Neighborhood Perception in Bar Charts

Mingqian Zhao
Hong Kong University of Science and Technology
mzhaoad@connect.ust.hk

Huamin Qu
Hong Kong University of Science and Technology
huamin@cse.ust.hk

Michael Sedlmair
University of Stuttgart
Michael.Sedlmair@visus.uni-stuttgart.de

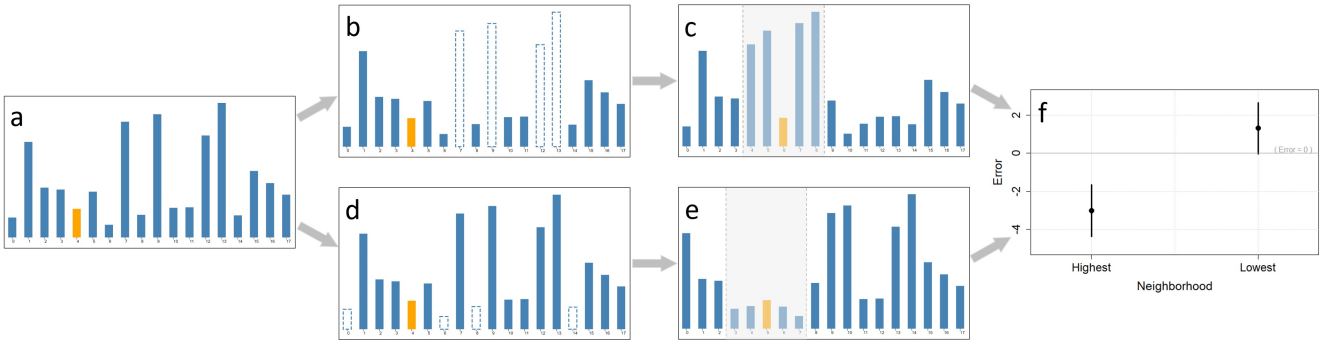

Figure 1: Neighborhood effects: the difference in perceived target bar rank between the highest and the lowest neighborhoods (from one dataset in study 1). The baseline chart before neighborhood manipulations is shown in (a). The target bar at 25th percentile is either surrounded by the top four highest bars in (c) or the top four lowest bars in (e). A substantial difference in estimation error is found in (f).

ABSTRACT

In this paper, we report three user experiments that investigate in how far the perception of a bar in a bar chart changes based on the height of its neighboring bars. We hypothesized that the perception of the very same bar, for instance, might differ when it is surrounded by the top highest vs. the top lowest bars. Our results show that such neighborhood effects exist: a target bar surrounded by high neighbor bars, is perceived to be lower as the same bar surrounded with low neighbors. Yet, the effect size of this neighborhood effect is small compared to other data-inherent effects: the judgment accuracy largely depends on the target bar rank, number of data items, and other data characteristics of the dataset. Based on the findings, we discuss design implications for perceptually optimizing bar charts.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300462

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in visualization*;

KEYWORDS

Perception, Neighborhood, Order, Bar Chart, MTurk Study.

ACM Reference Format:
Mingqian Zhao, Huamin Qu, and Michael Sedlmair. 2019. Neighborhood Perception in Bar Charts. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300462

1 INTRODUCTION

Bar charts are arguably on the most ubiquitous visualization techniques. Their usage ranges from casual users that seek to understand their personal data, to professional decision makers making delicate choices on such data representations.

As a natural consequence, many researchers have focused on gaining a detailed understanding of how bar charts and other common statistical graphics are perceived and interpreted by users. This line of research has unveiled many interesting and surprising facets. Previous work has, for instance, shown how contextual factors, such as biased social signals [17], priming and anchoring effects [14, 40], framing effects [22], decorations [26], luminance [37], plotting density [16, 37], and neighboring objects' size [12, 42], can

influence visual perception. Better understanding such factors is critical to design proper statistical graphics, and to avoid potential misinterpretations of the data.

The goal of our work is to contribute a novel factor to this line of work: understanding neighborhood effects in bar charts. Given the prevalence of other visual biases, we were wondering in how far the perception of a bar's height is influenced by the height of its neighboring bars. For instance, can we observe any perceived differences when the very same target stimulus bar is surrounded by the top highest vs. the top lowest bars?

The main psychological foundation that guides the neighborhood manipulation in our study is Parallel Line Illusion (PLI). PLI illustrates how the perceived length of a target stimulus line is altered by different contextual lines positioned in parallel to it [11, 18–20]. There are two distortions in PLI: length contrast and length assimilation [18, 20]. Length contrast occurs when the target line is distorted away from the contextual line, *e.g.,* the target line perceived to be shorter when accompanied by a long contextual line as opposed to a short one; see Fig. 2. Length assimilation occurs when the target line is distorted towards the contextual line, *e.g.,* target line perceived to be longer when accompanied by a long contextual line than by a short one. These two distortions shift from one to the other under certain conditions.

A similar work to ours is Peebles' empirical investigation on whether two adjacent bar values affect the perception of a target bar value in a simple five-bar chart [29]. The author observed a significant difference in target bar value judgments between two conditions (the target bar measuring '5' surrounded by bars measuring '4, 4' vs. by '1, 1'). Our goal is to expand on this work and provide a more systematic investigation into potential neighborhood effects in bar charts. Simultaneously, we aim at moving from this highly-controlled, artificial setting towards a more realistic setting by examining real world datasets. Therefore, in this work we change the order of bars to keep datasets the same (constraint from real applications) instead of only changing the neighboring bar values (highly-controlled research setting).

Towards that goal, we contribute three Amazon Mechanical Turk (MTurk) experiments with overall 52,101 trials and 587 participants, in which we study these neighborhood effects. We tasked participants with *rank estimation* tasks. Rank estimation is a common, yet so-far mostly overlooked task in bar charts, in which users seek to relate a specific target item to the whole dataset. For example, one might want to learn about the ranking of one's own university, company, or sports club, according to some performance metric.

The results of our studies show that neighborhood effects exist in rank estimation tasks. When the same target bar is surrounded by the highest bars, the perceived target bar rank is lower than when it is surrounded by the lowest neighbors.

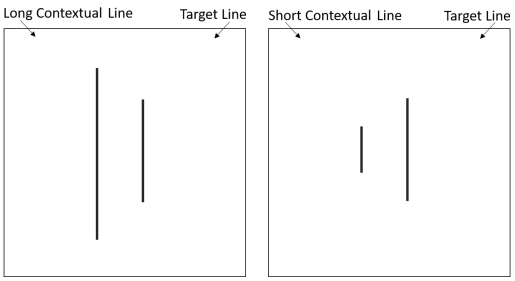

Figure 2: Parallel Line Illusion. The target line on the right side in both figures has exactly the same lengths. However, when put next to a long contextual line (left figure) it is perceived shorter as it actually is. In contrast, when the contextual line is short (right side) the target line is overestimated as too long. Length contrast is at play.

This finding is in line with the length contrast distortion in PLI. Yet, the effect size of neighborhood effects is small. Altering from the highest to the lowest neighborhood leads to an approximate 1.32 increase in the target bar rank. This means that a target bar being perceived as at rank 60 in its highest neighborhood will be perceived as at rank 61.32 in its lowest neighborhood. However, this small effect is largely masked by other data-inherent effects, which reveal that the accuracy of this task mostly depends on true target bar rank, number of data items and two other data characteristics (skewness and kurtosis) of the datasets. After isolating, identifying and quantifying both data-inherent effects and neighborhood effects, we draw three possible implications for future bar chart designs based on our results:

- Compared to random order, sorting bars according to real heights is of higher accuracy for conveying rank-related information.
- Despite their existence, there is no need to worry too much about neighborhood effects in bar chart rank estimation tasks.
- Skewness has a comparatively large effect on rank estimation though; additional clues might be used to convey rank-related information more precisely.

2 RELATED WORK

Empirical Studies on Ordering

The neighborhood manipulation is related with ordering graphical components (*i.e.,* how to map data to different graphical characteristics). This kind of mapping is a fundamental challenge in visual design, such as mapping rays in star plots. Various empirical studies exist that investigate effects of such ordering decisions. Klippel et al., for instance, studied how aligning variables to rays in star plots influences perceptual comparison tasks. They found that solely altering

the assignments of two rays in an 8-ray star plot resulted in significant differences in response time and participants' grouping behavior whereas this alter only evoked very few difference in perceived similarity [23]. Saket et al. compared between five different basic visualization types and revealed how they differ in effectiveness in ordering tasks [32]. To the best of our knowledge, no empirical studies had been conducted on ordering effects within bar charts.

Graphical Perception in Bar Charts

Much empirical work has been conducted to better understand the widely adopted visual encoding of bar charts. Utilizing a common scale, bar charts for instance were found to offer higher accuracy for estimating component parts than other square-, circle-, or cube-based encodings [5, 7, 35, 38]. These works focused on tasks such as larger/smaller comparison, difference comparison and proportion extraction between a pair of bars [5, 16, 24, 35]. Cleveland & McGill studied 10 different pairs of bar values in simple bar charts ranging from .18 to .83 and found that the accuracy would decrease with increasing distances between the graphical elements [5], a finding that was further confirmed by Talbot et al. [39]. These studies focused on the accuracy of judging absolute bar heights or comparing the relative height between a pair of bars. No work had shed light on the visual judgment of bar ranks, namely the relative position within the whole group.

Parallel Line Illusion

Theoretically, our work is grounded in neighborhood effects in Parallel Line Illusion (PLI) research. PLI was used to explain many different illusions [19], such as the famous Miller-Lyer [21], Baldwin [30], and Ponzo illusions [4].

Within the PLI framework also two different length distortions were described. The perceived length assimilation describes perceiving a test line length towards a context line. Length contrast describes distorted perception of a test line length away from the context line. Which direction an effect has (assimilation or contrast) depends on the spatial [18, 19] and temporal [20] distance between the two lines. Based on PLI, Zacks et al. [44] attributed the difference in accuracy of judging absolute bar height to potential length assimilation and length contrast distortions. They hypothesized that these effects might be produced by the relative heights of bars in a figure and the relative heights of judged bars to its surrounding graphical frame [44]. Closest to our work, Peebles [29] explored whether the surrounding bar values affected perceived bar values in a simple bar chart (comprised of five bar components). They compared the perceived values of a target bar measuring '5' under two conditions (the two adjacent bar values to be '1,1' vs. '4,4') [29]. Results showed a significant difference that the mean perceived value for '1,1' condition was '5.07' while for '4,4' was '4.96'. This is consistent with length contrast in PLI. Yet, this study only tested one, highly artificial stimulus chart. Building on PLI and these initial findings, we study length assimilation and length contrast effects in much more realistic settings of bar charts, with an eye towards making this theory applicable in visual design.

3 OVERVIEW

The three studies were conducted in chronological order between May 2018 and September 2018. Study 1 (pilot) explored the existence of neighborhood effects with target bars at three different ranks being surround by three neighborhood conditions (highest, similar and lowest neighborhood). After identifying that other data-inherent effects had a strong confounding influence on our results, we conducted study 2 (pilot) to isolate and quantify these other effects by testing each bar in the chart under two baseline orders. In study 3 (main study), we offset the impact from these other data-inherent effects and validated the presence of neighborhood effects with target bars at six different ranks being surrounded by two neighborhood conditions (highest and lowest neighborhood). All three studies were conducted on MTurk. Details of each study are shown in Tab.1.

Hypothesis

Following Peebles' results [29] and fundamental PLI research [18], we hypothesized that neighborhood effects exist in bar charts. Jordan et al. [18] found that in very close neighborhoods/spatial separation (5mm) assimilation effects dominate, while in larger neighborhoods (100mm) contrast effects are more common. In our case, the width of the neighborhood around the target bar (distance from the leftmost neighboring bar to the rightmost neighboring bar) was roughly in the range of 100mm. We thus hypothesized that contrast effects would occur. Surrounded by the top highest bars, the target bar rank would be perceived lower than that surrounded by the top lowest bars.

Datasets and Participants

To increase ecological validity, we carefully handpicked 15 nominal datasets which could be applied to a univariate bar chart from online resources statistics.com [36] and New York Times [28]. These 15 datasets were equally distributed into three groups with 9-13 bars (referred to as approx. 10 bars group), 18-24 bars (20 bars group), and 30-32 bars (30 bars group). Participants were recruited from MTurk and compensated at an hourly payment of $7.5.

Experimental Setting

We task participants with the rank estimation task. We regard it as the simpler testbed (mainly involves one target bar and

Table 1: Overview of studies. Details include main results, number of participants recruited and remained after quality control, and number of tasks in each MTurk HIT.

Study	Result	Recruited	Used	Tasks
Pilot 1	Neighborhood effects seem to exist; however, various confounds mask the effects.	105	94	135
Pilot 2	Modeling data-inherent effects that explain the confounds from study 1.	200	184	63 (61)
Main 3	With controlled confounds, the neighborhood effect is present with small effect size.	283	200	72

the others) for investigating neighborhood effects than other tasks, *e.g.,* to compare heights between two bars (involves two targets and the rest). We calculate rank as

$$Rank = \frac{OR}{N} \times 100 \quad (1)$$

where OR denotes the ordinal rank, and N denotes number of data items in the dataset.

We defined a manipulation neighborhood covering 20% of the entire chart (*i.e.,* 2, 4, and 6 bars in the three groups); see gray shadow in Fig. 1 (c) & (e). Fig. 1 exemplifies our neighborhood manipulation process. Starting with a quasi-random alphabetical order, we first identified the set of highest or lowest neighboring bars (Fig. 1 (b) & (d)); we then swapped them with the initial neighborhood bars around the target (Fig. 1 (c) & (e)). The target bar (orange) was randomly chosen within the possible range. Choices that would cut off the neighborhoods at the sides of the chart were excluded.

We formulated the rank estimation task as "How do you think about the height of the target bar highlighted in orange among the whole group?" We chose this formulation to avoid misunderstanding the task as comparing the target bar with another bar such as the highest one. We only showed each chart to participants for exactly one second. With this setting, we sought to prevent participants from precisely counting how many bars are above or below a target. We tested different time durations before the studies and found that one second gave the best tradeoff between bar chart perception and avoiding counting strategies.

The whole chart (measuring about 960:500 pixels) was rendered in the center of the screen, below a control panel containing the question and a slider (measuring about 600 pixels). The target bar was highlighted in orange while the other bars filled in blue. Titles of data items are replaced with ordinal numbers. For all conditions, the visual appearance of bars and tasks were exactly the same. The auxiliary materials contain all stimuli and screenshots of the setup.

Participants' responses were collected on a continuous slider with #bar-times tick marks. The two ends were clearly marked as 'Lowest' and 'Highest' in order to avoid participants accidentally swapping the scale direction. Every time a new task started, the slider handle was visually hidden. The handle appeared after a click on the slider and could be further adjusted then, always snapped to the nearest tick position though (ordinal rank). We measured estimation error in target bar rank as follows:

$$Error = \frac{Perceived\ OR - True\ OR}{N} \times 100 \quad (2)$$

Perceived OR denotes ordinal rank collected on sliders and True OR denotes the actual ordinal rank of the target bar.

Quality Control

We set quality control tasks to ensure serious participation from crowdsourcing workers. In study 1 and 3, quality control trials which shared the same task (rank estimation) and setup with the main experiment were randomly inserted into the study. The target bars in quality control trials were either the top highest, the top lowest or the median bar in a sorted order according to heights. If a participant failed (absolute error of OR exceeded 25%) more than 50 percent of all the quality control tasks, his/her entire response was discarded. In study 2, we control data quality based on outlier detection in main study tasks.

Statistical Analysis

Following advise on good statistical practices [2, 6, 8, 25], we use a graphical presentation approach to present effect sizes and 95% confidence intervals (CIs). Specifically, we adopt the 95% corrected Cousineau-Morey CI generating method [2, 6], which is recommended for within-subject designs as it can strip out individual differences [2].

In regression analysis, we use adjusted R^2 and stepwise model selection which is based on Akaike Information Criterion (AIC) [1] from both backward and forward directions to compare models. ANOVA tests (F-test and p-value) are adopted to test significance of models and predictors.

4 EXPERIMENT

Study 1

Objective: We set out to investigate neighborhood effects with target bars at three different ranks. Each target was surrounded by three neighborhood conditions. We expected to see substantial differences in the perceived rank task for different neighborhood conditions.

Setup: The three target bars were chosen to be low bars ($rank = 25, mean = 27.07, SD = 2.83$), median bars ($rank = 50, mean = 51.10, SD = 1.84$), and high bars ($rank = 75, mean = 77.52, SD = 1.73$). The three neighborhood conditions consisted of the shortest bars, bars similar to the target (*e.g.,* for target bars ranked at 50%, neighboring bars ranged from 40% to 60%), and the highest bars. Followed by manipulation procedures discussed in Section 3, reordering was performed on the alphabetical baseline.

A within-subject study was conducted with 105 MTurk workers. The whole study lasted about 10 minutes. After 9 training tasks presented in a fixed order at the beginning, the main experiment contained 135 perceptual ranking tasks (135 $Trials = 3\ Targets \times 3\ Neighborhoods \times 15\ Datasets$) and five quality control tasks was presented. Task order in the main experiment was randomized. To eliminate fatigue, we gave participants a 30s break after 5 minutes.

Results: We accepted 94 responses after quality control and plotted the mean error (effect size and CI). Fig. 3 shows the results aggregated over all trials and datasets: mean errors above zero indicate overestimation, while errors below indicate underestimation. The figure is first organized by the target bar rank condition (25%, 50%, 75%), and then by neighborhood conditions (lowest, similar, highest).

Looking at the three different target bar rank groups (25%, 50%, 75%), we see a neighborhood effect within each of them. The highest neighborhood condition is always more underestimated than the other two (lowest, similar). That is, people tend to perceive target bar rank in its lowest neighborhood to be higher than that in its highest neighborhood. The similar neighborhood conditions have a substantial overlap with the lowest neighborhood conditions, so we cannot make reliable judgments about potential differences between them.

Considering the results for each dataset separately (screenshots are available in the auxiliary material), we observe these neighborhood effects in 11 out of 15 datasets (19 out of 45 targets). In three other datasets (three out of 45 targets), we see a reversed pattern that is consistent with length assimilation distortion. Bar orders in the alphabetical baseline could be the confounds for inducing the reversed pattern.

While we could see neighborhood effects within each of the three groups, we were surprised by the much stronger effects that exist between these groups. In Fig. 3, it is obvious that a large difference in error exists between the three target rank groups (25%, 50%, 75%). There is a substantial underestimation at 75%, a slight underestimation at 50%, and a slight overestimation at 25%. These are likely caused by other factors such as the choice of the target bar, or potentially also other data-inherent characteristics. One possible explanation for these effects might be anchoring and insufficient adjustment [10]. It is possible that participants swipe a horizontal line between the highest and lowest bars to be the 'median'

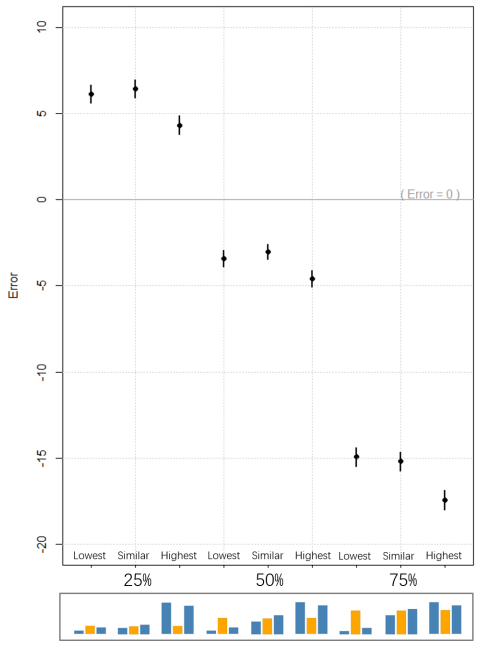

Figure 3: CIs for means of Error. We tested nine conditions: Three target bars (25%, 50%, 75%), each tested with three neighborhoods (lowest, similar, and highest).

height and then conservatively indicate whether the target bar is higher or lower than this height. Another possible explanation is the statistical phenomenon regression to the mean [9, 15] in repeated measurements. When viewers make decisions based on subjective probability distribution [13] (*e.g.,* generating a quantification of certainty on how likely this target bar to be at certain rank), the estimates for an unknown target tend to cluster to the mean of population distribution, *i.e.,* the median 50^{th} percentile in our rank estimation task. Both these two explanations help us understand that estimations at 75% group are under- and those in the 25% group are overestimated.

Study 2

Objective: The purpose of study 2 was to isolate and quantify other data-inherent effects that might influence the rank estimation task. Two baseline conditions were chosen to model these data-inherent effects: (B1) random condition with data items randomly assigned to bar components and (B2) sorted condition with bars sorted in an ascending order according to real heights.

Hypothesis: Based on the lessons learned from study 1, we hypothesized that a large proportion of error could be explained by the actual target bar rank and number of data items. To some degree also data characteristics such as coefficient of variance (ratio between standard deviation to mean), skewness (asymmetry of the probability distribution),

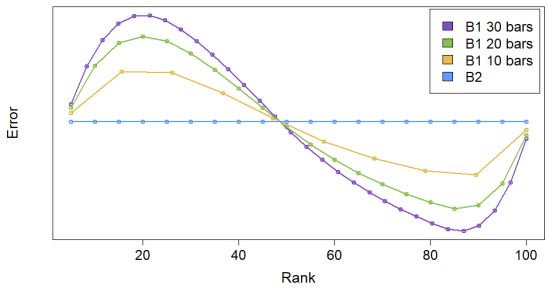

Figure 4: Hypothesized data-inherent effects from true rank and number of data items in two baseline conditions.

kurtosis ("tailedness" of the probability distribution) and normality might play a role.

We expected to see differences between B1 and B2 conditions. Errors in B1 condition would have larger variances and the same data-inherent effects would appear as in study 1. Errors in B2 condition would have smaller variances and only little data-inherent effects would be observed considering the additional visual assistance from bars being sorted along the x-axis.

Fig. 4 illustrates our anticipated effect of true target bar rank. We represent it as a curvilinear relationship between errors and true target bar ranks. We hypothesized that viewers would anchor and adjust their answers on median bars (ranked at 50) and the two extrema (ranked at 1 and 100). Means of error near these anchors would be close to zero, while for the ranks in-between we expected their perception to regress towards the anchors. The expected relationship in B2 was a horizontal line very close to x-axis, which means no effect. In terms of number of bars in a chart, we expected that with increasing number of data items, also the absolute error would increase in condition B1. This is illustrated with different colors in Fig. 4.

In terms of regression analysis, we expected to see a contribution to the model's goodness of fit (increased adjusted R^2, and a smaller AIC value) from target bar rank and number of data items when adding each of them as variables into the model. An increase in R^2 indicates that the model better explains the variance in the data. AIC deals with the tradeoff between model's goodness of fit and complexity (number of predictors), with a lower AIC indicating a superior model.

Setup: We conducted a new study to have every single bar component in 6 out of 15 datasets (randomly chosen) being assessed under both B1 and B2 conditions. For practical reasons we split the study into two blocks completed by two groups of participants (Block 1: charts were chosen to be N = 13, 18, 32; Block 2: N = 10, 20, 31). After the same 9 training tasks in pilot study 1, there were 126 tasks for each participant in Block 1 and 122 tasks for Block 2. The presentation order of main study tasks within each block was randomized. Each block averaged 10 minutes. 200 MTurk workers were randomly assigned to one block.

According to factorial analysis, the total number of permuted orders in B1 condition for a 10-bar chart is more than 10^6. To increase randomness, bar orders in B1 condition were randomized when the charts got rendered in the front end. Ideally, with 100 participants in each block, we could average between 100 random orders. For the sorted condition (B2), the order was always the same.

Result: To remove speeders (workers who rushed through the study with low data quality), we used $1.5 \times IQR$ rule to detect outliers in B2 sorted baselines. There remained 95 and 89 participants for Block 1 and Block 2 respectively. In total we kept 22,828 samples for data analysis. Comparing between errors, we observe smaller variances in B2 ($mean = -1.62, SD = 11.40, CI[-1.79, -1.46]$) than in B1 ($mean = -2.13, SD = 19.34, CI[-2.30, -1.97]$). Large variances in B1 are plausible as we let participants judge different charts for the same condition and there can be large individual differences between participants in empirical studies.

We first tested the hypothesized curve by modeling relationship between error and target bar rank. We applied the $1.5 \times IQR$ rule in each B1 condition to remove outliers. 10,863 out of 11,414 samples remained. We applied two regression models: a polynomial regression model and Locally Weighted Regression (LOESS) [27]. For the polynomial regression model, we chose a cubic fitting function as the simplest function capable of modeling our hypothesized relationship. By contrast, LOESS is a non-parametric regression method that automatically finds the curve that best describes the relationship between independent and dependent variables without the need to specify the fitting function.

To derive prediction models, we averaged all the repeated measurements for the same condition (124 conditions) instead of using the 10,863 original samples. While models derived from different #regression samples have different R^2, their coefficients and the ANOVA test results are identical. Thus, we reported the model derived from 124 averaged samples as it introduces less uncertainty in the prediction.

After applying stepwise model selection, we derive a cubic model with the lowest AIC (denoted as Model 1). Fig. 5 illustrates the fitting line and prediction intervals (*i.e.,* 95% CI and 95% PI) of Model 1, fitting results from LOESS and regression samples. LOESS results (Residual Standard Error is 11.07) are consistent with the hypothesized curvilinear relationship between error and target bar rank. The results from the cubic model and LOESS are identical. All predictors in Model 1 are significant, revealing their associations with the dependent variable. The curvilinear relationship can be expressed as:

$$Error = 6.79 - 0.00735 \times Rank^2 + 0.0000640 \times Rank^3 \quad (3)$$

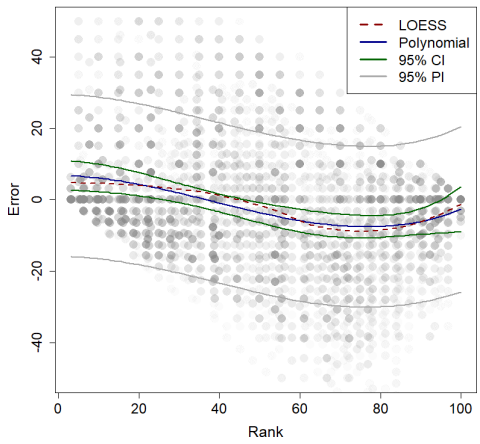

Figure 5: Model 1 predictions and regression samples (scatterplots, alpha = 0.03): Errors against true target bar ranks, indicating the influence from ranks.

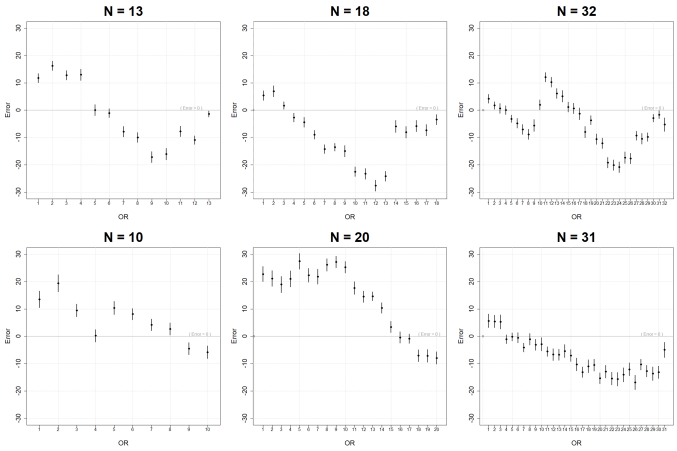

Figure 6: Error against rank, indicating the data-inherent effects from datasets. Error means and 95% CIs are plotted for each target bar.

However, this model does not yield a good model fit, $F(2, 121) = 11.51, p = 2.66e^{-5}$, with adjusted R^2 to be 0.14 and AIC to be 957.74. Even though the curvilinear relationship we see in Fig. 5 follows our hypothesis, only using rank to account for estimation errors is insufficient.

To investigate the potential influence of other data characteristics, we plotted error means and CIs for each target bar (keep all the raw samples to avoid biases in CIs) as shown in Fig. 6. The pattern varies across datasets. For charts N = 13 and 32, there is the curvilinear relationship that we hypothesized. In the other charts, however, we only see large slopes or other curvilinear relations. In the two charts N = 13 and 32, x-intercepts are very close to 50, but the other four charts do not show similar patterns. We cannot confirm the hypothesis about the effects from number of data items with these graphical presentations in that the plots differ vastly.

We further hypothesized that besides target bar rank and number of data items, other data characteristics might also influence the accuracy of rank estimation tasks. We thus added Coefficient of Variation (CV), skewness, kurtosis and the p-value in Shapiro-Wilk test [34] (a preferred method to test whether the dataset is drawn from normal distribution [43]) of each dataset to the model to see whether these additions can further explain the errors we see. After stepwise model selection, we derive a new multiple cubic regression model denoted as Model 2. As shown in Tab. 2, p-values for each predictor in Model 2 are statistically significant, meaning that changes in true target bar ranks, number of data items, skewness, kurtosis and the interaction between target bar rank and number of data items are related to the changes in the estimation error. Model 2 yielded a good model fit on the same 124 samples, $F(6, 117) = 48.62, p < 2.2e^{-16}$, with adjusted R^2 to be 0.70 and AIC to be 832.22.

Table 2: Coefficients of multiple cubic regression using six predictors, *i.e.*, two polynomial components of target bar rank (Rank), linear components including number of data items (N), skewness, kurtosis, and the interaction between Rank and N (N:Rank).

	Estimate	Std.Error	t value	Pr (>|t|)
(Intercept)	6.37*e*+00	6.65*e*+00	0.96	0.34
$Rank^3$	8.50*e*-05	1.47*e*-05	5.80	5.81*e*-08***
$Rank^2$	-1.08*e*-02	1.78*e*-03	-6.06	1.71*e*-08***
N	-1.26*e*+00	1.53*e*-01	-8.26	2.61*e*-13***
Skewness	-2.39*e*+01	2.07*e*+00	-11.53	< 2*e*-16***
Kurtosis	1.50*e*+01	3.02*e*+00	4.97	2.31*e*-06***
N:Rank	6.41*e*-03	2.47*e*-03	2.59	0.01*

Comparing between the two models, we find that Model 2 is superior to Model 1. Adjusted R^2 increases from Model 1 to Model 2 ($Adj.R^2_{\text{Fit1}} = 0.14, Adj.R^2_{\text{Fit2}} = 0.70$) while AIC drops ($AIC_{\text{Fit1}} = 957.74, AIC_{\text{Fit2}} = 832.22$), which validates that number of data items, skewness, kurtosis and the interaction between rank and number of data items are explaining a large and unique proportion of variances that target bar rank related terms cannot. Adding rank related terms as the last additions results in an increase in adjusted R^2 (from 0.61 to 0.70) and a decrease in the AIC score (from 862.08 to 832.22), which shows the contribution from true target rank.

In Fig. 7, we plot the predictions of Model 2 together with the data we trained it with. As reported in Tab. 2, there is a negative correlation between error and number of data items, which means a larger number of bars in the chart will bring more underestimation in rank. However, in Fig. 7, the chart of 20 data items seems to be an outlier as it does lie on top, not

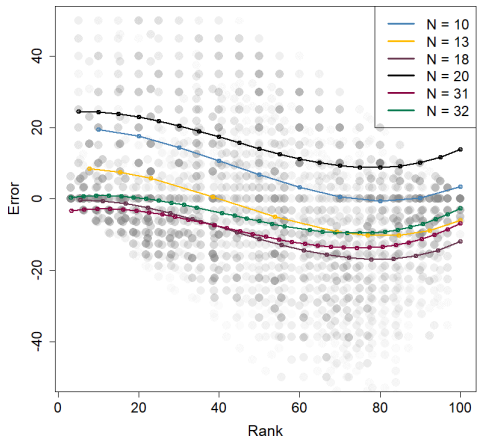

Figure 7: Model 2 predictions (color encodes datasets) and regression samples (scatterplots, alpha = 0.03) for B1 random condition. Data-inherent effects are salient.

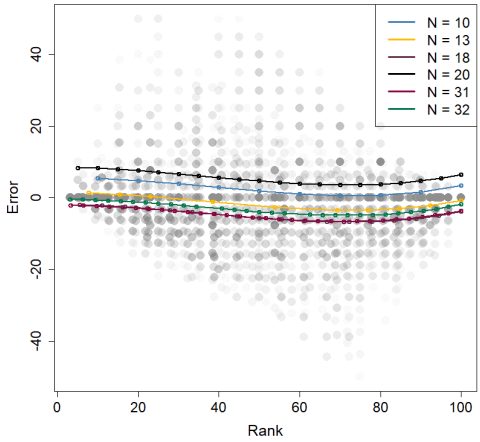

Figure 8: Predictions (color encodes datasets) and regression samples (scatterplots, alpha = 0.03) for B2 sorted condition. Data-inherent effects are of smaller effect size.

following this negative relationship. When looking into more details, we find that this is related to the predictor skewness. Skewness has a large negative coefficient -23.86 and the chart N = 20 has the smallest skewness score -0.62 ($mean = 0.17, SD = 0.60$) among all the datasets. Thus, we see the largest amount of overestimation in this chart, which further indicates the large influence from other data characteristics. We will leave a full and systematic characterization of the influence from such data-inherent factors to future work.

In addition, Model 2 satisfies the assumptions on both the curvilinear relationship between predictors and dependent variables and the independence of predictors. Stepwise model selection helps us guard against multicollinearity. Model 2 also satisfies other Ordinal Least Square (OLS) assumptions with *residuals* (*i.e.,* vertical distances between observed values and fitted values) randomly scattered around zero, which means there is no heteroscedasticity. We performed out-of-sample testing on Model 2 with the same 6 datasets from pilot study 1 and adjusted R^2 is 0.81. In out-of-sample testing with the 9 datasets not leveraged for deriving Model 2, adjusted R^2 averages 0.87 ($SD = 0.16$). We also test against overfitting. Observations per term in Model 2 are 20.67, which is sufficient (rule of thumb: 10-15 observations per term) in multiple linear regressions (including polynomial terms and interactions). The predicted R^2 for Model 2 is 0.68, indicating only 2% (70% - 68%) of variances is explained by too many factors or random correlations. As predicted R^2 is close to adjusted R^2, we know that our model is not severely overfitting.

For B2 sorted baseline, we applied the same fitting function as that for Model 2 (see predictions in Fig. 8). Adjusted R^2 for B2 equals 0.48 with only the interaction between target bar rank and number of data items to be non-significant. Compared with B1, there is less data-inherent effects in B2.

We conclude that there are different data-inherent factors that influence rank estimation tasks. We can build a rough model based on our data as a means to filter out noise in our investigations of neighborhood effects. Yet, further work is needed to clearly investigate these data-inherent effects.

Study 3

After accounting for data-inherent effects, we conducted a new study to validate the effect from neighborhoods.

Hypothesis: We hypothesized that neighborhood effects still exist after offsetting the influence from data-inherent effects. We hoped to explain the difference between two neighborhood conditions by a new linear predictor: neighborhood rank (100 for the highest vs. 1 for the lowest) and anticipated its coefficient to be negative.

Setup: Similar to pilot study 1, we generated two neighborhood conditions (*i.e.,* the highest and the lowest) on randomly ordered bar charts. Target bar ranks were chosen to be 30, 40, 50, 60, 70, 80 in the 6 datasets we used in study 2. In total, there were 9 training tasks, 72 main study tasks, and five quality control tasks in the study. Task order was randomized. We increased the number of target bar ranks to 6 for eliminating potential learning effects. 283 MTurk workers were recruited and the study lasted about 5 minutes.

Results: 200 participants remained in the sample after data cleaning. The results we saw in study 3 were very similar to study 1. Plotting aggregated mean errors and CIs from a total of 14,400 samples, there is no overlap between the highest and the lowest neighborhood conditions (see Fig. 9). Neighborhood effects show up at five out of six target bar rank levels. There is a substantial overlap between CIs of the two neighborhoods at target bar rank = 60. Looking into more details, we see neighborhood effects in 13 targets (in

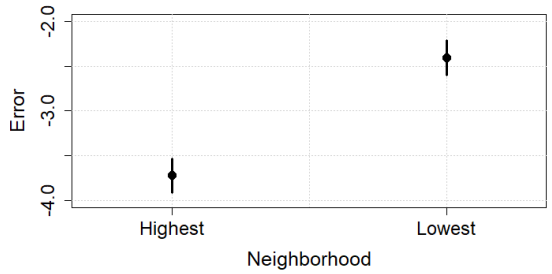

Figure 9: Neighborhood effects: Mean errors and CIs in two neighborhoods with all 14,400 samples from study 3.

total 36) from 5 datasets while other two targets showed a reversed pattern corresponding to length assimilation in PLI.

Next, we performed regression analysis over the 14,400 collected samples. First, we only use the linear predictor neighborhood rank to explain residuals. Results show that this predictor is significant (coefficient is estimated to be $-1.34e^{-2}$, p = $3.15e^{-6}$) but it only accounts for a small percentage of all the variances in error ($Adj.R^2 = 1.44e^{-3}$). Further exploration shows that the target bar rank and other factors related to datasets are also associated with changes in residuals. Then we extended Model 2 by adding neighborhood rank as a new linear predictor to directly account for variances of error in study 3. After stepwise model selection, we derived Model 3 ($F(7, 14392) = 1208, p < 2.2e^{-16}, AIC = 121329.8, Adj.R^2 = 0.37$). As shown in Tab. 3, all predictors are significant, meaning that they are likely to be meaningful addition to the model. The coefficient of neighborhood rank is negative as hypothesized. A target bar being judged as at rank *e.g.,* 60 in its highest neighborhood will be judged as at 61.32 in its lowest neighborhood. Adding neighborhood rank as the last predictor improves model fit (AIC drops by 24.8 and Adj.R^2 increases by 0.001)[1]. Based on Model 3, we also sorted predictors according to their effect sizes using Cohen's f2 (Large effect: 0.35; Medium: 0.15; Small: 0.02 [33]): skewness (0.42) > target bar rank^3 (0.12) > N (0.021) > target bar rank^2 (0.015) > kurtosis (0.010) > neighborhood rank (0.002). Unit increase (value +1.00) in skewness brings -25.00 in rank; unit increase in number of data items results in -0.57 while unit increase in kurtosis brings +7.69 in rank. This shows how each data characteristics influences the estimation error. The power analyses we conducted for each predictor in Model 2 and Model 3 revealed that all the power values exceeded 0.99. This means that the sample sizes are adequate to catch effect sizes for all predictors.

Note that although Model 3 only explains 37% of all the variances, there is still a reliable relationship between the significant predictors and the dependent variable. Large inter-individual differences and higher level cognitive tasks [5] can be the underlying reasons for such unexplainable high-variability. Predicted R^2 for Model 3 equals 0.3695, which is close to the adjusted R^2, meaning that Model 3 is not overfitting. Both the predictions from Model 3 (see Fig. 10) and Cohen's f2 suggest that compared to the data-inherent effects, the impact from neighborhoods is rather small.

With all these analyses, we conclude that neighborhood effects exist but their effect size is small on the rank estimation tasks.

[1]We only compared models trained with the same regression samples.

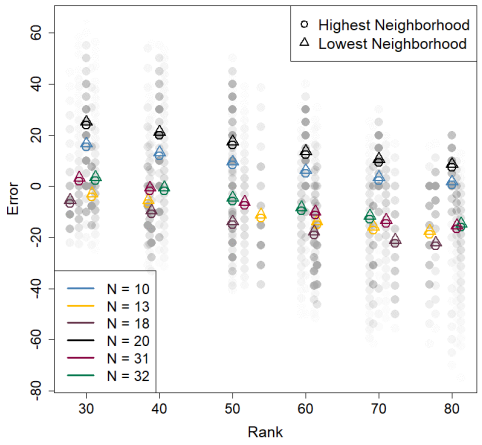

Figure 10: Model 3 predictions (color encodes datasets) and regression samples (scatterplots, alpha = 0.01) showing both neighborhood effects and data-inherent effects.

Table 3: Coefficients of Model 3, extending Tab.2 with the predictor neighborhood rank (Neighbor).

	Estimate	Std.Error	t value	Pr (>|t|)
(Intercept)	1.38*e*+01	1.52*e*+00	9.10	< 2*e*-16***
Rank^3	5.95*e*-05	7.18*e*-06	8.28	< 2*e*-16***
Rank^2	-7.65*e*-03	6.99*e*-04	-10.95	< 2*e*-16***
N	-5.73*e*-01	5.60*e*-02	-10.22	< 2*e*-16 ***
Skewness	-2.51*e*+01	4.74*e*-01	-53.08	< 2*e*-16***
Kurtosis	7.70*e*+00	6.54*e*-01	11.77	< 2*e*-16***
Neighbor	-1.33*e*-02	2.75*e*-03	-4.84	1.31*e*-06***
N:Rank	-3.27*e*-03	9.56*e*-04	-3.42	0.000629***

5 DISCUSSION

In this section, we discuss design implications from our results, limitations, and future work.

Design implications: Random (alphabetical order as one typical case), and sorted orders are the two most commonly used approaches in bar charts. In study 2, we hypothesized that we would see a perceptual effect of data-inherent factors in random orders while see very little of this effect in sorted orders. This hypothesis is partially confirmed. We also

see some data-inherent effects in sorted orders although the effect size ($Adj.R^2 = 0.48$) is smaller than that in random orders ($Adj.R^2 = 0.70$). Nevertheless, the sorted order is still the more accurate approach to convey rank-related information as the mean of all estimation errors is closer to zero with smaller variances. Besides, results suggest that—despite their existence—designers do not need to worry too much about neighborhood effects. Our findings show that a local neighborhood manipulation does not substantially influence the perceptual rank estimation tasks in bar charts.

To the best of our knowledge, no one else has investigated perceptions on rank estimation tasks so far. A surprising result was the large amount of biases that is introduced by data-inherent effects. Following Cleveland & McGill [5], we calculate the accuracy defined on estimation errors for our task, $Accuracy = log_2(|estimated\ value - true\ value| + .125)$. The accuracy can exceed 4.0 (marked as a large log absolute error [5]) when the estimation error goes beyond 20 in rank. This means the accuracy can be rather low in our task. Considering the comparatively large effect size of data-inherent factors such as skewness, we encourage designers to provide additional clues if they would like to convey rank-related information precisely.

Limitation: We understand neighborhood effects as an interplay between all bars instead of the influence only from neighboring bars. Although we hypothesized to see length contrast effect in our study (based on previous work [18, 29]), there is no reason to rule out length assimilation effect under other experimental settings. Confounding factors could come from the 'shape' in the baseline charts (*e.g.,* influence from Gestalt symmetry and continuity laws [31, 41]) and target bar positions in the chart. In three out of five assimilation cases (from 5 different datasets where length contrast effect also showed up), we found that the top highest bar was just positioned at/very close to the border of the lowest neighborhood (study 1 screenshots #96 and #123, study 3 screenshot #58 in auxiliary material), which could be the confounding factor reversing the effect.

Following previous work in quantifying perceptual biases in visualizations [40], we used multiple regression to quantify both the data-inherent and neighborhood effects. We simplify the models by neglecting potential interactions between predictors. From further analysis we found that these complex but currently excluded interactions (*e.g.,* the interaction between target bar rank and skewness) also contributed to model fit and helped in explaining varying patterns between datasets (Fig. 6), which demonstrates the challenge and importance of using real datasets. Also, the relationship between estimation errors and target bar ranks is simplified by using a cubic fitting function and the inter-individual differences [3] are not taken into account in current models.

One of the differences from our work to others on bar chart perceptions is the one-second time limit for presenting charts. Cleveland & McGill pointed out that a substantial danger in graphical perceptual research is to perform judgments differently from the way people adopt in real life [5]. A possible solution is to have participants judge much more quickly than they do in real life in order to prevent them from performing higher level cognitive tasks other than basic graphical perceptual tasks. This time limit in our work helped in preventing precise counting on the bars, but it might also result in other confounds such as random judgements.

Another difference is that we used several real datasets instead of simple, highly-controlled stimuli. Our goal in doing so was to bring this line of research closer to applied visualization design. As a consequence, however, our results also did not lead to a simple, clear-cut picture. We found that patterns varied between datasets, and even observed patterns contrary to our hypothesized effects. Moreover, one could argue that the observed neighborhood effects may only be exclusive to the six datasets in our study. Although including more datasets could decrease the consistency and accuracy of a study, we still believe our current results provide some actionable insights into the question of how to account for neighboring effects when designing bar chart visualizations.

Future work: One direction of future work is to expand this work by including more datasets. As our study started from an exploratory end with no benchmark, we also suggest exploring other parameters, *e.g.,* chart exposure time, neighborhood size, neighboring bar ranks, and measurements (*e.g.,* sliders in our study). Although we found evidence of different rank estimation effects, we haven't figured out any counter-acting methods yet. Find counter-acting methods might become even more complicated due to constraints from real world requirements such as efficiency (*e.g.,* to find every target item quickly) and saliency (*e.g.,* to notice a distinct target item quickly). We regard our work as a first step and hope it can inspire more work on investigating solutions for counter-acting perceptual biases.

6 CONCLUSION

We studied the perception of ordering in bar charts on rank estimation tasks. Our results showed that neighborhood effects can be observed yet their effect size is small. Accuracy on these rank estimation tasks is dominated by other effects stemming from the choice of target bars and certain dataset characteristics. We regard our work as a further piece to better understand the perception of the widely-used bar chart encoding.

7 ACKNOWLEDGEMENT

We thank all study participants and anonymous reviewers. This work is partly funded by RGC GRF grant 16213317.

REFERENCES

[1] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19, 6 (1974), 716–723.

[2] Thom Baguley. 2012. Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior research methods* 44, 1 (2012), 158–175.

[3] Daniel J Bauer. 2011. Evaluating individual differences in psychological processes. *Current Directions in Psychological Science* 20, 2 (2011), 115–118.

[4] Richard W Brislin. 1974. The Ponzo illusion: Additional cues, age, orientation, and culture. *Journal of Cross-cultural Psychology* 5, 2 (1974), 139–161.

[5] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.

[6] Denis Cousineau. 2017. Varieties of Confidence Intervals. *Advances in cognitive psychology* 13, 2 (2017), 140.

[7] Frederick E Croxton and Harold Stein. 1932. Graphic comparisons by bars, squares, circles, and cubes. *J. Amer. Statist. Assoc.* 27, 177 (1932), 54–60.

[8] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[9] Clarence E Davis. 2007. Regression to the mean. *Wiley Encyclopedia of Clinical Trials* (2007), 1–2.

[10] Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science* 17, 4 (2006), 311–318.

[11] Volker H Franz, Manfred Fahle, Karl R Gegenfurtner, and Heinrich H Bülthoff. 2000. Effects of visual illusions on grasping: The Parallel-Lines Illusion. *Perception* 48 (2000), 50.

[12] Patricia P Gilmartin. 1981. Influences of map context on circle perception. *Annals of the Association of American Geographers* 71, 2 (1981), 253–258.

[13] JM Hampton, PG Moore, and Howard Thomas. 1973. Subjective probability and its measurement. *Journal of the Royal Statistical Society. Series A (General)* (1973), 21–42.

[14] Lane Harrison, Drew Skau, Steven Franconeri, Aidong Lu, and Remco Chang. 2013. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2949–2958.

[15] MJR Healy and H Goldstein. 1978. Regression to the mean. *Annals of Human Biology* 5, 3 (1978), 277–280.

[16] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 203–212.

[17] Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1461–1470.

[18] Kevin Jordan and Peter W English. 1989. Simultaneous sampling and length contrast. *Perception & Psychophysics* 46, 6 (1989), 546–554.

[19] Kevin Jordan and Diane J Schiano. 1986. Serial processing and the parallel-lines illusion: Length contrast through relative spatial separation of contours. *Perception & Psychophysics* 40, 6 (1986), 384–390.

[20] Kevin Jordan and John Uhlarik. 1985. Assimilation and contrast of perceived length depend on temporal factors. *Perception & psychophysics* 37, 5 (1985), 447–454.

[21] Charles H Judd. 1905. The Muller-Lyer illusion. *The Psychological Review: Monograph Supplements* (1905).

[22] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, Kwon-Hyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 272–283.

[23] Alexander Klippel, Frank Hardisty, and Chris Weaver. 2009. Star plots: How shape characteristics influence classification tasks. *Cartography and Geographic Information Science* 36, 2 (2009), 149–163.

[24] Yi-Na Li, Kang Zhang, and Dong-Jin Li. 2017. How dimensional and semantic attributes of visual sign influence relative value estimation. *ACM Transactions on Applied Perception (TAP)* 14, 3 (2017), 18.

[25] Michael EJ Masson and Geoffrey R Loftus. 2003. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 57, 3 (2003), 203.

[26] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5421–5432.

[27] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons.

[28] Nytimes.Com. 2018. The New York Times - Breaking News, World News & Multimedia. Retrieved June 15, 2018 from https://www.nytimes.com

[29] David Peebles. 2004. Distortions of perceptual judgment in diagrammatic representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 26.

[30] Alexander W Pressey and Nancy E Smith. 1986. The effects of location, orientation, and cumulation of boxes in the Baldwin illusion. *Perception & Psychophysics* 40, 5 (1986), 344–350.

[31] Irvin Rock and Stephen Palmer. 1990. The legacy of Gestalt psychology. *Scientific American* 263, 6 (1990), 84–91.

[32] Bahador Saket, Alex Endert, and Cagatay Demiralp. 2018. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics* (2018).

[33] Arielle S Selya, Jennifer S Rose, Lisa C Dierker, Donald Hedeker, and Robin J Mermelstein. 2012. A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. *Frontiers in psychology* 3 (2012), 111.

[34] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.

[35] David Simkin and Reid Hastie. 1987. An information-processing analysis of graph perception. *J. Amer. Statist. Assoc.* 82, 398 (1987), 454–465.

[36] Statista.Com. 2018. Statista - The portal for statistics. Retrieved Jun 15, 2018 from https://www.statista.com

[37] Maureen Stone and Lyn Bartram. 2008. Alpha, contrast and the perception of visual metadata. In *Color and Imaging Conference*, Vol. 2008. Society for Imaging Science and Technology, 355–359.

[38] Marc Streit and Nils Gehlenborg. 2014. Bar charts and box plots. *Nature Methods* 11, 2 (2014), 117–117. https://doi.org/10.1038/nmeth.2807

[39] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2152–2160.

[40] Andre Calero Valdez, Martina Ziefle, and Michael Sedlmair. 2018. Priming and anchoring effects in visualization. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 584–594.

[41] Colin Ware. 2012. *Information visualization: perception for design*. Elsevier.

[42] Daniel J Weintraub. 1979. Ebbinghaus illusion: context, contour, and age influence the judged size of a circle amidst circles. *Journal of Experimental Psychology: Human Perception and Performance* 5, 2 (1979), 353.

[43] Bee Wah Yap and Chiaw Hock Sim. 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation* 81, 12 (2011), 2141–2155.

[44] Jeff Zacks, Ellen Levy, Barbara Tversky, and Diane J Schiano. 1998. Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of experimental psychology: Applied* 4, 2 (1998), 119.