Journal of Data Science xx (xx), 1–13 DOI: xx.xxxx/xxxx_xxxx
July 2020

# Testing Charts: viewer's perceptual accuracy in surveys

**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.

hello?

Kiegan, this is your editing color :) - you can choose a different color above

## Introduction

What do viewers see, when we show them a data chart? A crucial step in the process of 'understanding' a chart, is that a visual allows us to make comparisons between its parts. Cleveland and McGill (1984) defined for the purpose of their seminal study the better visual as the one that allows viewers to make more accurate comparisons. Based on mappings of quantitative variables to different graphical elements, Cleveland and McGill's study resulted in a ranking of perceptual tasks from most accurate to least accurate. (**mackinlayAutomatingDesignGraphical1986?**) took this approach and expanded the tasks and their order to ordinal and nominal scales, as shown in Figure 1.

What we need to remember, is that Mackinlay's work is purely theoretical. Cleveland and

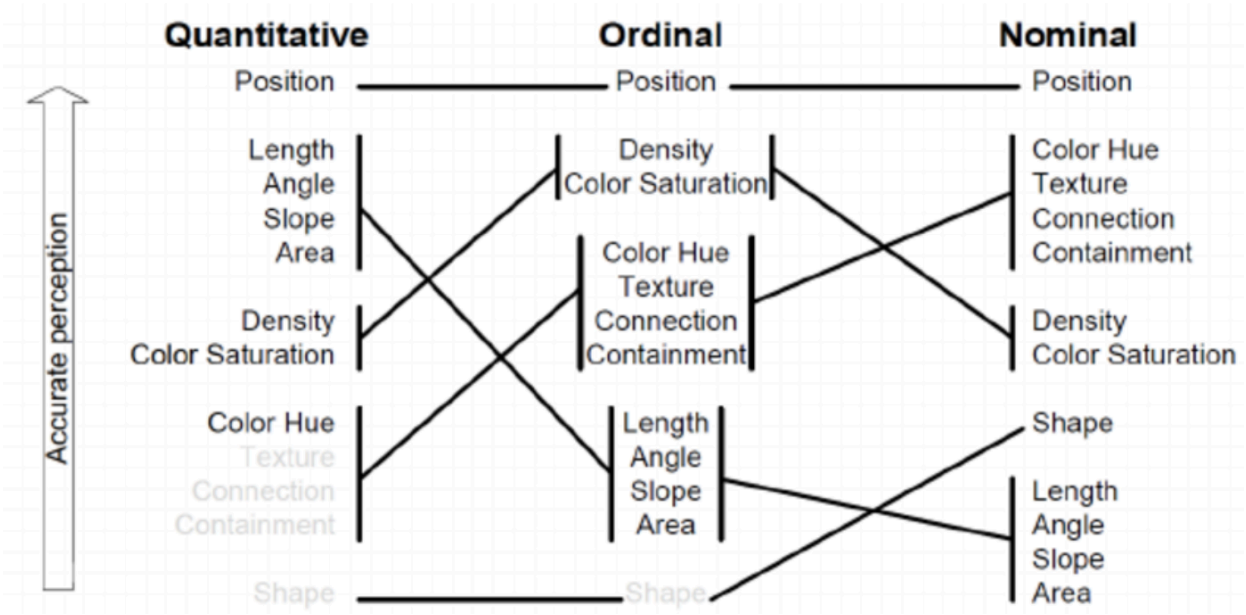



Figure 1: Ranking of perceptual tasks, as given by (**mackinlayAutomatingDesignGraphical1986?**). The ranking of tasks on the quantitative scale are empirically verified by Cleveland and McGill (1984).

McGill's rankings based on a study. However, this study is a small convenience sample, consisting of only a few individuals recruited from among the authors' coworkers and their spouses. Heer and Bostock (2010) have reproduced Cleveland and McGill's rankings using a crowd sourcing platform. A total of XXX amazon turkers were involved.

Crowd workers are known to be biased towards male, young and relatively higher education: Borgo et al. (2017)

Here, we seek to – first – answer the question whether it is possible to use a survey to reproduce (some of) the rankings.

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


(a) **Unaligned bars** (b) **Aligned bars**

Figure 2: The only difference between the two pairs of rectangles A, B and C, D is their alignment, i.e. A and C are of identical size, as are B and D. While the predominant response for the unaligned pair on the left is 'they are of the same size', more than half of the viewers respond with 'D is bigger' to the aligned pair of bars on the right.

the same (or similar) type of question with varied signals 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. We are using results from studies on barcharts and pie charts (Lu et al. 2022) to inform the differences in charts shown to survey panelists.

- 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?
- 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

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

Figure 3 shows the two stacked barcharts shown to participants in the survey. The marked tiles in each plot are 155 pixels apart. Based on Lu et al. (2022)'s model, a difference of 155 pixels leads to a just noticeable difference of 3.5 pixels. The heights of the bars 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 Figure 3 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.

DATA THAT MAY BE INCLUDED IN ANALYSIS:

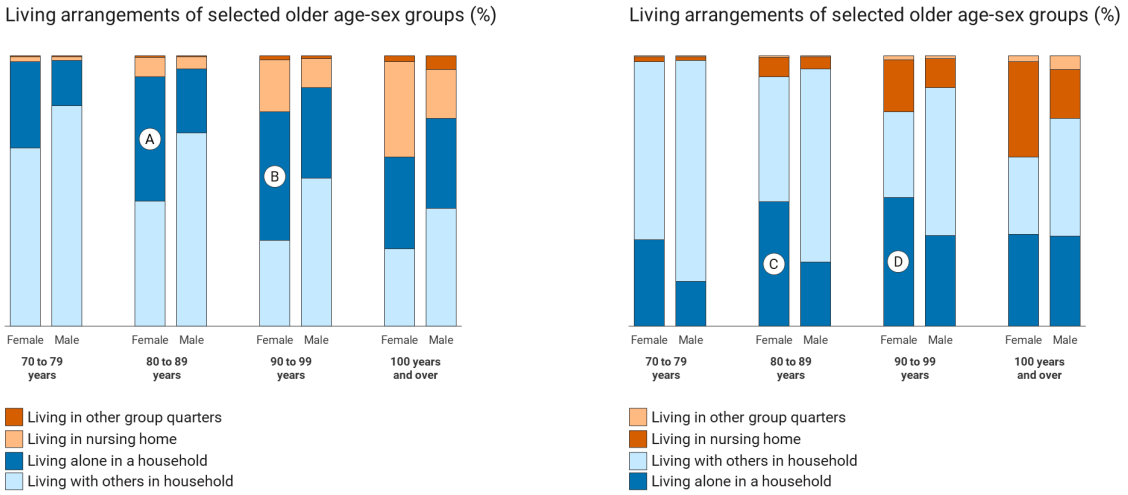
ROUNDS 1-2: Color variations on vertical stacked bar, aligned and unaligned

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)




(a) **Unaligned bars** (b) **Aligned bars**

Figure 3: 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 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 (O'Muircheartaigh and Pedlow 2002).

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.

All calculations in this paper are done in R (R Core Team 2022) using the `survey` package (Lumley 2004) version 4.0 (Lumley 2020) based on Lumley (2010).

The data used for assessing the accuracy of comparisons in **??** is collected in two rounds of the NORC Omnibus survey. Rounds 1 and 2 are combined by adjusting the weights with

$\lambda = 0.495$ for an effective sample size of 1004. Figure 4(a) shows that more than twice the number of responses is accurate, when the tiles are aligned along the same axis. Because each participant was shown both versions of the chart, we can use a paired $t$-test to compare mean accuracy between the two charts. The resulting $t$-statistic is highly significant ($t$ statistic: 16.1, df: 1656, $p$-value: $< 2.2e\text{-}16$).

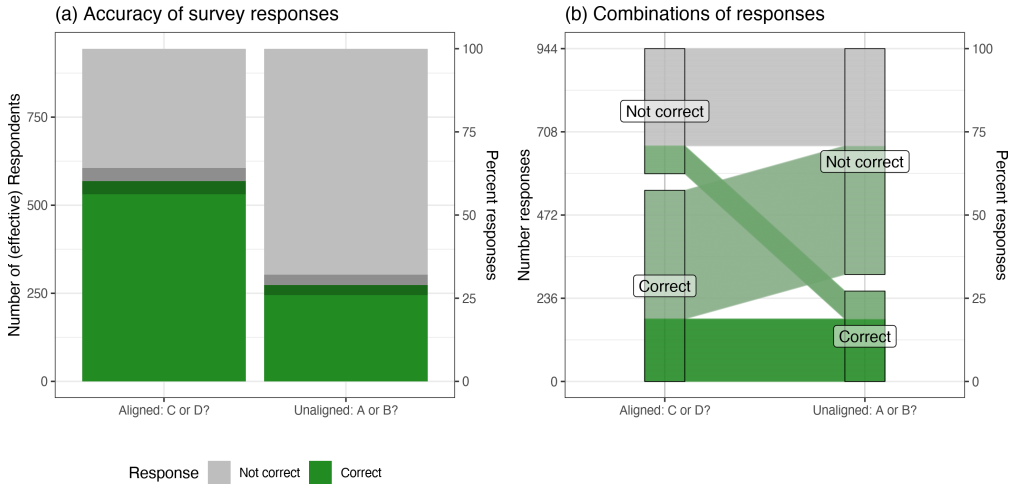



Figure 4: On the left (a), a stacked barchart shows the number of respondents with correct (green) and incorrect (grey) responses to the two comparison questions. When tiles are aligned along the same axis, more than twice the number of responses is accurate. The shaded area along the top of the green tiles corresponds to 95% confidence intervals around (marginal) correct responses. On the right (b), a parallel coordinate plot shows all combinations of responses. There's a huge asymmetry in the number of responses, where participants answered only one of the questions correctly. A lot more responses are correct when comparing aligned tiles than unaligned tiles.

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 **??**.

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

Borgo, Rita, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, et al. 2017. "Crowdsourcing for Information Visualization: Promises and Pitfalls." In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, edited by Daniel Archambault, Helen Purchase, and Tobias Hoßfeld, 96–138. Lecture Notes in Computer Science. Springer International Publishing.

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.

Heer, J., and M. Bostock. 2010. "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design." In *Conference on Human Factors in Computing Systems - Proceedings*, 1:203–12. ACM. https://doi.org/10.1145/1753326.1753357.

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): 718–26. https://doi.org/10.1109/TVCG.2021.3114874.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (1): 1–19.

———. 2010. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley and Sons.

———. 2020. "Survey: Analysis of Complex Survey Samples."

O'Muircheartaigh, Colm, and Steven Pedlow. 2002. "Combining Samples Vs. Cumulating Cases: A Comparison of Two Weighting Strategies in NLSY97." In *ASA Proceedings of the Joint Statistical Meetings*, 2557–62. http://www.asasrms.org/Proceedings/y2002/Files/JSM2002-001082.pdf.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.