Testing statistical charts for accuracy and interpretation

Anonymous

Keywords: Graphical Perception, Data Visualization, Scientific Communication

1 Introduction

Communication of scientific results to the general public is essential. The use of visuals is a key component in scientific communication; visuals can help provide context, explain scientific concepts, highlight findings, and display patterns in data.

Decisions about the design of a data visualization are made by the author, and may be driven by subject matter conventions, branding choices, or personal style preferences. When used to communicate with the general public, choices about data visualizations should be informed by knowledge on what best supports the audience in understanding the data and conclusions correctly.

Cleveland and McGill (1984) proposed a list of basic graphical perception tasks and experimentally determined an accuracy-based ranking of those tasks. This ranking has been reproduced by others (Heer and Bostock, 2010). The tasks proposed by Cleveland and McGill involve assessments of elementary chart elements which focus solely on the *structural* components of a chart: the mapping of a statistical quantity to a particular shape, size, and position in an image. These studies do not address the impact of *aesthetics* on perception: orientation of elements in a chart, context provided, color mappings, and supporting visual elements such as borders, labels, and reference or grid lines.

Cleveland and McGill recruited colleagues and their spouses for their evaluation, while the Heer study used respondents to Amazon Mechanical Turk[1].

We extended the tasks proposed by Cleveland and McGill to complete a series of tests on graphical perception in modern data visualization, incorporating and varying both structural and aesthetic elements within the charts presented to viewers. In addition, we utilized a nationally-representative sample of U.S. adults to better understand how the general public as a whole perceives graphics, in contrast to the convenience samples used by prior studies.

2 Methods

We conducted a series of tests focusing on the public's ability to perceive differences between two values displayed in a chart. Each test asked the respondents to identify which of two elements displayed in a chart was larger. We varied the *structure* of the visual elements as well as the *aesthetics* used in the chart.

2.1 Test population

We employ AmeriSpeak's Omnibus[2] survey, which utilizes a probability-based panel and surveys a sample of 1,000 nationally representative adults 18 and older. The advantage of using probability-based panel approach is two-fold. First, we have access to a large sample of survey participants and thus have greater power in making inference about graphical perception abilities. Second, the sample is representative of the general adult public in the U.S., which is an important target audience for scientific communication.

In some rounds, respondents were split into two groups and each group received a different stimulus; each stimulus in every round was seen and responded to by at least 465 respondents.

2.2 Test stimulus

The tasks focused on determining which of two very similar values were larger. The two values were chosen close to the *just noticeable difference* (the theoretical difference for which about half of the population is able to determine the difference correctly), making this a visually hard task at the boundary of our perception (Lu et al., 2022). Our assumption is that even small changes to this task can have a large effect on a viewer's perception resulting in measurable changes of the overall accuracy.

Values were depicted as pieces of a stacked bar chart, shown in Figure 1. Responses were collected from a series of variations to structure and aesthetics. Differences in structure included whether the marked pieces were aligned along a common baseline or not, as well as the orientation of the chart as a horizontal or vertical stacked bar chart (see Figure 1a,b). Differences in aesthetics included variations in the color scheme used, use of grid lines, and removal of all relevant context (not shown).

3 Results

Each one of the stacked barcharts in Figure 1 was shown to panelists in two versions: with bars A and B aligned along a common axis and the stacked version shown in the image. The results of these evaluations are shown in Figure 2. Like Cleveland and McGill, we find that assessing the difference between aligned bars is easier than between unaligned bars (McNemar test statistic 372.15, p-value = 2.2×10^{-16}).

[1] https://www.mturk.com/

[2] https://amerispeak.norc.org/us/en/amerispeak/our-capabilities/amerispeak-omnibus.html

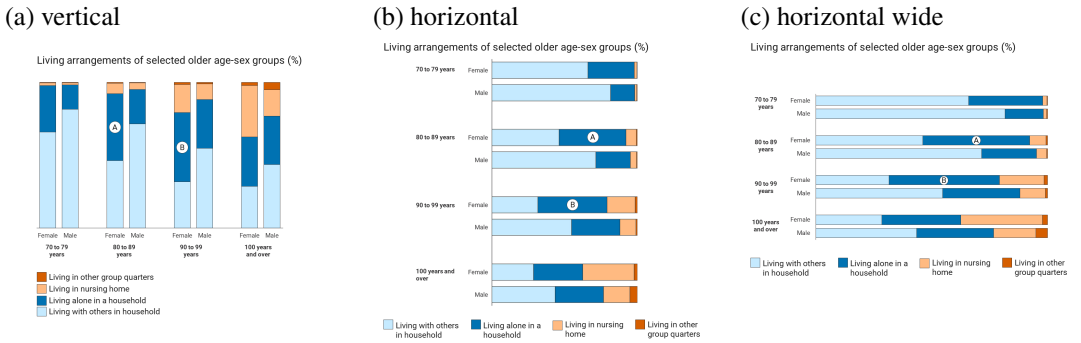

Figure 1: From left to right three stacked barcharts shown to panelists. The difference in value between A and B is the same throughout all charts (B is larger). The area of the representation is kept constant; the difference in heights/widths between bars A and B changes from 7 pixels in (a) and (b) to 11 pixels in (c).

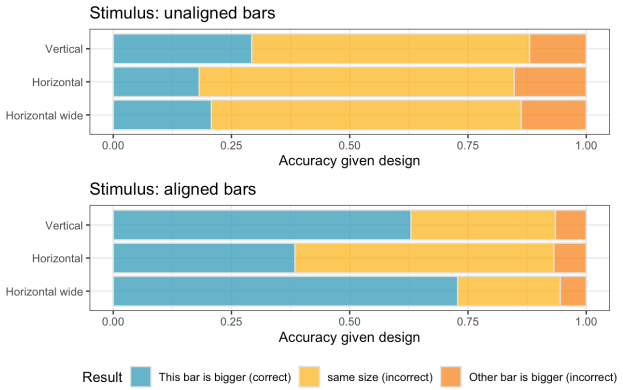

Figure 2: A large number of respondents chose the answer 'the bars are the same size'. While correct for the purposes of interpretation, technically this answer is not the most accurate. Differences between aligned bars are correctly identified at about twice the rate of stacked bars.

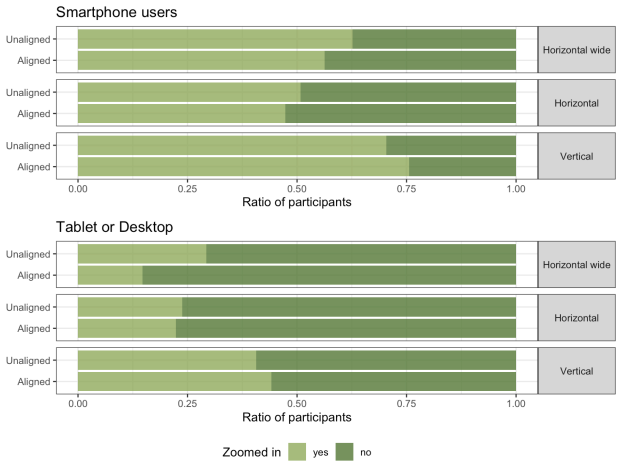

Figure 3: Devices with small screens lead to about twice the zooming rate as tablets or desktops. Panelists zoom into the vertical design, but less so into horizontal designs.

The design choices are clearly having an impact on accuracy. The change from a horizontal to a vertical design leads to a significant loss in accuracy for both aligned and stacked bars. The re-scaled design of the wide horizontal bars reclaims some of the loss for stacked bars and outperforms the vertical design by a similar margin in aligned bars. This improves performance of the wide horizontal design over the horizontal design is expected: bars are overall wider and closer together, both factors positively affect accuracy (Lu et al., 2022).

Another factor contributing to the difference in accuracy between the vertical and the horizontal is how participants interact with the different designs: generally, about half of all participants make use of the option to zoom into charts (which improves accuracy by about 5 percentage points on average). Figure 3 shows that zooming behavior of panelists changes depending on the design of the chart.

4 Conclusions/Further Work

We have shown that the format of the AmeriSpeak panel allows us to reproduce findings established in the literature. This provides us with a unique opportunity to create an experimentally validated portfolio of graphical insights: Grid lines improve accuracy in comparisons, isolating the visual task from its context in the chart seems oddly detrimental to an accurate assessment, forcing panelists to rank objects by size (rather than offering the choice of 'same') does not change the relative rate between the other choices, but increases time spent on an evaluation and reduces certainty in one's response.

References

W.S. Cleveland and R. McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*.

Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

M. Lu, J. Lanir, C. Wang, Y. Yao, W. Zhang, O. Deussen, and H. Huang. 2022. Modeling just noticeable differences in charts. *IEEE Transactions on Visualization & Computer Graphics*, 28(01):718–726, jan.