Four Experiments on the Perception of Bar Charts

Justin Talbot, Vidya Setlur, and Anushka Anand

Abstract—Bar charts are one of the most common visualization types. In a classic graphical perception paper, Cleveland & McGill studied how different bar chart designs impact the accuracy with which viewers can complete simple perceptual tasks. They found that people perform substantially worse on stacked bar charts than on aligned bar charts, and that comparisons between adjacent bars are more accurate than between widely separated bars. However, the study did not explore *why* these differences occur. In this paper, we describe a series of follow-up experiments to further explore and explain their results. While our results generally confirm Cleveland & McGill's ranking of various bar chart configurations, we provide additional insight into the bar chart reading task and the sources of participants' errors. We use our results to propose new hypotheses on the perception of bar charts.

Index Terms—Graphical perception, bar charts.

1 Introduction

In their classic graphical perception paper, Cleveland & McGill [1] studied how accurately people can estimate relative heights in different bar chart designs. They used these results to rank the designs; for example, suggesting that aligned bar charts should be preferred to stacked bar charts since the former allow more accurate estimates. Their work introduced a quantitative and experimentally-grounded approach for choosing between alternative visualization designs.

In this paper, we describe a series of four follow-up experiments to further explore and understand Cleveland & McGill's results. Our primary goal is to understand *how* different bar chart designs impact accuracy. In particular, we find:

- In simple bar charts, comparisons between non-adjacent bars are difficult due to the separation between them. Separation makes comparison of short bars particularly difficult. Intervening distractor bars may also increase difficulty, but our estimates suggest that this effect is smaller.
- In stacked bar charts, distractors substantially increase difficulty, perhaps because they reduce the visual saliency of the lengths to be compared. Also, in Cleveland & McGill's study, bars were marked with a small dot. Surprisingly, our results suggest that the placement of this dot confounds their results.
- Comparisons between adjacent bars in the same stack have much higher error than non-adjacent comparisons. We speculate that this is due to a bias towards making part-of-whole comparisons.
- In aligned bar comparisons, responses vary widely across subjects. In general, our results suggest that people are much better at 50% comparisons than other ratios, that multiples of 5 and 10 are more common responses than simple fractions, and that increased response time correlates weakly with higher accuracy.

The next section summarizes the previous work on the graphical perception of bar charts. This is followed by a description and discussion of the four experiments we conducted. Finally, we draw conclusion from the studies and outline future work.

- *Justin Talbot is with Tableau Research. E-mail: jtalbot@tableausoftware.com.*
- *Vidya Setlur is with Tableau Research. E-mail: vsetlur@tableausoftware.com.*
- *Anushka Anand is with Tableau Research. E-mail: aanand@tableausoftware.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication 9 Nov. 2014; date of current version 31 Jul. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Related Work

The Cleveland & McGill study that forms the basis for this paper is described in the following section. Then we discuss two other pieces of closely related work—an approximate replication of the original study done by Heer & Bostock [6] and a paper by Zacks *et al.* [12], which evaluates the same task on somewhat different bar chart stimuli, including 3D charts. Finally, we discuss selected related work in graphical perception.

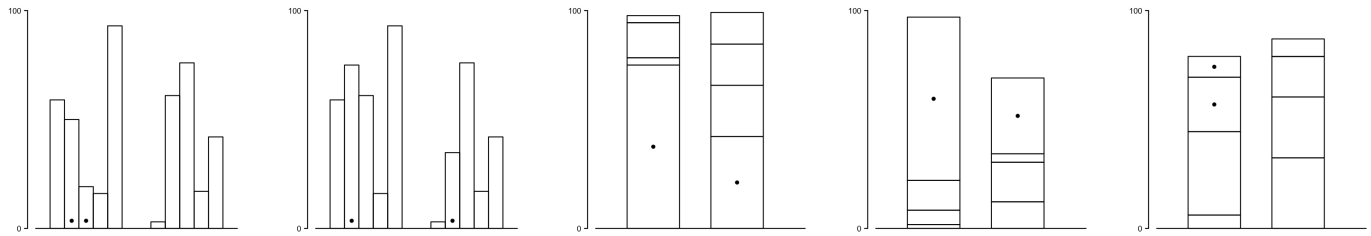
2.1 Cleveland & McGill

Cleveland & McGill's bar chart experiment studied how well participants could estimate the ratio of the lengths of two bars [1]. Participants were shown bar charts in five different configurations (Figure 1) and were asked to estimate the height of the shorter marked bar as a percent of the height of the taller marked bar (the "reference bar"). The first two conditions test comparisons of adjacent and separated bars in simple bar charts. The third and fourth conditions test aligned and unaligned comparisons in stacked bar charts. The fifth condition tests comparisons across divisions in a single stacked bar.

Fifty-five subjects were shown 10 variants of each of the five chart types. The 10 variants were made using 7 distinct true percents, 3 of which were used twice with different absolute heights—17.8%, 26.1%, 38.3%, 46.4% (twice), 56.2%, 68.2% (twice), and 82.5% (twice). Cleveland & McGill's description of the design does not specify the absolute heights of the judged bars. The heights of the distractor bars (those not marked for comparison) were chosen at random.

Fifty-one of the subjects provided usable results. For each chart type, Cleveland & McGill computed the average log absolute error (the difference between a participant's response and the true percent). This measure was used to rank the chart types, from Type 1 (lowest error) through Type 5 (highest error), as shown in Figure 1. For all chart types, they found that the average log absolute error was highest for true percents around 60%–80% and fell off for lower and higher percents.

Cleveland & McGill noted that the three designs (Adjacent Bars, Separated Bars, and Aligned Stacked Bars) in which the compared bars are aligned along a common baseline scored substantially better than the two designs (Unaligned Stacked Bars and Divided Bar) where the comparisons were unaligned. They hypothesized that this difference results from the use of two different visual estimation strategies—for aligned bars, viewers make visual comparisons of positions, while for unaligned bars, viewers must make less accurate visual comparisons of lengths. In a later paper, Cleveland and McGill [2] looked at the same estimation task on much simpler stimuli consisting of just points and lines, rather than complete bar charts. In these somewhat more artificial conditions, they confirmed that position judgments were more accurate than length judgments.



(a) Adjacent Bars (Type 1) (b) Separated Bars (Type 3) (c) Aligned Stacked Bars (Type 2) (d) Unaligned Stacked Bars (Type 4) (e) Divided Bar (Type 5)

Fig. 1: The five bar chart tasks studied in Cleveland & McGill [1]. Study participants were asked to estimate the height of the shorter marked bar as a percent of the taller marked bar. Cleveland & McGill's ranked these tasks from lowest error (Type 1) to highest error (Type 5).

2.2 Heer & Bostock

Recently, Heer & Bostock [6] performed an approximate replication of the first Cleveland & McGill study while examining the crowdsourcing of perceptual experiments. Their experimental design was similar to the original experiment, but they used Amazon Mechanical Turk (`https://www.mturk.com`) participants and they studied a somewhat different set of true percents (18%, 22%, 26%, 32%, 39%, 47%, 55%, 56%, 69%, and 85%). In addition to bar charts, they also evaluated pie charts, bubble charts, and tree maps.

They confirmed the relative rankings of the same 5 comparison tasks, but their absolute accuracy results are somewhat better than in the original study. While some of this difference may be due to the populations studied, the study design may also contribute. Cleveland & McGill's true percents were rounded to tenths, but Heer & Bostock's were whole numbers. If we make the plausible assumption that participants in both studies largely responded with whole numbers, then Heer & Bostock's results would appear to have lower error. This difference might appear to be small, but both studies use a log-transformed error metric which exaggerates small differences.

2.3 Zacks *et al.*

In separate work focused on understanding the effect of 3D perspective bars on graphical perception, Zacks *et al.* ran a similar experiment which focused on percent estimation between pairs of axis-aligned bars without extraneous distractor bars [12]. They tested 15 different true percents (1%, 8%, 15%, 22%, 29%, 36%, 43%, 50%, 57%, 64%, 71%, 78%, 85%, 92%, and 99%), which provides much broader coverage of the percent space than the previous two studies. Since they were trying to understand 3D bar perception, their work focused on varying the presentation of the bars, rather than on the configuration of the bar chart as a whole. Their stimuli included simple lines, rectangles, 3D bars, and somewhat abstract geometric shapes.

In contrast to Cleveland & McGill's finding that the error function has a maximum around 60%–80%, Zacks *et al.* found that the error function is symmetric around 50% with a local minimum at 50%.

2.4 Other graphical perception work on bar charts

Simkin and Hastie develop an explanatory model of visual comparisons in bar charts [10]. In their model, proportion judgments begin with *anchoring*, in which the two bar charts are jointly segmented to allow for comparison, followed by a scanning step which gives the final estimate. They provide some experimental results supporting their model. Elzer *et al.* present a mental processing model for bar chart reading [5]. Their preliminary eye tracking studies show that comparison of non-adjacent bars requires more saccades than comparison of adjacent bars. Newman and Scholl [8] show that bars representing the means of a sample create a false "within-the-bar bias" that incorrectly suggests that values within the bar are more likely than values outside the bar.

3 Experiments

To better understand Cleveland & McGill's results, we ran four experiments to clarify *how* bar chart design impacts accuracy:

- Experiment 1 compares the *Adjacent* and *Separated Bars* tasks to disentangle the effects of visual separation and intervening distractors on error.
- Experiment 2 compares the *Aligned* and *Unaligned Stacked Bars* tasks to isolate the effects of unaligned comparison, distractors, and dot position.
- Experiment 3 compares variants of the *Divided Bar* task to understand why comparisons of bars in the same stack are more difficult than in different stacks.
- Experiment 4 examines the discrepancy between Cleveland & McGill's and Zacks *et al.*'s error funtions and explores how participants make estimates.

We use a consistent analysis procedure across all four experiments. First, to deal with outliers, we use the same robust aggregation procedure as in Cleveland & McGill and Heer & Bostock. We compute 25% trimmed means (the "midmeans") for each experimental condition and then average across them to get marginal and grand means. Since trimmed means discard outliers, our estimates should be interpreted as the mean of "typical" responses, not of the entire distribution. The choice of 25% trimming was made *a priori* to match the previous work. We also ran our analysis with a more conservative trim of 15% and verified that our results are robust to this choice.

Second, rather than rely on null hypothesis significance testing, which has been challenged on numerous grounds [9, 7], we instead report estimates of simple effect sizes and associated confidence intervals (CIs) computed via bootstrapping [4]. This method of analysis has been advanced in psychology [3] to address the shortcomings of significance testing and we adopt it here for similar reasons.

By reporting effect size—the change in an outcome of interest, usually response error, across experimental conditions—we can situate our results within the previous work and we can make judgements about the practical significance of the effects we find. By reporting CIs we can communicate the uncertainty in our results. We consistently provide 95% CIs in square brackets after our effect size estimates. CIs that are narrow compared to the estimated effect size imply that we have strong evidence for the size, while those that are wide imply that the size is uncertain. CIs that do not include 0 indicate that we have strong evidence for the sign of the effect and those that include 0 indicate that we have weak or no evidence for the sign.

One challenge is that, since we use a within subjects design in each experiment, our CIs are necessarily correlated, which means that they cannot be directly compared to each other. To emphasize this limitation, we plot each CI in its own frame. When we want to compare two values, we directly estimate the *difference* between them.

3.1 Experiment 1: *Adjacent Bars* vs. *Separated Bars*

Both Cleveland & McGill and Heer & Bostock found that the absolute error (|subject response – true percent|) for comparisons between *Separated Bars* (Figure 1b) was higher than between *Adjacent Bars* (Figure 1a) by roughly 0.6–1 percentage points[1]. In this experiment, our goal is to better understand this *Separation Effect*. In particular, we are interested in three questions:

1. *Why are Separated Bars more difficult to compare than Adjacent Bars?* Is this *Separation Effect* due to (a) the visual separation between the bars, (b) the presence of distractors, or (c) some combination of these two?

2. *If distractors contribute to the Separation Effect, how do they do so?* Does simply adding visual clutter decrease accuracy, or do the distractors have to interfere with the visual comparison task by, e.g., making it harder to mentally draw a line connecting the tops of the two compared bars?

3. *Does the height of the reference bar (the taller compared bar) impact the Separation Effect?* And does it interact with the height of the distractors?

3.1.1 Method

The bar chart variants studied in Experiment 1 are shown in Figure 2. To address our first two questions, we include factors for separation and distractors. In the top row, we show adjacent and separated bar comparisons without distractors. In this case, any increased difficulty in making the right comparison can only arise from the increased distance between the bars (195 pixels). In the second row, we introduce short distractors, which are shorter than almost all the compared bars. While these distractors add to the visual clutter of the plot, they do not interfere with visually comparing the tops of the two bars. Any increased difficulty here will result from a combination of the effects of separation and the intervening distractor bars. In the bottom row, we add tall distractors. These distractors visually interfere with making a comparison between the tops of the two marked bars. Again, any increased difficulty here will arise from a combination of separation and the distractors. The heights of the distractor bars are fixed to avoid possible confounding.

To address our third question, we also include three conditions for the height of the reference bar—125, 250, and 375 pixels—and tested all other factors at all three heights. In the original study, the reference bar height depended on the true percent. If reference bar height has a strong effect on accuracy, this dependence on the true percent may confound the results. We revisit this potential issue with the Cleveland & McGill study in Experiment 4.

Since Cleveland & McGill found a clear effect of true percent in some of their tasks, we use the same 7 true percents (17.8%, 26.1%, 38.3%, 46.4%, 56.2%, 68.2%, and 82.5%) to ensure that our results are comparable. In our analysis, we average over this factor since we are interested in the overall accuracy. In Experiment 4, we explore the effect of true percent in detail.

These choices of factors result in a total of 126 conditions (2 separations $\times$ 3 distractor variants $\times$ 3 reference bar heights $\times$ 7 true percents). We used a within subjects design where all subjects saw all 126 conditions.

In contrast to the Cleveland & McGill design, we did not randomize whether the reference comparison bar appeared on the left or right. In Zacks *et al.*, they found a small effect for the side of the reference bar, which they hypothesized may be due to reading order in their population. To avoid possible confounding due to this effect, we fixed the reference bar to be on the left side. The Cleveland & McGill design also included small letters beneath the x-axis, labeling the two bar groups. Since these letters were not essential to the task we omitted them.

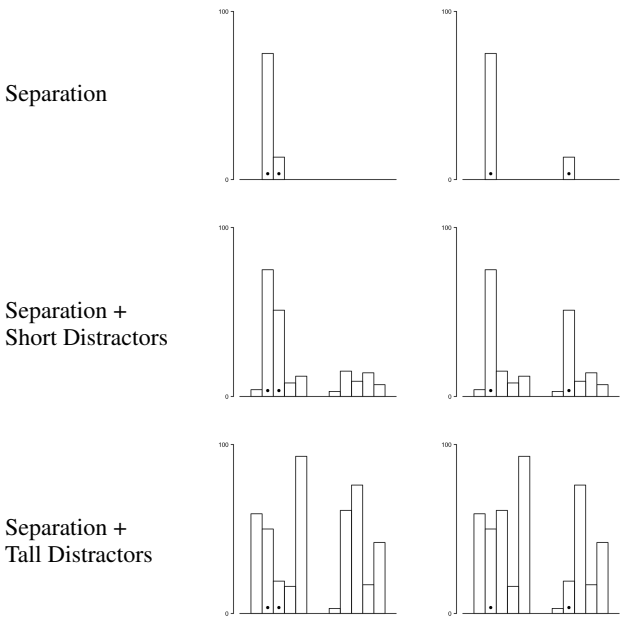


Fig. 2: Bar chart variants used in Experiment 1, which contrasts adjacent bar comparisons (left) with separated bar comparisons (right). To tease apart the effects of separation and distractors we include three distractor conditions: (top) no distractors, to evaluate the impact of separation alone, (middle) short distractors, which add visual clutter, and (bottom) tall distractors, which may visually interfere with the comparison task.

Fifty Amazon Mechanical Turk users were recruited to participate in this study. Only users with a 95% or higher Human Intelligence Task (HIT) acceptance rating were allowed to participate. The study requirements indicated that users should be between 18 and 65 years old, and would need to be comfortable with English instructions and with using an online interface to perform relative judgments. Participants were first shown an instruction page that specified the task—to make a quick visual judgement of what percent the smaller marked bar is of the larger one. They were instructed that answers should range between 0% and 100%, and that comparisons should be made unaided by fingers, rulers, or other external tools. They were told to target an average of about 7–9 seconds per response. Three example plots were provided with the corresponding true percents—70%, 59%, and 23.2%. Examples with various amounts of rounding were selected to suggest to participants that they could use as much precision as they desired. After choosing to participate in the study on the Amazon Turk site, we redirected participants to our own website where we could measure response time, and ensure that a single user completed all 126 tasks. In contrast, Amazon Turk's default website allows multiple users to collaborate in completing a replication making within subject designs difficult. Participants were paid $0.02 per comparison, or $8 an hour at 9 seconds per response. The HIT, once accepted by a participant, was set to expire in 60 minutes.

We validated that subjects understood the task by looking at the correlation between subject responses and the true values. In our first run of 50 subjects, 46 had correlations higher than 0.8, while 4 had correlations ranging from -0.4 to 0.5. We considered these extremely low correlations strong evidence that these four subjects did not understand the task. We rejected their responses and recruited four additional subjects. The replacement subjects all had correlations greater than 0.8.

[1]Reanalysis of Heer & Bostock's data [6] gives an estimate of 0.6 percentage points. We do not have access to Cleveland & McGill's original data, but rough approximation based on Figure 16 in their first paper [1] suggests an estimate of around 1 percentage point. (Cleveland & McGill report the mean log absolute error, which we can only approximately invert to get the mean absolute error.)

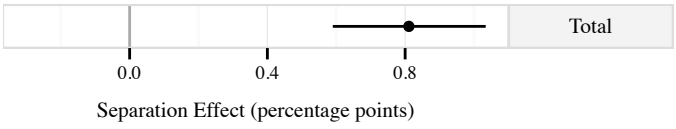


Fig. 3: Estimate of the total *Separation Effect* from Experiment 1, including the effect of both separation and distractors. The point estimate, 0.81, is in line with estimates from previous work.

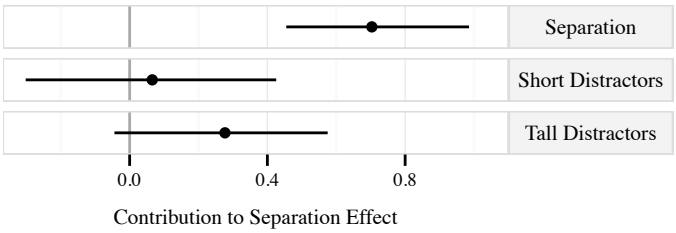


Fig. 4: Estimates of the contributions of separation and distractors to the *Separation Effect*. Separation between the comparison bars makes the percentage estimation task more difficult. The impact of distractors, whether short or tall, is estimated to be smaller; however, our CIs are relatively wide.

3.1.2 Results

In Figure 3, we show the total *Separation Effect*, including the increased difficulty due to both distractors and the separation between the comparison bars. Our estimate of 0.81 [0.59, 1.03] percentage points is between the previous estimates due to Cleveland & McGill and Heer & Bostock. We confirm that comparison of separated bars is more difficult than comparison of adjacent bars.

Our first experimental question asks whether this increased difficulty arises from the separation between the bars or from the presence of intervening distractor bars. In Figure 4, we estimate how much each contributes individually to the total effect. The contribution of separation is estimated by looking at the separation-only variants shown in the top row of Figure 1. These conditions have no distractors, so any difference in error must come from the increased separation between the bars. Our estimate of this error is 0.70 [0.46, 0.99] percentage points. The contributions of the short and tall distractors are computed using a difference-in-difference approach. Comparing the conditions in the second or third rows of Figure 1 will result in an estimate that includes the impact of both distractors and separation. To get the impact of distractors alone, we subtract out the separation estimate from the first row. The effect of short distractors is estimated to be very small, 0.07 [-0.30, 0.43], but with a wide CI. The estimated effect of the tall distractors is larger, 0.28 [-0.04, 0.58]. Pairwise comparison of the effect of separation with the effects of distractors indicates that it has a larger impact than short distractors by 0.72 [0.18, 1.29] percentage points and a larger impact than tall distractors by 0.33 [-0.12, 0.83] percentage points, though the CI covers 0 so the evidence is weaker.

Our second question asks how distractors impact the *Separation Effect*; in particular, do tall distractors add more difficulty than short distractors? Pairwise comparison of the effect of short and tall distractors indicates that tall distractors add 0.10 [-0.29, 0.45] percentage points of error over short distractors. Given the small estimate and the breadth of the CI, there is little evidence that tall distractors make separated comparisons more difficult than short distractors.

Our third question is about the effect of the reference bar height. In Figure 5 we show estimates of the *Separation Effect* conditioned on the height of the reference bar. Regardless of the height, separated bar comparisons are more difficult than adjacent bar comparisons. However, for short reference bars (125 pixels), our estimate of the effect is much larger. Pairwise comparison of these three conditional effects confirms that comparison against a short bar is substantially more difficult than comparison against a medium bar, adding 0.74 [0.22, 1.29] percentage points of additional error, or against a tall bar, adding 0.88 [0.41, 1.4] points. Pairwise comparison provides only weak evidence that comparison against a medium bar is harder than against a tall bar, adding 0.24 [-0.25, 0.68] percentage points of error. We also looked for interactions between the reference heights and the heights of the distractors; but we did not find any clear patterns in our data.

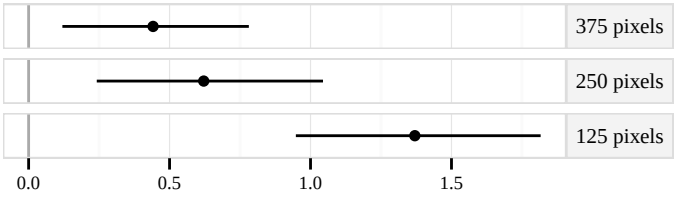


Fig. 5: Estimates of the *Separation Effect* conditioned on the height of the reference bar. Regardless of the height, separated bar comparisons are more difficult than adjacent comparisons. However, for short reference bars (125 pixels), our estimate of the effect is much larger.

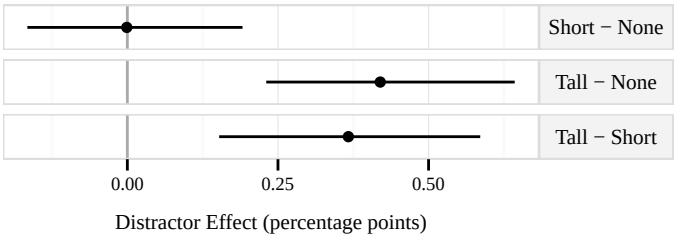


Fig. 6: From post hoc analysis, estimates of the effect of distractors averaged over both adjacent and separated bar tasks. The top row shows that short distractors are estimated to have little impact compared to no distractors. In contrast, tall distractors substantially increase the error over no distractors and short distractors.

Finally, while our data does not clearly show whether distractors differentially increase the difficulty of separated bar comparisons over adjacent comparisons, post hoc analysis of our data does show that tall distractors increase the difficulty of *both* comparison tasks. Figure 6 shows estimates of the effect of distractors averaged across both adjacent and separated bar comparisons. In the top row, we see that short distractors have no clear impact over no distractors, 0 [-0.17, 0.19]. But in the bottom two rows we see that tall distractors increase the error over both no distractors (0.42 [0.23, 0.64] percentage points) and short distractors (0.37 [0.15, 0.59] percentage points).

Discussion Our results from Experiment 1 show that separating bars in space makes comparison of their heights more difficult. In contrast, the effect of distractors was more ambiguous. While our point estimates were consistent with distractors not substantially increasing the difficulty of separated comparisons, our CIs were relatively wide. We conclude that the *Separation Effect* results from either the separation between the bars alone, or from some combination of separation and distractors. Since we did not find a clear difference between short and tall distractors, we cannot determine whether distractors increase error by adding visual clutter or by interfering with the visual task.

Our results also show that comparisons against small reference bars are particularly difficult when separated. A clearer understanding of this interaction between separation and height may give insight into the visual mechanism by which these comparisons are being made. An obvious follow-up experiment would vary the reference bar heights and the separation distance more finely than we did here. Also, given the impact of distance, particularly for small reference bars, interaction techniques that bring distant bars closer together, such as sorting, drawing reference lines, or windowing [11] are likely to improve the readability of bar charts.

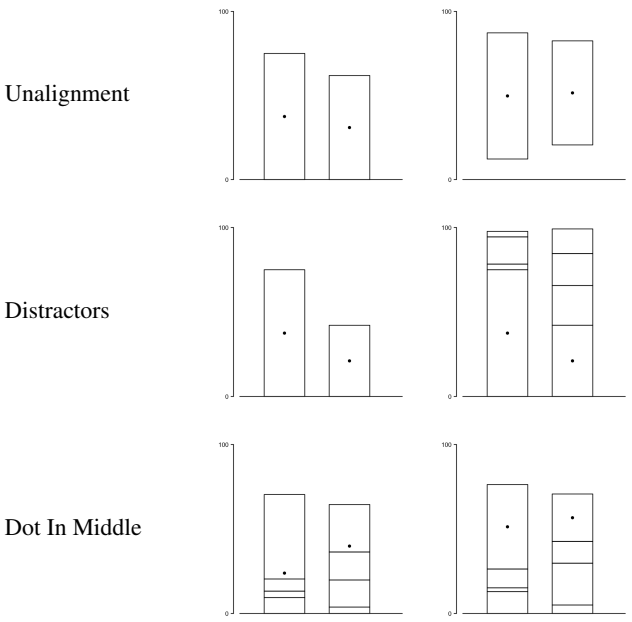


Fig. 7: Factors studied in Experiment 2. The top pair contrasts aligned and unaligned bar comparisons. The second pair contrasts comparisons with and without distractors. The bottom pair contrasts placing the marking dot at the bottom or in the middle of the bar.

We found in post hoc analysis of our data that short distractors do not increase the difficulty of adjacent or separated bar comparisons. This suggests that subjects are relatively robust to at least some kinds of visual noise. However, we did find that tall distractors increased the difficulty of *both* adjacent and separated bar comparisons. That this effect occurred for both comparisons suggests that the increased difficulty is not due to visual interference when comparing the tops of the bars. However, more study will be necessary to determine the mechanism by which tall distractors interrupt the height comparison task and to understand if visual changes such as highlighting the comparison bars can effectively decrease the impact of tall distractors.

3.2 Experiment 2: *Aligned Bars* vs. *Unaligned Bars*

Our second experiment explores the difference between comparisons of *Aligned* and *Unaligned Bars* in stacked charts (Figures 1c and 1d). Cleveland & McGill found that it is substantially harder to make height comparisons between unaligned bars than between bars aligned to a common baseline. They hypothesize that this is due to position comparison being a fundamentally easier visual task than length comparison. The Heer & Bostock estimate of this *Unalignment Effect* is 1.2 percentage points; Cleveland & McGill's estimate is roughly 2.0 percentage points. In this experiment, we want to confirm this result while exploring three questions not considered in the previous work:

1. *What is the source of the Unalignment Effect?* Does it arise from unalignment only or do distractors also play a role? This could happen if, for example, distractors increase the difficulty of length comparisons more than they increase the difficulty of position comparisons.

2. *Does the location of the marking dot make unaligned bar comparisons more difficult?* In simple bar charts, Cleveland & McGill mark their bars with a dot at the bottom, but in stacked bar charts, they mark the bars in the middle. Placing the dot in the middle provides a convenient 50% reference point, which may be more useful when the bars are aligned than when they are unaligned.

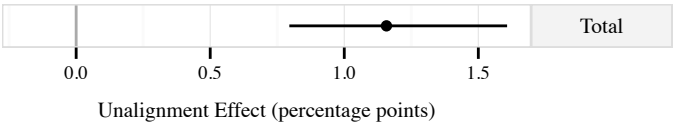


Fig. 8: The estimate of the *Unalignment Effect* and corresponding CI from Experiment 2. The point estimate of 1.15 is marginally smaller than estimates in previous work.

Fig. 9: Estimates of the contribution of unalignment, distractors, and placing the marking dot in the middle of the bar to the *Unalignment Effect*. Both unalignment itself and the distractors in a stacked bar arrangement make unaligned bar comparisons difficult.

3. *Does the height of the reference bar impact the Unalignment Effect?* In Experiment 1, we found that the height of the reference bar substantially impacted the accuracy of responses for separated comparisons in simple bar charts. Does this effect also occur for length comparisons in stacked bar charts?

3.2.1 Method

Some of the bar chart variants considered in Experiment 2 are shown in Figure 7. In the first row, we compare aligned and unaligned bars. In the second row, we vary the presence of distractors. With stacked bars it is difficult to simultaneously control the heights of all the bars, so we randomly generate the distractor heights. We randomly vary the overall heights of the stacks to ensure that the unaligned bars do not align at the top of the chart. In the third row, we also vary the position of the mark, either at the bottom, as in Cleveland & McGill's simple bar charts, or in the middle of the bar, as in their stacked charts. We use the same 3 reference heights and 7 true percents as in the previous experiment. This results in 168 conditions (2 alignments $\times$ 2 distractors $\times$ 2 mark positions $\times$ 3 reference bar heights $\times$ 7 true percentages).

We again used a within subjects design with 50 Mechanical Turk users. The criteria for participation and the experimental set up were the same as in the previous experiment. We evaluated the subjects' task understanding by again looking at the correlation between their responses and the true values. In this experiment, all 50 subjects had correlations higher than 0.8 and all were accepted.

3.2.2 Results

Our estimate of the *Unalignment Effect* is 1.16 [0.80, 1.61] percentage points (Figure 8). To be comparable to the previous work, this estimate only includes our conditions with distractors and with the marking dot in the middle of the bar. Our estimate is marginally lower than Heer & Bostock's estimate of 1.2. We can confirm that unaligned bar comparisons are more difficult than aligned comparisons in stacked bar charts.

In Figure 9, we plot the impact of unalignment, distractors, and placing the marking dot in the middle of the bar, rather than at the bottom, on the *Unalignment Effect*. Making an unaligned bar comparison has a clear strong effect, increasing error over aligned bar comparisons by an estimated 0.84 [0.59, 1.1] percentage points. As in the first experiment, we use a difference-in-difference approach to examine the effect of distractors and dot position. In contrast to the first experiment, distractors have a larger and more robust contribution to the *Unaligned Effect*, adding an estimated 0.50 [0.14, 0.91] additional

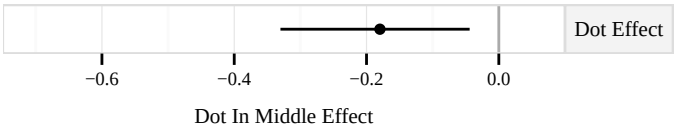


Fig. 10: Estimate of the overall effect of placing the marking dot in the middle of the bar, rather than at the bottom. That the estimate is negative indicates that placing the dot in the middle makes bar comparison tasks *easier* than placing it at the bottom.

percentage points of absolute error to unaligned bar comparisons over aligned bar comparisons. But there is little evidence that placing the dot in the middle of the bar disproportionately increases the difficulty of unaligned bar comparisons (estimated 0.12 [-0.19, 0.42]).

Unlike in Experiment 1, we found no clear evidence that the reference bar height impacts the *Unalignment Effect*. We did find some evidence that distractors increase the difficulty of aligned comparisons, but only mildly (estimated 0.2 [-0.01, 0.4] percentage points).

While the marking dot position does not have a clear impact the *Unalignment Effect*, we did find in post hoc analysis that placing the dot in the middle of the bar makes *both* aligned and unaligned comparisons *easier* (Figure 10), decreasing absolute error by 0.18 [0.04, 0.33] percentage points.

Discussion In contrast to the simple bar charts of Experiment 1, in stacked bar charts, distractors have a clear effect on error in unaligned bar comparisons. If Cleveland & McGill are right that aligned bar comparisons are made based on position and unaligned bar comparisons are made based on length, this means that stacked distractors disproportionately impact length comparisons. We speculate that one possible reason for this effect is that stacked distractors change visually salient bar corners into less visually salient T-junctions, which may make length estimation more difficult, but more study is clearly needed. A possible implication of this result is that if visualization users must make unaligned bar comparisons in a stacked bar charts, interactive visual techniques such as highlighting, which help the user visually separate the bars of interest from the surrounding distractors should help users make more accurate comparisons.

In post hoc analysis we found evidence that the positioning of the marking dot likely does have an effect on the difficulty of the task. This means that Cleveland & McGill's results for simple bar charts (with a dot at the bottom of the bar) are not directly comparable with their results for stacked bar charts (which have dots in the middle of the bar). We estimate this effect to be about a fifth of a percentage point, which, while small, is potentially enough to impact Cleveland & McGill's ranking of bar chart types. If all chart types used a marking dot at the botton, this would increase the error of all the stacked bar variants, potentially swapping the ranking of *Aligned Stacked Bar* comparisons (Figure 1c) and *Separated Bar* comparisons (Figure 1b). Since this effect was identified in post hoc analysis, a follow up study reevaluating the rank order of Cleveland & McGill's chart types with a consistent dot position or other visual mark (e.g. highlighting) is needed.

If marking the middle of the bar results in higher accuracy, this suggests an interesting interactive tool for bar charts. When visually highlighting a bar, a visualization system might automatically add subtle lines or tick marks to the bar itself at common ratios (e.g. 25%, 50%, and 75%) to aid in comparisons with that bar.

3.3 Experiment 3: *Divided Bars*

This experiment explores relative height comparisons in *Divided Bar* scenarios (Figure 1e). Cleveland & McGill find that this arrangement has the highest error of all the comparison types they considered—they estimate an overall absolute error of roughly 6.6 percentage points, while Heer & Bostock's estimate is around 4.7 percentage points. This is 0.6 (Heer & Bostock) or 1.4 (Cleveland & McGill) percentage points harder than making unaligned comparisons across different stacks.

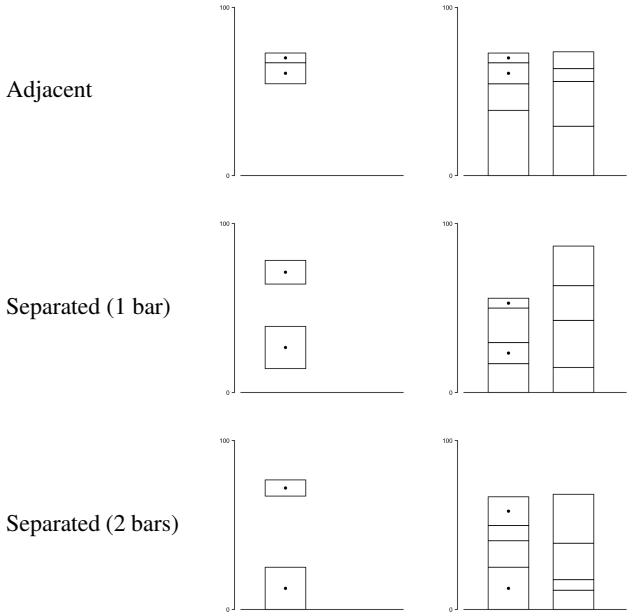


Fig. 11: Bar chart variants used in Experiment 3, without distractors on the left and with distractors on the right. In the top row the comparison bars are adjacent as in the the Cleveland & McGill design. In the middle row, the comparison bars are separated by a single bar. And, in the bottom row, they are separated by 2 bars.

Cleveland & McGill do not suggest a perceptual reason for this high error. In this experiment, we explore the question: *What contributes to the high error of Divided Bar comparisons?*

3.3.1 Method

The divided bar variants studied in this experiment are shown in Figure 11. We include conditions with and without distractors to see if they negatively impact divided bar comparisons. Also, since separation had a substantial effect in Experiment 1, we include a condition where the comparison bars are adjacent as in Cleveland & McGill's experiment, and conditions where the comparison bars are separated by 1 or 2 intervening bars.

As in the previous experiments, we control the height of the reference bar. However, since the compared bars have to be stacked on each other while still fitting within the vertical limit of the chart, we cannot use a height of 375 pixels. We instead use the heights 62.5, 125, and 250 pixels. We continue to use the same set of 7 true percents. This leads to a total of 126 conditions (2 distractors $\times$ 3 intervening bars $\times$ 3 reference bar heights $\times$ 7 true percents) which we ran with a within subjects design on 50 Mechanical Turk users. The qualifications and instructions remained the same as in the previous studies.

We again validated user understanding of the tasks by looking at the correlation of their responses with the true values. The correlations of 49 of the subjects were at least 0.71. One outlying subject had a correlation of 0.44; their responses were rejected and rerun.

3.3.2 Results

Our estimate for the absolute error of divided bar comparisons is 7.38 [6.65, 8.23] percentage points, which is higher than the estimates reported by Heer & Bostock (4.7) and by Cleveland & McGill (6.6).

In Figure 12 we show estimates for the marginal effects of distractors and separation on divided bar comparisons. Distractors add roughly 0.37 [0.05, 0.66] percentage points of additional error. This is similar to our estimate of the effect of distractors on unaligned comparisons in the previous experiment. Our estimates for the impact of separation are, perhaps surprisingly, negative, which means that subjects did better when comparing separated bars than when comparing

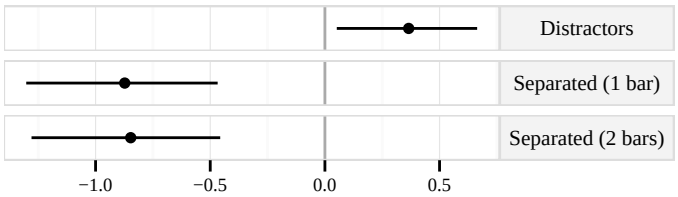


Fig. 12: Estimates of the marginal effects of distractors (compared to no distractors) and separation (compared to adjacent) when making divided bar comparisons. As in Experiment 2, stacked distractors clearly increase the error when making length comparisons. Somewhat surprisingly, separating the compared bars actually *decreases* the error.

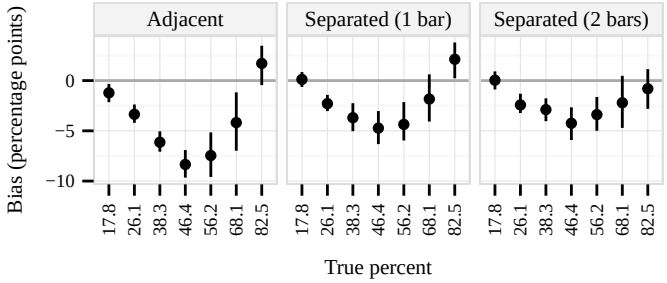


Fig. 13: Bias in Experiment 3 conditioned on separation and the true percent. The bias is large when the compared bars are vertically adjacent, but when they are separated, the bias is considerably reduced.

adjacent bars. The effect of one bar of separation is to make divided bar comparisons 0.87 [0.47, 1.3] percentage points *easier*.

Discussion While Cleveland & McGill found that comparison of bars in the same stack is substantially more difficult than other stacked bar configurations, our results suggest that this may only be true when the compared bars are immediately adjacent to each other. An intervening bar or gap reduces the average absolute error. The observed effect size in our data (0.87 percentage points) falls between the previous estimates of the difference between unaligned comparisons (Figure 1d) and divided bar comparisons (Figure 1e).

A clue to the source of the higher error of adjacent comparisons can be found in Cleveland & McGill's analysis of the bias (subject response − true percentage) in these types of comparisons. They found that, unlike the other chart types studied, divided bar comparisons had a large negative bias for middle true percents. This bias also replicated in our results for adjacent bars, but was substantially attenuated for non-adjacent ones (Figure 13). One possible explanation for this bias is that subjects are influenced by the part-to-whole comparison. For example, if we show two bars of size 60 and 30 units respectively, then the correct response to the height ratio task is $30/60 = 50\%$. However, if the bars are adjacent, subjects may instead respond closer to the part-to-whole ratio $30/(30+60) = 33\%$.

If this explanation is correct, an interesting research question is how this bias is affected by the size of the gap between the bars, or if there are other visual differences (e.g. color) which help reduce its impact. One possible design implication is that height comparisons in a stacked bar chart can be aided by the interactive introduction of gaps between adjacent bars to reduce this bias.

As in the previous experiment, the presence of distractors increases the difficulty of the divided bar task. Again, this may be related to how people make visual length comparisons. But a comparison of the distractor effect sizes found in Experiment 2 and in this experiment doesn't suggest that distractors make divided bar comparisons disproportionately harder than other unaligned stacked bar comparisons.

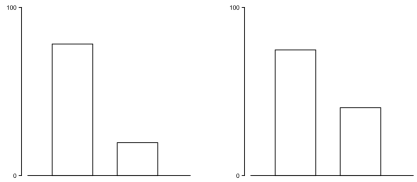


Fig. 14: Two examples of the bar chart design used in Experiment 4. The layout is similar to Cleveland & McGill's aligned stacked bar charts, but without distractors or dots marking the bars. The height of the reference bar (the taller bar) was varied slightly between tasks.

3.4 Experiment 4: *Effects of True Percent and Rounding*

Like the Cleveland & McGill study, our first three experiments were largely focused on understanding how bar chart design factors influence performance on the height ratio estimation task. In this final experiment, we focus on better understanding the height estimation task itself. We are interested in two questions:

1. *How does error vary with the true percent?* Cleveland & McGill and Zacks *et al.* have proposed very different shapes for this function; can we resolve this discrepancy?

2. *What estimation strategies do participants use and which are most accurate?*

3.4.1 Method

The design of the stimuli are similar to Cleveland & McGill's aligned stacked bar charts, but without distractors (see Figure 14). This design is also very similar to that used by Zacks *et al.* Since only two bars are shown, we omitted the marking dots. This allows us to make very short bars without having to reserve room to place a dot inside. The only controlled factor in this experiment is the true percent. We used all integer true percents from 1 to 100, for a total of 100 tasks (we omitted 0 since we felt that this case would likely be confusing for our participants). We randomly varied the height of the reference bar between 350–400 pixels. This variation discourages users from forming a fixed mental scale that could be reused from task to task. However, we kept the range of height variation relatively small to avoid possible confounding from the reference height effect seen in the previous experiments. The left bar was always taller.

We used a within subjects design on 50 Mechanical Turk users. The study qualifications and instructions were similar to those in the previous experiments. One subject did not finish the study and their partial responses were rejected and rerun. We validated understanding by looking at correlation between subject responses and the true values. All 50 subjects had correlations higher than 0.86; so were accepted.

3.4.2 Results

We first examine the results of Experiment 4 by true percent. In the top plot in Figure 15 we show our estimate of the absolute error as a function of true percent in blue with the pointwise 95% confidence interval in gray. Our error function has minimums near 0% (which we did not test), 50%, and 100%. The error function is roughly symmetric around 50%. For context, we also plot the results of Cleveland & McGill (orange) and of Heer & Bostock (teal), which are only approximate since we had to convert from their reported log absolute errors to absolute errors, and the results of Zacks *et al.* (brown). Our results are clearly most similar to those of Zacks *et al.* The middle plot shows the raw error as a function of the true percent. There is some evidence of overestimating for small true percents and underestimating for large true percents. The bottom plot shows response time, which is similar in shape to the absolute error.

We next look at the results of Experiment 4 by participant. In Figure 16, each panel shows the histogram of responses of the 10 most

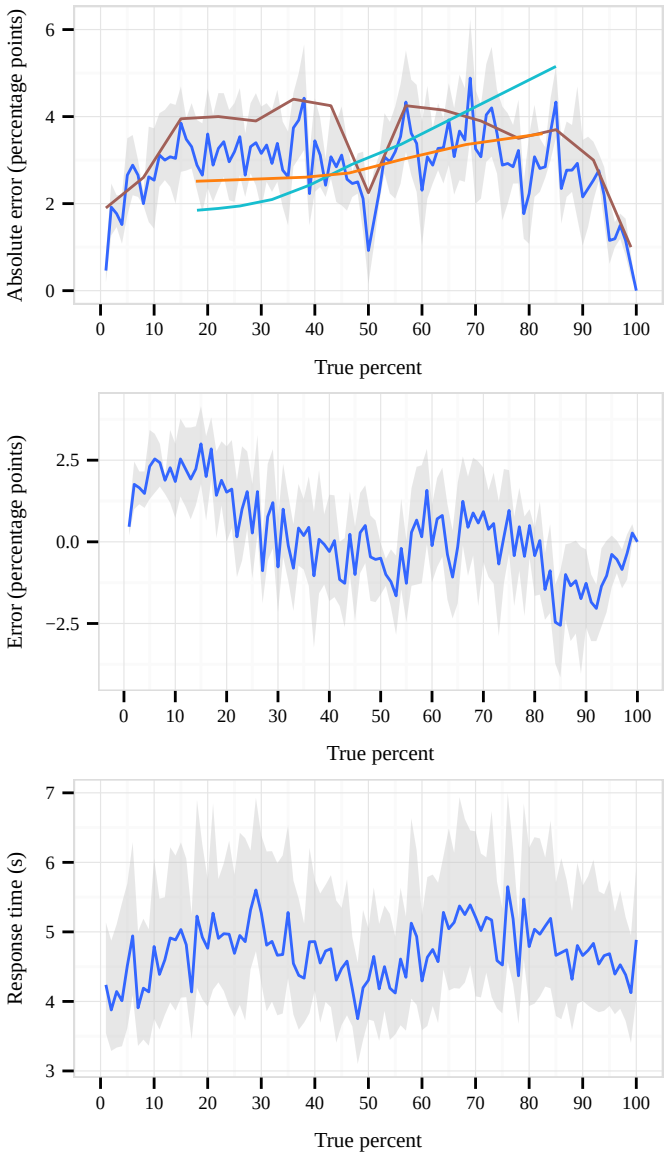


Fig. 15: (Top) Absolute error as a function of the true percent (blue, with confidence interval) compared to results from the previous work—Cleveland & McGill (orange), Heer & Bostock (teal), and Zacks *et al.* (brown). (Middle) Raw error, showing that small true percents are overestimated. (Bottom) Response time by true percent.

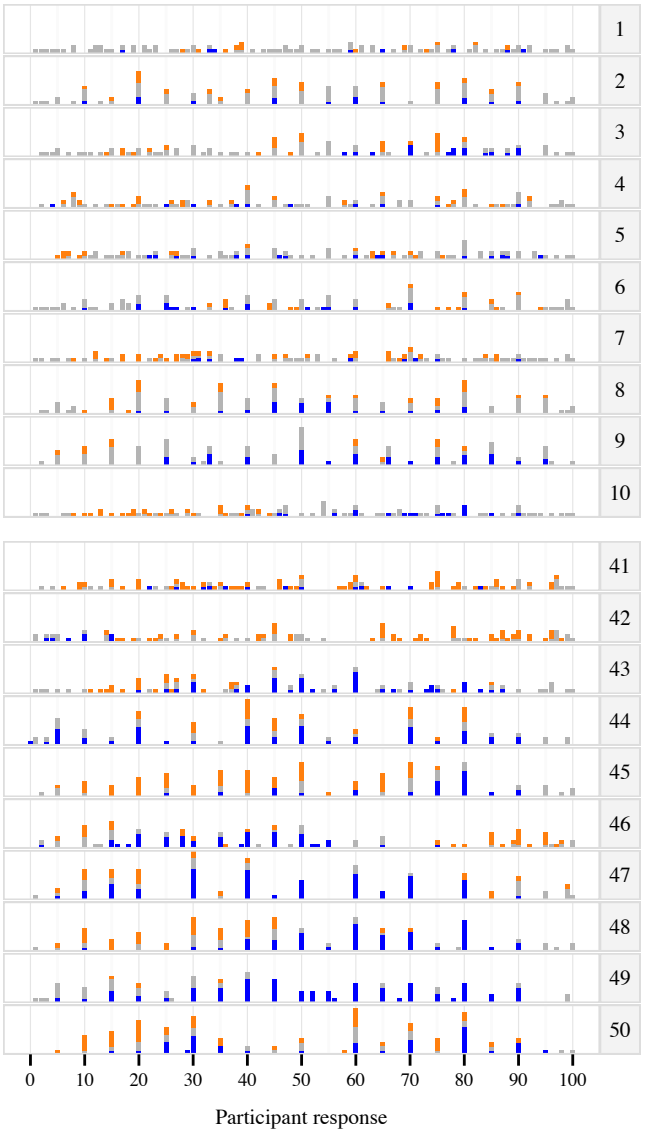


Fig. 16: Histograms of the responses for the ten most accurate (top) and ten least accurate (bottom) participants in Experiment 4. Blue indicates underestimates by at least 2.5 percentage points, orange indicates overestimates by at least 2.5 percentage points. Note the strong evidence of rounding by many, but not all participants. However, rounding is not directly predictive of overall rank.

accurate participants (numbered 1–10) and the 10 least accurate participants (numbered 41–50), as measured by their average absolute error. Orange indicates responses that overestimated the true percent by more than 2.5 percentage points, blue indicates responses that underestimated the true percent by more than 2.5 percentage points, and gray indicates "close" responses. If a participant always rounded to the nearest multiple of 5, their plot should be entirely gray. This plot reveals a considerable diversity in participant strategies. At a high level, participants can be split into those who provided "rounded" responses (e.g. participants 2 and 45) and those who provided more precise responses (e.g. participants 1 and 42). However, perhaps counterintuitively, rounding is not clearly associated with overall lower accuracy. As can be seen from the prevalence of orange and blue in the lower half, the between subject performance difference is largely caused by errors much larger than the amount of rounding that occurred.

In Figure 17, we plot the trimmed mean absolute error for each participant against their trimmed mean response time. As might be expected, the data suggests that increasing time spent on the task did lead to improved performance. However, there are a number of outliers from this trend. Also, some subjects achieved the same accuracy as others while spending less than half the time per response.

Discussion Our absolute error results are quite similar to those of Zacks *et al.*, but very different from the two other studies. The other studies included distractors, which may confound comparing results. However, a possible explanation for the discrepancy is the fact that both this experiment and that of Zacks *et al.* explicitly control for the height of the reference bar. In contrast, Cleveland & McGill and Heer & Bostock chose the height of the reference bar conditioned on the true percent. Given the strong effect of height seen in the previous experiments, it is plausible that the asymmetric error function found by Cleveland & McGill and Heer & Bostock is due to confounding of reference bar height and the true percent.

The minimum in the absolute error function at 50% is very prominent and was previously noted in Zacks *et al.* (the dip is too narrow

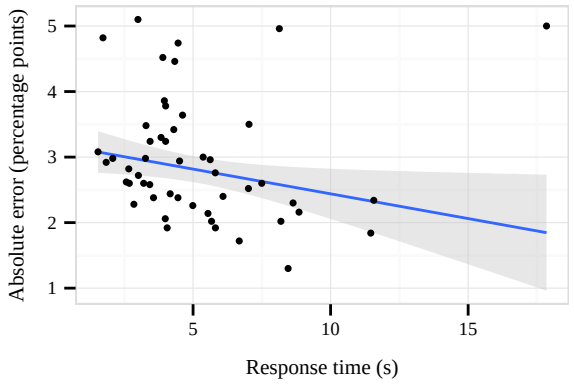


Fig. 17: The average absolute error against the average response time for each participant in Experiment 4. While there are some outliers, there is evidence of a downward trend in error for increasing response times as shown by the robust linear fit.

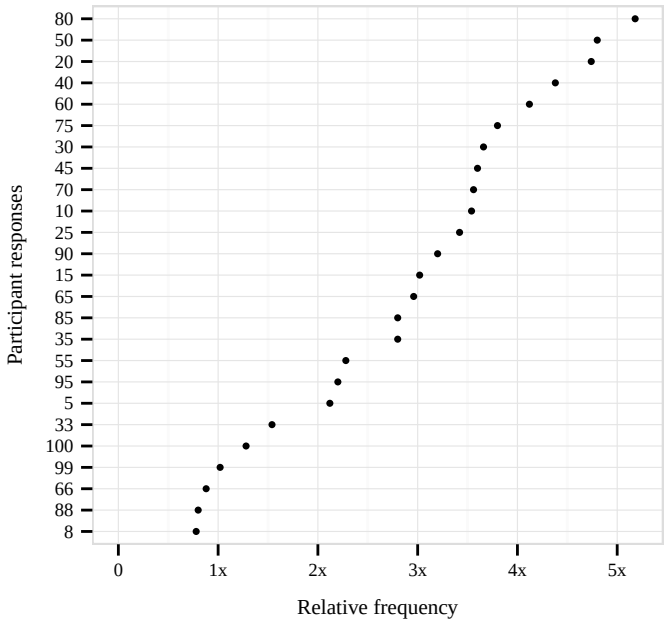


Fig. 18: The 25 most common user responses in Experiment 4 plotted against their relative frequency.

to have been identified by the coarser sampling used in Cleveland & McGill or Heer & Bostock), where they hypothesized that it resulted from rounding on the part of participants. To explore this, in Figure 18 we plot the most common participant responses in Experiment 4. Participants responded 50% nearly 5 times as often as expected in this study, indicating a strong rounding bias towards this value. However, participants responded 80% more frequently and 20% nearly as often, yet our data in Figure 15 do not show any evidence of a minimum at either point. Thus, while there is substantial rounding in our data, the minimum at 50% appears to be caused not by rounding, but by subjects actually being more accurate there.

Figure 18 also provides insight about the common estimation strategies in use. The most common responses are multiples of 10, followed by multiples of 5. This suggests that many participants are mentally dividing the reference bar into tenths followed by twentieths if necessary. Simple fractions, such as quarters (25% and 75%) and thirds (33% and 66%), do appear on this list, but less frequently than tenths.

4 Conclusion

The experiments in this paper explore variations of the bar charts originally studied by Cleveland & McGill and lead to insight into the sources of bar chart interpretation error. We found that short bars are more difficult to compare. Distractors have different effects in simple bar charts and stacked bar charts. The way bars are marked can influence accuracy. The introduction of a gap between stacked bars can prevent erroneous part-of-whole comparisons when desired. These results highlight the fact that small design changes can significantly affect how bar charts are perceived, and that even for simple visualizations, such as bar charts, we still do not have a complete understanding of what impacts chart perception.

Our experiments also raise new questions that will need to be addressed with additional studies. Future experiments will help understand the perceptual and mental strategies used in making bar height comparisons. Study designs that gather richer quantitative data, such mouse or eye movements, or more qualitative data, such as participant introspection on strategy, might provider deeper insight. Future directions might include exploring the effect of distractors in sorted bar charts, the impact of bar heights in constrained displays, and the effect of common interactions on bar charts, such as proportional brushing.

Predicting how people will perform on more complicated bar chart designs or how well people interpret bar charts "in the wild" remains difficult. However, we believe that concrete, experimentally-supported progress in understanding basic graphical perception effects is the most promising avenue towards improving visualization practice and towards a high-level theory of visualization.

Acknowledgments

The authors wish to thank Maureen Stone for substantial help in improving the presentation of this paper and the anonymous reviewers for suggesting ways to greatly improve the rigour of our analysis.

References

[1] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, September 1984.

[2] W. S. Cleveland and R. McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, Nov. 1986.

[3] G. Cumming. The new statistics: Why and how. *Psychological Science*, 25(1):7–29, 2014.

[4] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, 1996.

[5] 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*, 16(1):1–30, Mar. 2006.

[6] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Human Factors in Computing Systems (CHI)*, pages 203–212, 2010.

[7] R. B. Kline. *Beyond significance testing: Reforming data analysis methods in behavioral research*. APA Books, Washington, DC, 2004.

[8] G. E. Newman and B. J. Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, 19(4):6014–607, August 2012.

[9] F. L. Schmidt and J. E. Hunter. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors, *What if there were no significance tests?*, pages 37–68. Lawrence Erlbaum Associates, New Jersey, USA, 1997.

[10] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82(398):454–465, 1987.

[11] C. Tominski, C. Forsell, and J. Johansson. Interaction support for visual comparison inspired by natural behavior. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2719–2728, 2012.

[12] 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*, 4(2):119, 1998.