Testing Charts: viewer's perceptual accuracy in surveys

KIEGAN RICE[*1], HEIKE HOFMANN[†2], NOLA DU TOIT[‡1], EDWARD MULROW[§1], AND [2]

[1]*NATIONAL OPINION RESEARCH CENTER (NORC)*
[2]*DEPARTMENT OF STATISTICS, IOWA STATE UNIVERSITY*

Abstract

The use of visuals is a key component in scientific communication, and decisions about the design of a data visualization should be informed by what design elements best support the audience's ability to perceive and understand the components of the data visualization. We build on the foundations of Cleveland and McGill's work in graphical perception, employing a large, nationally-representative, probability-based panel of survey respondents to test perception in statistical charts. Our findings provide actionable guidance for data visualization practitioners to employ in their work.

Cleveland and McGill (1984)

This is Heike's color for making changes

Introduction

How do structural design choices in a data visualization impact viewers' ability to identify the larger of two elements?

How do aesthetic design choices in a data visualization impact viewers' ability to identify the larger of two elements?

[*]Corresponding author. Email: rice-kiegan@norc.org
[†]Email: hofmann@iastate.edu
[‡]Email: dutoit-nola@norc.org
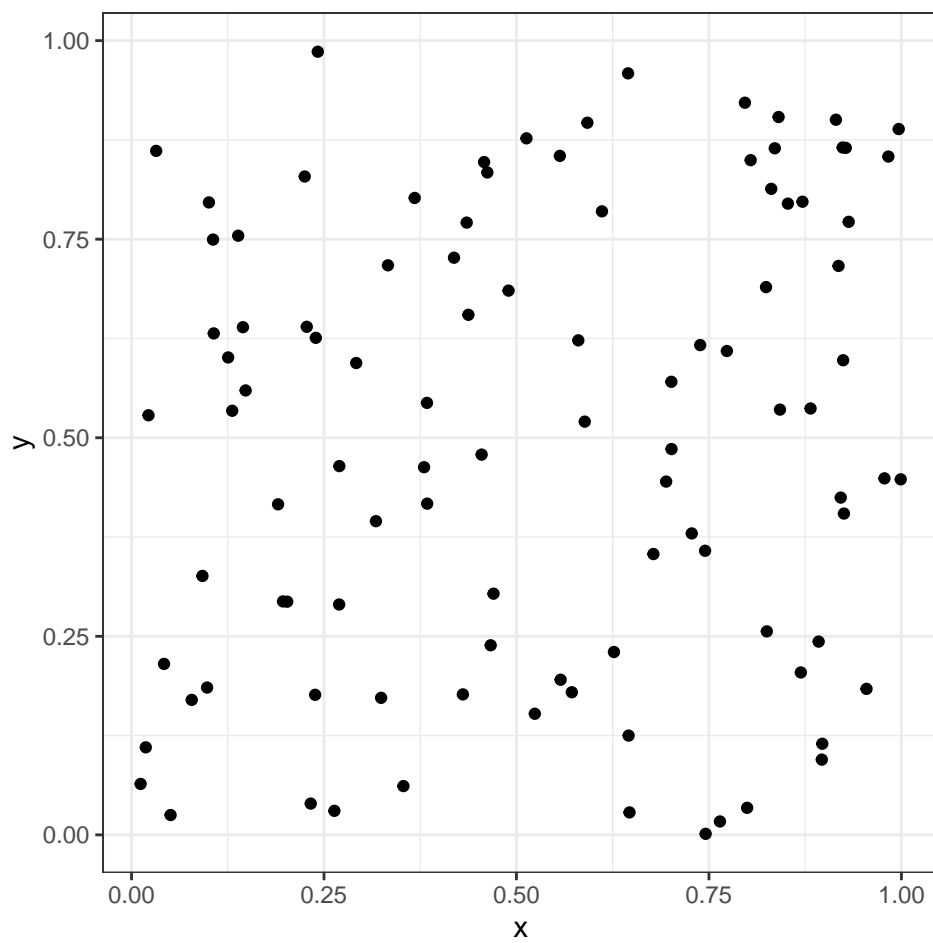[§]Email: mulrow-edward@norc.org

Figure 1: quarto figure caption.

How is viewer behavior (zooming, time spent on question, certainty of response) impacted by structural and aesthetic design choices in a data visualization?

Structural design choices:

Structural design choices

Mapping, Stacked bar, Vertical, Horizontal, Horizontal wide

Facetted bar

Only have a split sample for this

Pie, Alignment

We have this for all above mappings, but the setup is a little different for facetted bar

Aesthetic design choices (structural choices seems stronger/there could be a lot to talk about there… should we skip aesthetic on this one?)

Colors

Use of gridlines

Outcomes/responses for modeling:

Binary accuracy (correct/incorrect – 'they are the same' is incorrect here)

Ordinal response (a/b/they are the same)

Zooming behavior (zoomed/did not zoom)

Time spent on question (continuous, in seconds)

Certainty

I've noted this below, but: how to model? Ordinal response? Binary (certain or very certain vs everybody else)?

Survey setup - Stimulus description

Asking perceptual questions in a survey is different from the controlled environment of a cognitive lab, where these kind of questions would usually be addressed. This means, that instructions to participants have to be delivered in a very short and easily understandable, because questions arising from the task can not be answered. Similarly, rather than asking the same (or similar) type of question multiple hundred times, in a survey we can ask only a few questions. In order to observe any effect, we need to ask questions that are perceptually hard, which means that we need to ask questions about stimuli that are close to our perceptual threshold. The **Just-Noticeable Difference** (JND) is defined as the smallest difference that will be detected 50% of the time. For the height difference of barcharts, Lu et al. (2022) determined that the distance between bars is the main contributor to differentiating sizes. Based on their models, a tile difference of 155 pixels leads to a just noticeable difference of 3.5 pixels. Table 1 shows the two stacked barcharts shown to participants in the survey. The marked tiles in each plot are 155 pixels apart. Their heights are 205 (left) and 213 pixels (right), respectively, corresponding to about twice the JND. This difference should lead to a relatively high accuracy rate for participants and simultaneously limit the amount of frustration resulting from a task that is perceived as 'too hard'.

Both charts in Table 1 show the same data with slight modifications to the order of the levels – the first and second level in each of the bars are reversed between the left and the right chart. Participants were asked to compare the relative sizes of the tiles marked A and B (C and D, respectively) and select the correct response out of the possible choices:

1. A is bigger
2. B is bigger

3. They are the same

Answer 2 is the correct answer for both charts. Both charts are shown at the same size, i.e. in both cases the difference in size between the bars is exactly the same, the vertical distance between the bars is the exact same amount. This leaves the vertical positioning of the bars as the only difference between the charts. Any differences in observed accuracy can therefore be attributed to this difference in presentation.

Table 1: The two stacked barcharts every participant got to see. In each barchart, the two marked tiles are to be compared for their size. in both instances, the tile on the right is (very slightly) larger.
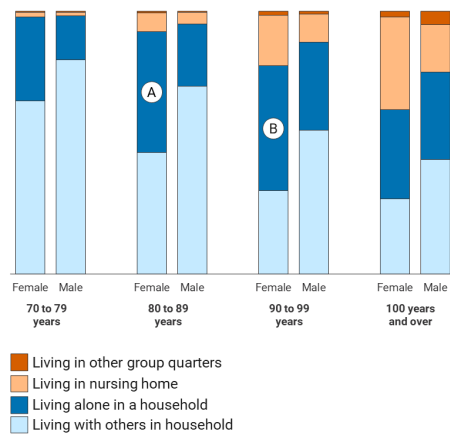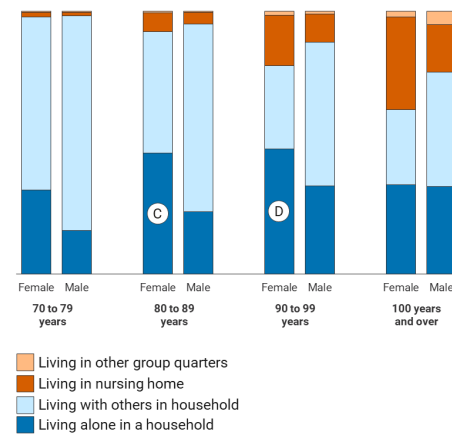

DATA THAT MAY BE INCLUDED IN ANALYSIS:

ROUNDS 1-2: Color variations on vertical stacked bar, aligned and unaligned

ROUND 3: Horizontal and horizontal wide, aligned and unaligned

ROUND 5: Horizontal wide gridlines (only dark grid split sample)

ROUND 6: Facetted bar (split sample w/o forcing choice)

ROUND 7: Aligned vs unaligned pie (full sample)

Data Analysis and Results

COMBINING SAMPLES AND WEIGHTING NOTES:

We can combine responses across samples into one combined dataset, but we need to adjust weights accordingly so that each sample is weighted equally in the model [@omuircheartaigh-CombiningSamplesVs2002].

Question for Ed: If we compare a full sample to a split sample, do we still want to weight these 'equally'?

Analysis should be done using the 'survey' package and weights should be taken into account.

ANALYSIS PLAN:

Models below structured as:

Response

Covariates to use in each model

STRUCTURAL VARIATION – START HERE

Binary accuracy across structural choices

Model 1:

Alignment only, just vertical stacked bar

Using the AmeriSpeak survey tool, a total of 1902 participants were exposed to two barcharts each, as shown in Table 1.

Model 2:

Alignment

Bar vs pie (comparable question for pie is A vs B)

Model 3:

Alignment

Vertical x horizontal x horizontal wide

Model 4:

Alignment

Every structure (vertical bar, horizontal bar, horizontal wide bar, facet bar, pie)

Visuals:

% yes across each different structural condition

Facet by aligned/unaligned?

Model estimates + CIs

Ordinal response

Model 1:

Alignment only, just vertical stacked bar

Model 2:

Alignment

Bar vs pie (comparable question for pie is A vs B)

Model 3:

Alignment

Vertical x horizontal x horizontal wide

Model 4:

Alignment

Every structure (vertical bar, horizontal bar, horizontal wide bar, facet bar, pie)

Visuals:

All responses across each different structural condition

Facet by aligned/unaligned?

Model estimates + CIs

Zooming behavior (zoomed/did not zoom)

Model 1:

Device type

Alignment

Vertical x horizontal x horizontal wide

Visuals:

% zoomed by device + alignment (already have this chart)

Model estimates + CIs

Time spent on question (in seconds)

Model 1:

Device type

Zoom

Alignment

Vertical x horizontal x horizontal wide

Model 2 (this may not be feasible for comparison depending on what level the 'TOTALTIME' is captured at):

Device type

Zoom

Alignment

Every structure (vertical bar, horizontal bar, horizontal wide bar, facet bar, pie)

Visuals:

Distribution of time spent variable

Facet by device type, zoom, structural condition, alignment? Play around with it

Average time spent by each of the conditions

Certainty?

Same models as above, but I'm not sure how we want to do the response. Ordinal response? Binary (certain or very certain vs everybody else)?

AESTHETIC VARIATION – ONLY IF TIME

Binary accuracy (correct/incorrect – 'they are the same' is incorrect here) across structural choices

Model 1:

Dark grid vs no grid (only have for horizontal wide)

Response choice (ordinal response)

Model 1:

Dark grid vs no grid (only have for horizontal wide)

Zooming behavior (zoomed/did not zoom)

Model 1:

Device type

Dark grid vs no grid

Time spent on question (in seconds)

Model 1:

Device type

Zoom

Dark grid vs no grid

Certainty?

Same models as above, but I'm not sure how we want to do the response. Ordinal response? Binary (certain or very certain vs everybody else)?

Conclusion

Supplementary Material

- **Participant Data (Linear):** Link to csv file with the data.
- **Data Analysis Code:** Link to an html document with annotated code chunks.

References

Cleveland, William S., and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–54. https://doi.org/10.1080/01621459.1984.10478080.

Lu, Min, Joel Lanir, Chufeng Wang, Yucong Yao, Wen Zhang, Oliver Deussen, and Hui Huang. 2022. "Modeling Just Noticeable Differences in Charts." *IEEE Transactions on Visualization and Computer Graphics* 28 (1, 1): 718–26. https://doi.org/10.1109/TVCG.2021.3114874.