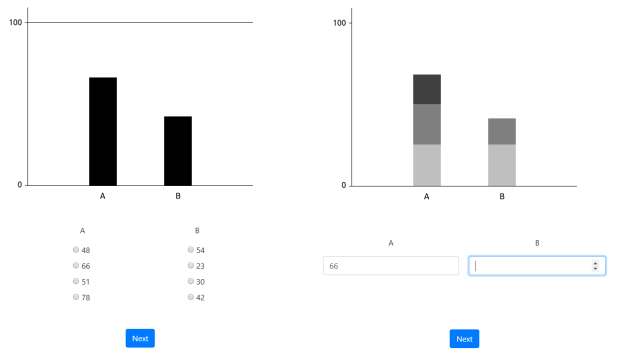
B. Chart Designs

The experiments conducted in the proposed study are based on the task of determining the values represented in bar charts with different styles and framings: basic frame (with a tick line on the Y-axis indicating the top value of 100), basic frame with a gridline at the top value and basic frame with bars boxed within the maximum range (see Figure 1). The styles used to draw the charts were:

- *Standard bar charts*: bars are filled with a solid color. This style is one of the most commonly employed when visualizing bar charts.
- *Gradient-enhanced bar charts*: the values of the bars are represented by their opacity with a color gradient from the base (0 – white) to the value at the top (eventually 100 – black). Since no contours are used in this design, bars representing small values may be almost transparent. To make the top always visible, a black line is drawn at the top of the bar.
- *Gradient with contours bar charts*: this design is based on the same idea than the previous one but the contours are displayed. Filling the bars with gradients is a commonly used style when visualizing bar charts.
- *Quantized gradient bar charts*: this style is a quantized version of the gradient-enhanced one. Every quarter of the bar, the opacity (*i.e.*, color) changes producing the effect of having inner borders in the bars. A maximum of four different shades are shown for each bar filled with an opacity of 25%, 50%, 75% and 100% respectively.
- *Tufte's encoding of bar charts*: bars are filled with a solid color with a contrasting inner line every 25%. This bar chart style is described by Tufte in [7] and it is referred to in the rest of the paper as Tufte's encoding.
- *Variant flat shades bar charts*: this design fills each bar with a flat shade at a darkness level that varies in correspondance to the value of the bar. Contours are also displayed to ensure visibility at lower values.

In our study, standard bar charts are used as baseline. The other styles are intended to provide additional visual cues (such as codifying the value in the opacity or through shading techniques) to help users better estimate quantities, but without adding external features such as gridlines, numerical values, and other aides. In all cases, we provide extra information inside the same bar, with proposals that have appeared in the literature (such as Tufte's encoding), or some that may be not frequent in visualization software but which seem helpful.

As opposed to other experiments, where the estimated quantity is the ratio length of one bar with respect to the other, we are evaluating absolute value estimation. Taking into account the work of Talbot *et al.* [6], the charts have been designed with two *separated* bars, so that each one is evaluated individually. In order



a) Initial training task b) Main task of the experiment

Figure 2. Experiment setup tasks. The experiments of the study consist of a training task (a), where the participants have to choose the correct values of the bars among a set of options, and the main task (b), where users have to introduce the values directly.

to avoid the potential effect of the distance from the bar to the reference scale axis, the left bar is placed at a reasonable distance of it (and will subsequently analyze whether estimating its value differs from the right one). Moreover, in the presence of the top gridline and the boxed version of the bars, the chart's frame has an horizontal line indicating value 100 in one case, or the bars are surrounded by a box, following Cleveland & McGill's approach [5] in the other. This makes the estimation to be in the vertical direction instead of horizontal. Charts have been generated at a resolution of 640x480 pixels (Heer and Bostock used a resolution of 380x380 pixels in [9]) to be able to properly see the whole chart at a glance without moving the head when the subject is at an approximate distance of 60 cm from the display [19], [20]. However, note that the conditions of Amazon Mechanical Turk make it difficult, if not impossible, to ensure that charts are visualized at this size by the participants. On the other hand, and if the screen sizes declared by them are to be trusted, it seems that most users were in similar conditions to those described. In any case, we assessed the validity of the proposed charts' design through our experiments.

C. Experimental Setup and Procedure

The differences between the setup of the three experiments of the study were the type of charts and framings shown in each experiment, the number of charts used, and how the users were selected (the pilot experiment was conducted in our lab in a controlled environment, while the other two experiments were performed via Amazon Mechanical Turk). All the experiments follow the same structure: a main task where the users have to guess the values of different bar charts, preceded by an initial training task. In training, the users have to select the answer among four predefined values for each column (Figure 2 a). The aim of this stage is to determine if the users understand the task to perform and discard possible outliers. In these tasks, charts are shown as is, without explanations about the encodings or framings, and users are asked to select the answers which reflect the values encoded in bars A and B, respectively. From the four values to choose, one is the correct answer, another one is within the range of the bar's exact value ± 10, and the other two are outside this range. If the user selects more than two answers outside the ± 10 range, the test fails, and the user is not invited to continue to the next part of the experiment. Users that pass the training task then start the main task, where instead of selecting one of the possible choices, they must introduce the estimated value of the bars in whole numbers (Figure 2 b). Charts in this task are presented in random order to avoid the learning effect and one control chart is shown for each chart type displayed (i.e., one of the charts per type is presented twice). These control charts are used to discard careless users. When the difference of the answers between the first and second time the control chart is shown exceeds a certain threshold, the user's data is discarded. To avoid visual fatigue, the main task is divided into six segments, where users can take a break at the end of each segment. The timings used to perform the statistical analysis do not take into account the time spent by the participants during the breaks. The values introduced by the user and the time spent to complete each chart (two values, one for each bar) are recorded to perform the statistical analysis.

The pilot experiment was run by means of a desktop application, whereas the two other experiments were performed via a web app (executed through AMT), which consisted of 5 sections:

- *Generic information and rules*: a brief explanation of the experiment and general rules to follow were provided (for instance, users were advised to not user rulers or other external tools to complete the tasks).
- *Demographic information survey*: participants were asked to fill a form with personal information (age, gender, education level, quality of eyesight and their display size in inches).
- *Training task*: instructions to accomplish the task were provided and then the training task was administered.
- *Main task*: instructions to accomplish the main task were provided and then the main task was administered.
- *Personal evaluation survey*: users were asked to answer different questions related to the understanding of the task to perform, its difficulty, and others. Additionally, they could introduce other comments about the experiment.

Pilot experiment. The first experiment served us to select the most promising chart types presented in Section III-B and to evaluate the effect of two framings styles: top gridline and boxed bars. Thus, the six internal contrast-based bar styles were used in combination with the two framings, producing a total of 12 different bar chart configurations. For each of them, 5 different charts were displayed and one was shown twice as a control chart. This amounts a total of 72 charts, with 2 bars per chart, which makes 144 answers per user. The values of the bars were generated randomly in the range [3, 97], to prevent overly simple judgments where the bars might be completely aligned with the axis, as suggested in [1]. This experiment was performed in a controlled environment with displays of 21-24 inches in good lighting conditions. Since we had access to the participants, the information and the rules of the study, as well as their demographic information and personal evaluation were collected on site.

Basic frame vs. top gridline experiment. In the second experiment, our goals were to assess whether the framing using the top gridline improved perception compared to using the basic framing, and whether the best internal contrast-enhanced method had less negative bias than the standard solid bars. To do this, we chose

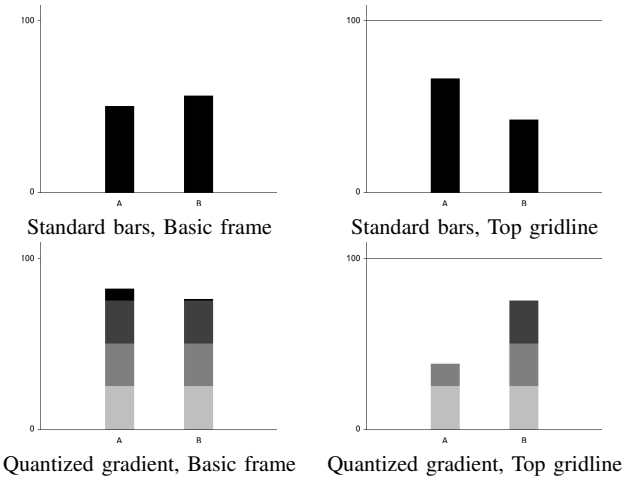


Figure 3. Chart types used in the second experiment, where the effect of the top gridline and the accuracy of the quantized gradient encoding were evaluated.

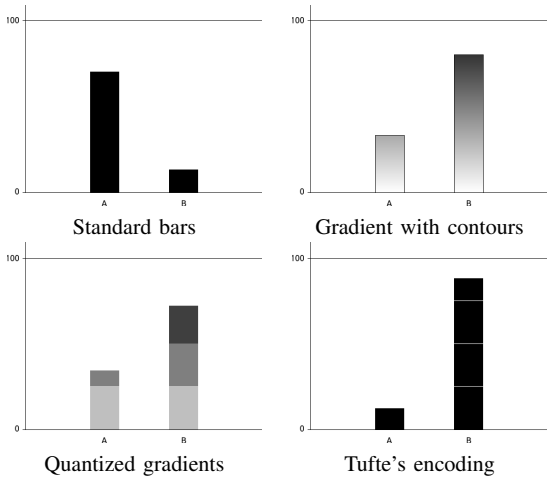


Figure 4. Chart types displayed in the third experiment, where the main goal is to determine if some internal contrast-enhanced styles support a more accurate estimation of the values than the standard style with solid color bars.

the internal contrast-enhanced type that provided the best results (see Section IV) in the pilot experiment: the quantized gradient bar style and compared it against the standard solid bars. Thus, the four different chart encodings shown in Figure 3 were used. For each configuration, we generated 14 charts plus one control chart that repeated one of them. This makes 60 charts with a total of 120 responses (2 bars for each chart). The values of the bars ranged from 3 to 97 and the distribution of the values was balanced among the whole [3,97] range: we divided the range in equal parts, and the values randomly generated for the two bars were distributed along each portion of the range until all the ranges had a similar amount of values. Furthermore, the cases where the left bar was higher than the second were also balanced, to prevent giving any advantage to any kind of judgment. This experiment was performed using Amazon Mechanical Turk.

Internal contrast enhancements experiment. The goal of the third experiment was to evaluate the degree of accuracy of the internal contrast-enhanced chart types against the standard bar design. The methods chosen for this experiment along with the standard design were: the quantized gradient and Tufte's encoding (users performed significantly better with them in the pilot experiment, as described in Section IV), together with the smooth gradient with contour style, for gradient is sometimes used in bar charts visualization, albeit not typically to encode a value. The Gradient-enhanced and the Variant flat shades styles were discarded because no significant differences in user's accuracy were found in the pilot experiment. Regarding the frame configuration, the use of the gridline on top was chosen for the four chart types to provide a level playing field, while taking into account that users' performed significantly better with it in the previous experiment. Examples of the four designs of charts displayed in this experiment are shown in Figure 4. In the same way as it was done in the second experiment, 14 different charts plus one control chart were displayed for each design, with bars' values randomly generated between 3 and 97, evenly distributed among the whole range and balanced with respect to the values of both bars. This produced 60 charts and 120 responses per user. This last experiment also consisted of a crowdsourced test using Amazon's Mechanical Turk.

D. Participants

We decided to conduct the first (pilot) test in a controlled environment, taking advantage of the proximity to obtain a clear idea of the amount of time the study actually consumed. However, to obtain ecological validity in the subsequent experiments, we used Amazon Mechanical Turk (AMT), as mentioned before. AMT has proven a good research tool, robust enough for perception experiments [9]. On the positive side, it provides an easy and affordable way to perform crowdsourced studies relatively quickly. On the other side, AMT is not free from the risk that users might answer carelessly. To avoid this, it is necessary to introduce controls inside the experiment to eliminate negligent users, as we have done. Moreover, when the experiments are separated in individual tasks (called HITs in Amazon's terminology, from Human Intelligence Tasks), for example, one task per answer, users may leave parts of the experiment uncompleted. As a result, in order to evaluate a high number of different conditions, properly randomized, and obtain enough answers, we need to either break down the study into several parts, or perform relatively long within subject experiments until all the configurations have enough answers. We chose the latter. However, to avoid frustrating users with a very long battery of questions, the main task was divided in six segments with breaks, to allow users to rest as much time as needed to resume the experiment focused. In the following paragraphs, we present the information relative to the users that took part in our study.

Pilot experiment. It is known that for low level perceptual studies, under controlled conditions, a low number of participants, such as 10, may be sufficient [21]. Since the pilot study was carried out in controlled conditions, the variance among the results was likely to be lower than with other crowdsourced tests [9]. For the experiments on AMT, we selected more participants. A total of 12 users (9 male, 3 female) participated in the pilot experiment, ranging from 14 to 45 years. All of them had an excellent or good

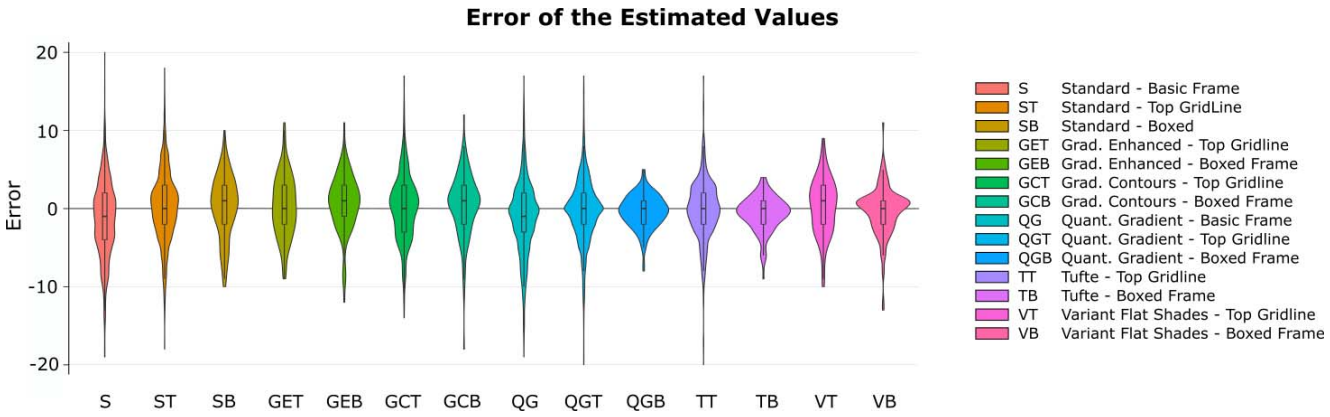


Figure 5. Estimation error commited in all the answers given in the study for each bar chart configuration. The empirical probability distribution of the estimation error is also shown in this violin plot.

eyesight quality and they understood the task to do. Each user was shown 5 versions of each chart plus a control one, so we got 72 charts, totalling 144 answers per user (each chart had two bars to estimate). After analyzing the data and the control charts, no user had to be discarded.

Basic frame vs. top gridline experiment. 58 *turkers* (*i.e.*, workers of Amazon Mechanical Turk) began the experiment and 50 finished it. 8 of them did not pass the training stage for poor performance or abandoned after completing the first segment of the main task. Among the 50 that completed the experiment, six were discarded: one of them introduced random answers and the other five failed in the control charts validation. These users were replaced with other six participants that provided consistent answers. In total, we had 50 valid participants (35 male, 15 female), with ages between 22 and 60, all of them claimed to have an excellent or good eyesight and all of them assured that they understood the task to perform. Each user was shown 60 charts, totalling 120 answers per user.

Internal contrast enhancements experiment. 59 turkers began the experiment and 50 completed it. In this case, none of the participants provided random answers but three of them did not pass the validation process with the control charts. These users where replaced by other 3 that passed the validation, totaling 50 valid users (35 male, 15 female) from 21 to 68 years. All of them reported that their eyesight was excellent or good and that they understood perfectly the task to do. As happened in the previous experiment, each user had to determine the values of 60 bar charts, providing a total of 120 answers per participant.

IV. Results

In this section we present the main results of our study. The empirical probability distribution of the estimation error committed in all the answers of the experiments conducted is shown in Figure 5. The data reveals normal and non-skewed distributions of the estimation error for all the different bar charts types considered in our work.

A. Data Analysis

In order to analyze the accuracy of the given answers, the absolute error of the estimated values was analyzed for each experiment by using a one-way analysis of variance (ANOVA) with a significance level of $\alpha = 0.05$. When significant differences between the means were found, we used a post-hoc Bonferroni's pairwise test with the same significance level ($\alpha = 0.05$). To determine if there was a negative bias in the answers provided by the participants, a hypothesis test for a proportion with significance level $\alpha = 0.05$ was used. Finally, to test linear correlation between the accuracy of the estimated values and the time spent in such a task, we used the Pearson's r statistic and assessed the linear model testing the regression coefficient β_1 with $\alpha = 0.05$.

B. Pilot Experiment

The result of the ANOVA test ($p < 0.0001$) led us to reject the null hypothesis that the means of the absolute errors were equal between the different chart types shown in this experiment (the 6 bar styles presented in Section III-B with 2 different framings: top gridline and boxed). Bonferroni's test revealed that in the majority of the cases, the answers' accuracy was significantly better when showing the quantized gradient (under both framings) and the boxed framing of Tufte's encoding. Concretely, these three designs presented significantly better results compared to the Standard bars, the Gradient-enhanced bars, the Gradient with contours bars (regardless of framing style) and the Variant flat shades bars (using top gridline framing). No other significant differences were found in the rest of the cases. The 95% confidence intervals for the absolute estimation error shown in Figure 6 (left) point towards the same conclusion. These results led us to choose the Quantized gradient and Tufte's encodings for further analysis in subsequent experiments.

The accuracy of the answers as a function of the framing style (Top gridline or Boxed) was compared using the same statistical analysis. In this case, the ANOVA test ($p = 0.0149$) revealed a significant difference between the means of their absolute errors ($\overline{Box} = 2.41$, $\overline{TopGridline} = 2.73$). As a consequence, we can conclude that users determined the values of the bars in a significantly more accurately fashion when the box was present, although by a very small margin of 0.32%.

Two main results were obtained through the pilot study:

- Boxed bars are more effective in reducing estimation error than charts with a top gridline to indicate maximum value for

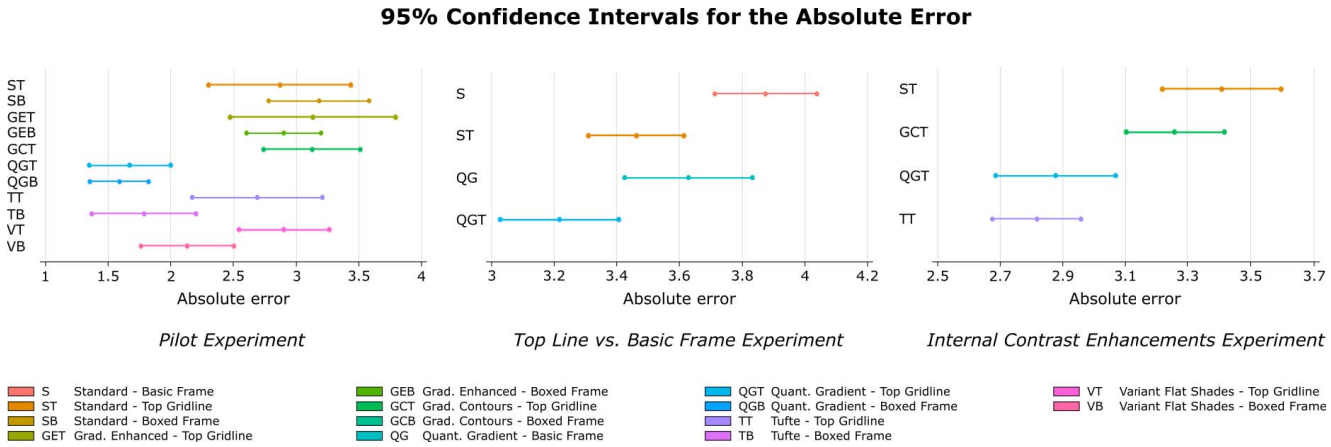


Figure 6. 95% confidence intervals for the absolute error committed when estimating the values of the bars shown in the *Pilot Experiment* (left), the *Top Gridline vs. Basic Frame Experiment* (center) and the internal *Contrast Enhancement Experiment* (right).

charts with a small number of bars.

- Quantized gradients and Tufte's internal contrast encodings improve significantly values estimation accuracy.

Although the experiment showed interesting, significant results, we wanted to confirm some of its findings and address open questions through the subsequent experiments.

C. Top Gridline vs. Basic Frame Experiment

The statistical analysis revealed differences (ANOVA: $p < 0.0001$) between the means of the absolute errors obtained with the four configurations compared in this experiment (see Figure 3). Concretely, users provided significantly more accurate answers when the quantized gradient style with the top gridline ($\overline{QGT} = 3.21$) was shown compared to quantized gradient bars with the basic frame ($\overline{QG} = 3.62$). Besides, standard bars with the top gridline framing ($\overline{ST} = 3.46$) presented a significantly more accurate estimate of the values than standard bars with the basic frame ($\overline{SB} = 3.87$). Figure 6 (center) shows the 95% confidence intervals for the absolute estimation error in this experiment. The analysis of the absolute error (ANOVA: $p < 0.0001$) confronting all the answers obtained in the presence of the top gridline ($\overline{TopGridline} = 3.33$) and in its absence ($\overline{NoTopGridline} = 3.72$) confirms the previous result: users provide more accurate answers when the top gridline is used in the charts.

In order to analyze if there was a negative bias in the answers provided by the users, a hypothesis test for a proportion with $\alpha = 0.05$ was used. The null hypothesis considered that the proportion of answers underestimating the value of the bars (negative error) was equal or lesser than the 50%. The results of the test only revealed a significant negative bias with the standard bars with simple framing ($p < 0.001$). To check if the estimated values showed a negative bias between the percentage 30 and 70, as it appears under some configurations in [5], we just considered the answers given by the users in the bars whose values lied in the range $[30, 70]$ (see Figure 7). The results of the test confirmed a negative bias in the absence of the top gridline (standard bars: $p < 0.001$, quantized gradient bars: $p < 0.001$). Instead, the null hypothesis could not be rejected when the top gridline was present (standard bars: $p = 0.5$, quantized gradient bars: $p = 0.583$).

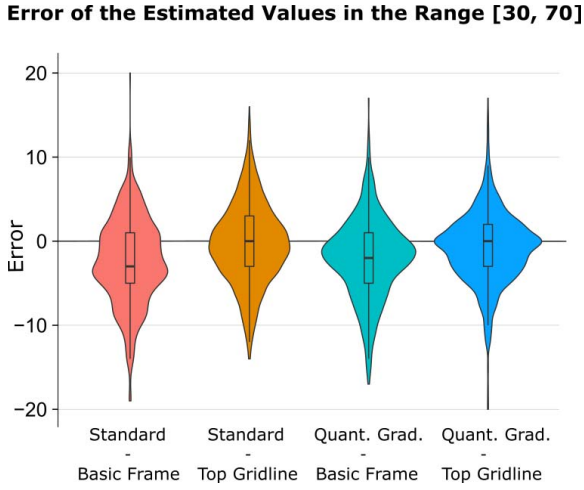


Figure 7. Error present in the Basic frame vs. Top Gridline experiment while estimating the values of the bars. The statistical analysis shows that there is a significant negative bias in the answers provided when using the basic frame, which disappears when using the top gridline framing.

Thus, there is no significant evidence to claim that a negative bias is present in the answers of the users when using the top gridline, but with the two configurations with the basic frame tested in this experiment, a significant negative bias is present.

The main result obtained through this experiment was:

- Showing a gridline on top of the graph at the maximum value improves the perception of the absolute value of bar charts.
- Similarly, the quantized gradients encoding is significantly better than solid bars, regardless of framing.

D. Internal Contrast Enhancements Experiment

We wanted to obtain further evidence that the internal contrast enhanced encodings selected after the pilot experiment could produce statistically significant improvements in accuracy. The goal of the third experiment was to obtain a more definitive answer to this question. The statistical analysis (ANOVA: $p < 0.0001$) revealed

a significant difference between the means of the absolute errors provided with the different chart designs displayed (see Figure 4). Bonferroni's test determined that the accuracy obtained with the quantized gradient charts ($\overline{QGT} = 2.87$) was significantly higher than the accuracy of answers with the gradient with contours ($\overline{GCT} = 3.25$) and the standard bar designs ($\overline{ST} = 3.40$). Moreover, users' performance with Tufte's encodings ($\overline{TT} = 2.81$) was significantly better than these two chart types. No significant differences were found between the means of the absolute errors obtained with the Quantized gradient style and Tufte's nor between standard bars and the gradients with contour (see Figure 6(right)). The conclusion is that quantized gradients and Tufte's encoding, both internal contrast enhanced configurations, provide significantly better estimations of the values than the standard solid bars. Our hypothesis is that the inner borders present in the quantized gradients and Tufte's style may make the estimation of values easier, compared to the standard solid bars encoding. Instead, although the gradient with contours configuration provide more accurate estimations than the standard bars, there are no significant differences between them.

The results of the third experiment are clear:

- Quantized gradients internal contrast is significantly better than solid bars and smooth gradients.
- Tufte's encoding is significantly better than solid bars and smooth gradients.
- There is no statistically significant difference between Quantized gradients and Tufte's style, nor between any other pair-wise combinations. We believe this is because both encodings effectively use internal contrast to display inner borders at every quarter of the range.

E. Further Analysis

Accuracy vs. completion time. As stated before (see Section III), the time to complete each chart was recorded to check if there was a linear correlation between the accuracy of the given answers and the time spent to provide them. Since participants have to estimate the value of two bars for each chart and we do not have information related to the time spent to determine a single value, we define the *average accuracy* of a given chart as the average of the absolute errors that occurred when estimating the values of the two bars. In this way, we can test for a linear correlation between the average accuracy and the time spent for each single chart. The time employed by the users to complete the pilot experiment was ca. 30 minutes (they spent an average of 25 seconds per chart). In this case, the statistical analysis by means of the Pearson's r statistic ($\alpha = 0.05$) does not reveal a linear correlation between the average accuracy and the time spent in each chart ($r = -0.007$, $p = 0.818$). In the second and third experiments, the average elapsed time to complete them was 37 and 34 minutes respectively (an average of 37 and 34 seconds per chart). As it happened during the pilot test, no linear correlation between the average accuracy and the time employed was found (experiment 2: $r = -0.025$, $p = 0.164$; experiment 3: $r = -0.004$, $p = 0.825$). Similarly, no such linear correlation was revealed when analyzing the average accuracy and the time spent for each chart design individually.

Accuracy of left vs. right bar estimates. We analyzed whether our data showed differences between the accuracy obtained estimating the magnitude of the left bar compared to that of the right bar, and found no support for this. We believe that this is due to the fact that the bar on the left was not as close to the vertical axis as to make it significantly easier to quantify compared to the bar on the right. In addition, conditions with enhanced framings would tend to reduce the effect of the proximity to the vertical axis, since the distance to the top gridline and the boxed bars top side is the same for both bars.

V. Discussion

In this section we discuss our results in light of the hypotheses that were introduced in Section III-A and finish with a set of guidelines for bar chart design.

A. Internal Contrast Enhanced Encodings

The first hypothesis, H1, proposes that internal contrast-enhanced encodings communicate quantities better that the standard solid bars. In this regard, we found that some internal contrast enhancements improve significantly the estimation of values. In particular, we found evidence that quantized gradients and Tufte's methods both significantly improve users' accuracy. The pilot experiment provided the initial evidence in support of quantized gradients and Tufte. In addition, we found that enclosing bar charts in a bounding box does provide support for users to increase estimation accuracy. Other internal contrast enhanced encodings, such a smooth gradient, with or without framing enhancements, do not contribute significantly to accuracy, whereas the main contribution of quantized gradients and Tufte's encodings is that users can take advantage of the inherent information encoded in these representations. Although the quantized gradients design may resemble stacked bars, it is worth noting that users received no instructions whatsowever about the meaning of the encodings, which is a strong indicator that users are able to discern whether the bars represent one quantity of multiple quantities. In addition, quantized gradients are presented as shades of the same color, while stacked bars are usually represented in different colors to emphasize the distinct values being encoded.

B. Basic Frame vs. Top Gridline

Hypothesis H2 states: "The line at the top of the chart improves the estimation of the bars' values". We found plenty of evidence to support this hypothesis through the second experiment, where we observed that, when considering all graph types with top gridlines versus all graph types with the basic frame, there was a statistically significant improvement in accuracy when the top gridline was present. Then, in pair-wise comparisons between individual chart types, we found that all representations with gridlines at the top performed significantly better that the graphs which used just the basic frame. The clearest result in this sense was that quantized gradients with top gridlines increase accuracy when compared to quantized gradients with the basic frame only. The same effect is observed when standard solid bars with top gridlines are compared against standard solid bars with the basic frame. That is, when comparing graphs of the same type, the presence of a top gridline significantly improves results.

C. Boxes vs. Top Gridlines

Hypothesis 3 (H3), proposes that Boxed bars have the same impact as Top gridlines as visual aids when estimating the absolute values of the bars. The pilot study addressed this question. There, half of the graphs had boxes, and the other half had the top gridline. We found that representations with boxes significantly improved the accuracy of the estimated values (see Section IV). From the three chart conditions that exhibited significantly lower error, two of them (Quantized gradients and Tufte) were with boxed bars, whereas only one (Quantized gradients) was not. Another evidence of the advantage of using boxes over the top gridline is that from 13 pair-wise comparisons that yielded statistically significant differences between a boxed against a top gridline representation, in 10 cases the representation with boxes produced better accuracy than the representation with a top gridline. The 3 exceptions were cases where Quantized gradients with top gridlines were better than other representations with boxes (smooth gradients, smooth gradients enhanced with top gridlines and solid bars). This suggests that, in the absence of quantized gradients, boxed framings tend to produce more consistent benefits that top gridline framing, and that the benefit obtained by using quantized gradients is greater than the benefit of using boxes instead of top gridlines.

D. Addressing Negative Bias

The fourth hypothesis (H4) states that internal contrast-enhanced encodings avoid the negative bias (*i.e.*, the bias where people tends to subestimate the value of the bars). We observed this negative bias as a subestimation error of -1.35% with respect to the actual average when people were presented with standard solid bars in a basic frame with no top gridline or other aids discussed above. Across several experiments, we found that the condition presenting quantized gradient bars, either with a gridline at the top, or boxed, is the most reliable way to reduce bias. We have also obtained interesting results from the second experiment: first, we found that in general standard solid bars with a basic frame exhibit a negative bias. In the range [30, 70] some chart types (solid and quantized gradients) used with the basic frame also show a negative bias. However, when these charts have the top gridline added, the bias effect is eliminated. This would suggest that having the gridline at the top would be the best and simplest strategy to use to remove the bias. However, in Tufte's design with a top gridline, a negative bias is still observed (at a significance level of 5%). This is a weak effect of 0.8%, but the fact is that it is still statistically significant. Even though the experiment was not designed to study bias in detail, the question of addressing bias could be a subject of more detailed research in the future.

E. Variant Flat Shading and Continuous Gradient

Internal contrast in the whole bar by encoding the depicted quantity as an opacity (variant flat shading) showed no improvement over solid bars. We also analyzed the effect of continuous gradient encoding over solid bars, where the estimation error seemed to diminish slightly, but found no statistically significant differences with respect to solid bars. While results might have changed if users had been given prior explanations or instructions about the meaning of the encodings, it is worth noting that neither Quantized gradients nor Tufte's method came with explanations, and nevertheless resulted in consistent improvements in perception accuracy.

F. MTurk Experiments

Due to the anonymous procedure used to recruit participants in the AMT, it is not possible to characterize the AMT population, thus, we cannot speculate much about the similarities or differences between the AMT population and the laboratory participants. While the experiment run in laboratory conditions went smoothly, for the MTurk experiments we had some issues. The most important one was time: users devoted more time than expected to the answers. We had counted for 20 minutes approximately, and the average was 34 minutes. The users took more than the 5-7 advised seconds to answer each chart. We had calculated the timing based on the first laboratory test and previous examples in literature. For further studies, we will adjust our advices accordingly. Some users responded either randomly or carelessly. This is expected, and that is why we added the controls to avoid their results to be included.

VI. Conclusions and Guidelines

We performed a set of perception studies to evaluate the accuracy in absolute value estimation for bar charts. The objective was twofold: first, get more insights on how accurately people perceive absolute values in bar charts. Second, finding ways to improve the estimation of values in bar charts using internal contrast encodings and minimalistic framing enhancements. The perceptual experiments showed many interesting results. For instance, that using a gridline at the top of the chart or enclosing the bars in boxes effectively helps improve the estimation of values. Another interesting result is that we can improve the estimation of values either with the visualization method proposed by Tufte [7] or with the proposed Quantized gradient style. Finally, we also found that the use of a top gridline helps reduce negative bias that occurs in bar charts without adequate internal contrast encodings.

A. Guidelines

Here we present the contributions of this study as guidelines for the design of bar charts, with the goal to provide the best support to communicate the values encoded accurately and without bias, while maintaining support to do visual comparisons between the bars.

- It is better to add a gridline at the top value in the chart. Most previous experiments in literature addressed the estimation of ratios between bars. However, in many cases the observer needs to estimate not only the relative size of the bars, but the absolute value represented by a bar. In such context, it is better to add a top gridline to improve accuracy.
- Tufte's encoding is significantly better than simple solid bars and smooth gradients. This was expected, because the internal gridlines provide additional reference points. Despite being a simple technique to use that adds no noise, it not commonly used. We would encourage designers to use it, keeping in mind that some negative bias may occur when used with a line on top.
- Quantized gradient encoding is significantly better than simple solid bars and smooth gradients as an aid to improve perception. This result was also expected because this encoding provides additional information that users can interpret intuitively.

The benefits of a quantized encoding are more significant than the effect of using boxes for encodings such as solid bars or smooth gradients. Quantized gradients, used with a gridline on top, or a box framing, consistently produces the least bias. For these reasons, we would encourage designers to use this encoding in addition to Tufte's.

- In case the quantized gradient encoding is not used, for charts with a small number of bars, such as those presented in this study, using boxes instead of top gridlines improves the estimation of the encoded values.

With regards to the last guideline, we have not studied the effect of using boxes in dense bar charts where visual clutter may become an issue, so it was not possible to advise on it.

In the future, we want to gain more insight on the perception of bar charts and other visualization modalities. For instance, the effect of boxes in dense bar charts has not been analyzed. We would like to see whether they are still beneficial for such designs, or they start to act as a distractor. All the tests we have designed follow the most common examples in literature, with no additional gridlines and only a tick indicating 100% at the top. We would like to analyze the use of a moderate amount of ticks, and see how they compare to Tufte's internal gridlines. Finally, we would also like to analyze the effect of color. Throughout our experiments, we found that encoding the quantity as the opacity or darkness of the bar did not improve the value judgment. However, this was carried out with monochromatic gradients, modifying the color, in addition to the opacity, may have different effects.

Acknowledgements

The authors want to thank O. Argudo and J.L. Díaz-Barrero for their valuable contributions. As well, we thank the volunteers who took part in the study. This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness and by the FEDER (EU) funds under the Grant No. TIN2017-88515-C2-1-R.

References

[1] 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, 2015.

[2] J. Zacks, E. Levy, B. Tversky, and D. J. Schiano, "Reading bar graphs: Effects of extraneous depth cues and graphical context." *Journal of experimental psychology: Applied*, vol. 4, no. 2, p. 119, 1998.

[3] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 519–528, 2016.

[4] J. Matejka, M. Glueck, T. Grossman, and G. Fitzmaurice, "The effect of visual appearance on the performance of continuous sliders and visual analogue scales," in *CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5421–5432.

[5] 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, 1984.

[6] 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, 2014.

[7] E. R. Tufte, *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press, 1986.

[8] J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim, "A systematic review of experimental studies on data glyphs," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 7, pp. 1863–1879, 2017.

[9] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 203–212.

[10] W. S. Cleveland and R. McGill, "An experiment in graphical perception," *International Journal of Man-Machine Studies*, vol. 25, no. 5, pp. 491–500, 1986.

[11] D. Simkin and R. Hastie, "An information-processing analysis of graph perception," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 454–465, 1987.

[12] 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 *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3237–3246.

[13] S. Elzer, N. Green, S. Carberry, and J. Hoffman, "A model of perceptual task effort for bar charts and its role in recognizing intention," *User Modeling and User-Adapted Interaction*, vol. 16, no. 1, pp. 1–30, 2006.

[14] E. Wu, L. Jiang, L. Xu, and A. Nandi, "Graphical perception in animated bar charts," *arXiv preprint arXiv:1604.00080*, 2016.

[15] M. Correll and J. Heer, "Black hat visualization," in *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, 2017.

[16] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques," in *ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1469–1478.

[17] M. A. Yalçın, N. Elmqvist, and B. B. Bederson, "Raising the bars: Evaluating treemaps vs. wrapped bars for dense visualization of sorted numeric data," in *Proceedings of the 43rd Graphics Interface Conference*, 2017, pp. 41–49.

[18] I. Spence and S. Lewandowsky, "Displaying proportions and percentages," *Applied Cognitive Psychology*, vol. 5, no. 1, pp. 61–77, 1991.

[19] P. Bodrogi and T. Khan, *Illumination, color and imaging: evaluation and optimization of visual displays*. John Wiley & Sons, 2012.

[20] G. W. Humphreys, *Attention, Perception and Action: Selected Works of Glyn Humphreys*. Routledge, 2016.

[21] D. W. Cunningham and C. Wallraven, *Experimental design: From user studies to psychophysics*. CRC Press, 2011.