Journal of Data Science xx (xx), 1–16 DOI: xx.xxxx/xxxx_xxxx
July 2020

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.

Introduction

Should the abstract match – or be close to – our SDSS short abstract?

What do viewers see when we show them a data chart? A data chart – at its core – maps quantitative values to graphical elements representing their relative values. Modern data visualizations are much more than a simple, objective mapping of values to a plane; they contain contextual and design elements, and are often structured to support the viewer in understanding a particular view of a set of data or specific pattern underlying the values. The design of a data visualization impacts a viewer's ability to achieve that understanding;

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

a poorly designed data visualization may leave viewers struggling to understand the content or context, or make it difficult to complete accurate and useful comparison of values across groups or time points. More broadly, the design of a data visualization can change how viewers interact with the chart.

A crucial step in the process of interacting with and understanding a chart is the viewer's employment of comparisons of the parts within. Cleveland and McGill (1984) observed as such, and in their seminal study defined the better visual among a pair 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, which was then extended by Mackinlay (1986) to a theoretical framework ranking tasks' order along their ordinal and nominal scales, as shown in Figure 1.

Cleveland and McGill's work – while a foundational user study in graphical perception – utilized a small convenience sample, consisting of only a few individuals recruited from among the authors' coworkers and their spouses. Heer and Bostock (2010) reproduced Cleveland and McGill's rankings using a larger sample from a crowd sourcing platform. A total of XXX Amazon Mechanical Turkers were involved. Crowd sourced samples were shown by Borgo et al. (2017) to be biased towards more male, younger, and relatively higher education relative to the adult U.S. population as a whole. These study populations are thus not representative of the general population, a common target audience for data visualization and scientific communication work. Further, the populations' emphasis on higher education individuals also leads to results which hold for groups of individuals who may be more likely to have prior exposure to data visualization in the context of scientific communication, or more exposure to data topics in higher education, but may not hold across other groups within the population. I don't know if the prior statement is going too far? Maybe we can rephrase.

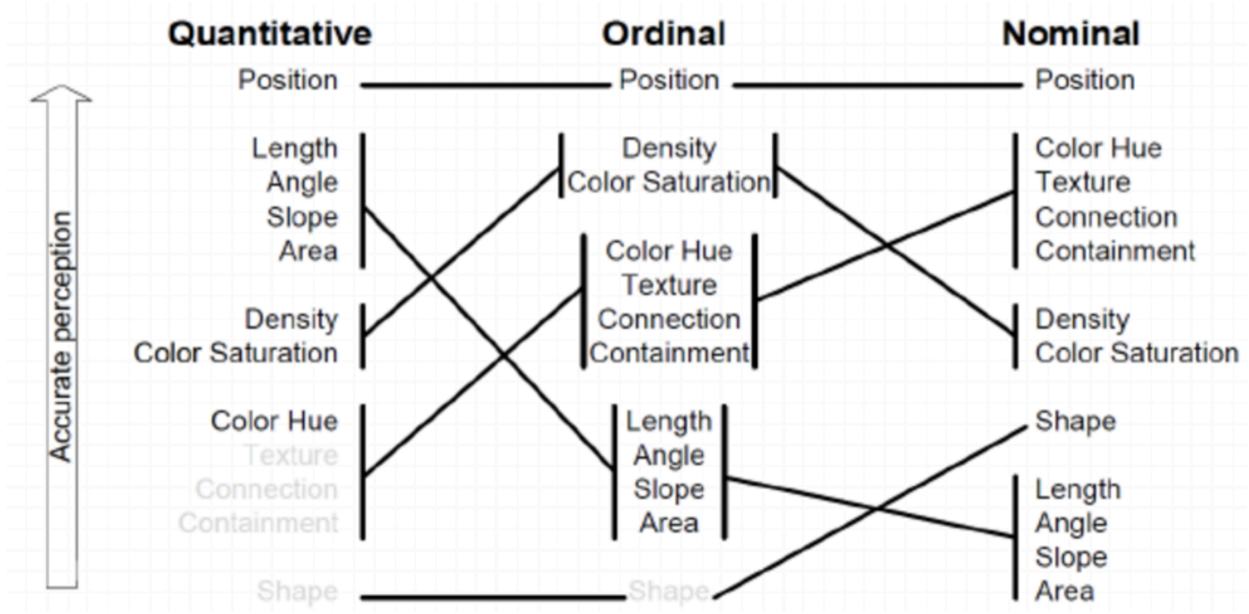
The mackinlay thing feels like maybe we are overemphasizing the ranking done by cleveland/mcgill and then mackinlay - I think we could emphasize less if we want

Our work – as that of Cleveland and McGill and Heer and Bostock – centers around studying data visualization design choices and their impact on viewer behavior and accuracy of viewers' responses. We seek to answer whether it is possible to reproduce some of their findings in the context of a survey with a large, nationally-representative set of respondents, and within that context we focus on the following research questions:

1. How do structural design choices in a data visualization impact viewers' ability to identify the larger of two elements?
2. How is viewer interaction with the task impacted by structural design choices in a data visualization?

We employ a probability-based survey panel and run a series of perception tests with nationally-representative samples of respondents from that panel. The advantage of using a probability-based 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 perceptional abilities. Second, the sample is representative of the general adult public in the U.S., which is an important target audience for scientific communication, and this allows us to test whether prior results from convenience samples hold with a nationally representative sample. Do we want a stronger statement about this here?

We present viewers with structural variations on bar charts and pie charts and ask them to answer questions comparing the size of elements within those charts. In this work, we present the series of tests we completed and the resulting findings. The remainder of the paper is organized as follows: first, we describe the design of the visual stimulus used in our perception tests. We then describe the population of study respondents and obtained survey sample. Subsequently, we share our analyses of the resulting survey responses across each of our tests, including analyses on accuracy of responses and response behavior. Finally, we discuss implications of this work and next steps.



The Mackinlay ranking of perceptual task.

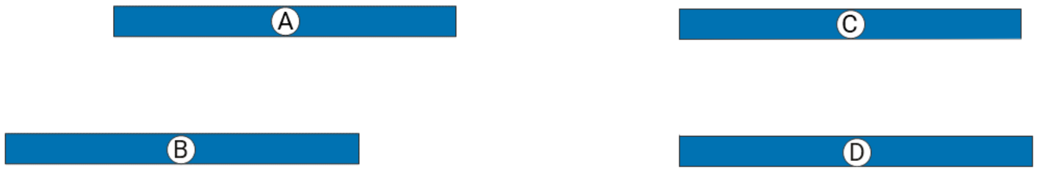
Figure 1: Ranking of perceptual tasks, as given by Mackinlay (1986). The ranking of tasks on the quantitative scale are empirically verified by Cleveland and McGill (1984).

The figure below is still a bit of an orphan, but I think we can revisit the intro ad study setup after we get more of the results in the paper

Study design – stimulus

Each task is made up of two elements: a visual stimulus and a question about that image that the viewer is asked to respond to. In our study, each visual stimulus is an image of a data visualization, while each question prompts viewers to identify which of two marked pieces in the data visualization is larger.

What specifically is each task? Each comparison between marked pairs is designed to be a difficult task, with the difference between the values represented in the two marked pieces being a just-noticeable difference. The **Just-Noticeable Difference** (JND) is defined as



(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. When participants are asked to compare the size of these tiles in barcharts, the predominant response for the unaligned pair on the left is 'they are of the same size'. In contrast, more than half of the viewers respond with 'D is bigger' to the aligned pair of bars on the right.

the smallest difference that will be detected 50% of the time. Prior results from studies on bar charts and pie charts (Lu et al. 2022) inform the differences in charts shown to our survey panelists.

Why do we utilize just-noticeable differences? We employ comparisons at the JND in our tasks in order to maximize our ability to identify the impact of design changes on viewer accuracy and behavior. Asking perception tasks in a survey differs from the controlled environment of a cognitive lab, where these kind of questions may usually be assessed. Rather than asking the same (or similar) type of question with varied signal strength dozens or hundreds of times, we are limited to only a few questions at a time. With a small set of tasks, we need to present tasks that are perceptually hard, and thus ask questions about stimuli that are close to our perceptual threshold. Therefore, we focus on questions which vary the presented image, but ask viewers to compare the same values across those varied images.

How do we vary the task? We ask participants to determine which of two just-noticeably different marked pieces is larger within the data visualization image, and we vary the structural design of that data visualization image. We focus on three main sets of structural variation in the design. First, we vary the alignment of the pieces in question. Viewers are presented with two marked pieces in a chart that do not share a common baseline, then two pieces that do share a common baseline. Second, we vary the orientation of the chart – a vertically oriented stacked bar chart, a horizontally oriented version of the same chart, and then a wider version of the horizontally oriented chart. We need to finalize what we end up showing in the results and then revisit this section... Third, we consider a facetted bar chart and facetted pie chart.

Not sure where should put this, parking it here for right now: expectations - start

What we call 'aligned' and 'unaligned', here, is similar to Cleveland and McGill's set of rankings, but with some modifications: both 'aligned' and 'unaligned' bars (in tests 1 and 2) or wedges (in test 3) share the same axis. Aligned tiles are additionally anchored in the same position in one dimension, i.e. the difference between their sizes can be reduced to a positional assessment. Unaligned tiles do not share this anchor, however, the context of the other tiles in the chart provide a frame, which *should* help with an assessment of the tiles' sizes beyond a comparison of (arc) lengths or areas.

We would expect that comparing unaligned tiles is a harder task (with correspondingly lower levels of accuracy) than a comparison of aligned tiles, with the framing given by the context of the other tiles in the same column (or the same pie) mitigating some of this difficulty. Figure 3 gives an overview of the comparisons of tasks 1 through 3 and the closest corresponding tasks in Cleveland and McGill.

expectations - end

How do we present the task? The format of a survey guides the format and design of the questions asked and how they are presented to respondents. First, participant instructions must be delivered in a very short and easily understandable format, because participants cannot ask clarifying questions about the task as they might be able to in a cognitive lab setting. I'm not sure how important this second point is... Second, participants should be

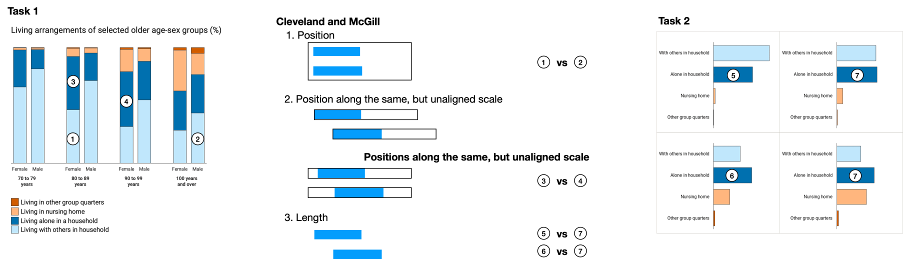


Figure 3: Comparisons made in charts across tasks 1 through 3 within the Cleveland and McGill ranking

given some context for the tasks they are being asked to complete; in a given round, our set of tasks appear as a group in the midst of other survey questions and topics, and without providing respondents some transition we risk a jarring shift and low respondent engagement in the task. Finally, to prevent viewers from being exposed to slight variations of the same stimulus in a row (and risk unforeseen order effects or respondents using prior questions to inform their responses), we either split a survey sample in two and show each subsample a distinct version of the chart or test variations of a chart across distinct rounds of the survey.

Somewhere here we should also talk about how we measure viewer behavior/interaction with the chart: certainty, time spent, zooming, devices, etc.

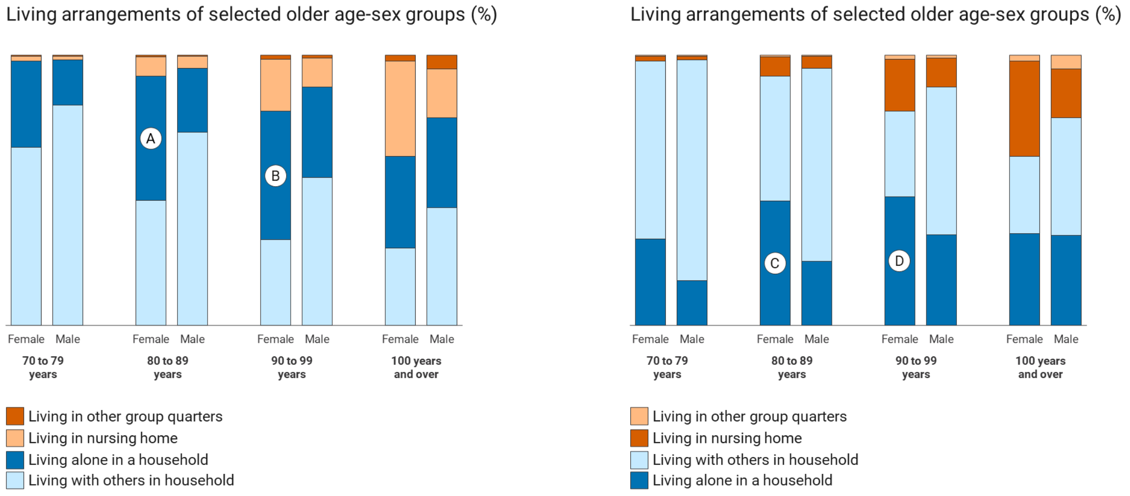
Test 1: Alignment in stacked bar charts

Figure 4 shows the two stacked bar charts shown to participants in the first test. 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 4 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 both the difference in size between the bars and the horizontal 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 responses can therefore be attributed to this difference in presentation.



(a) **Unaligned bars**

(b) **Aligned bars**

Figure 4: Two stacked (vertical) barcharts. In each barchart, two tiles are marked. In both instances, the tile on the right is (very slightly) larger.

Test 2: Orientation of stacked bar charts

In Figure 5 we see two structural variations on the stacked bar chart seen in Test 1: first, the Task 1 chart sized equally but rotated 90 degrees with bar lengths shown horizontally. Second, the horizontal chart with a smaller aspect ratio so the bars appear thinner and longer. Note that the wide version of the horizontal bars increases the pixel difference between the two marked pieces relative to the standard versions of the horizontal and vertical bars.

This test utilized a split sample approach, where 50% of participants in the sample were shown the tall horizontal bar chart and 50% were shown the wide horizontal chart. Participants were asked to complete the same task as in Test 1; selecting the correct response among:

1. A is bigger
2. B is bigger
3. They are the same

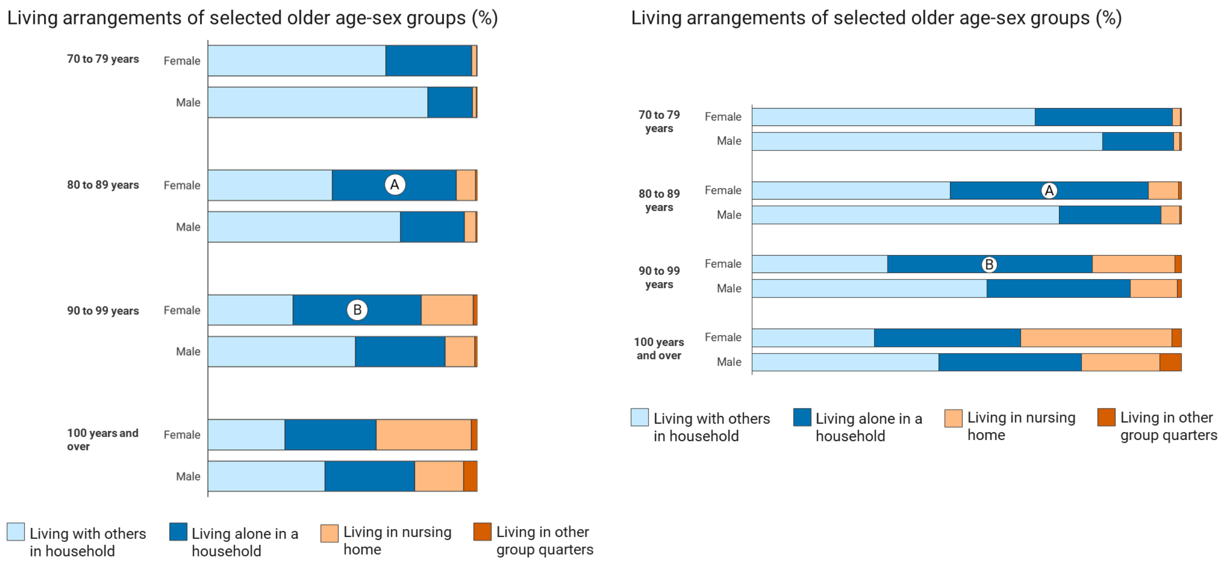
Again, answer 2 is the correct answer.

Test 3: Alignment and orientation in pie charts

Let's talk about the pies here

Study design – participants

Participants were recruited as part of NORC's AmeriSpeak panel, which utilizes a probability-based sampling methodology and samples U.S. households from NORC's National Sample Frame that provides coverage of over 97% of U.S. households. The current panel size is 54,001 panel members aged 13 and over residing in over 43,000 households **CITE DENNIS 2022** Dennis (2019). I don't think I've got the right citation Each test was conducted using the AmeriSpeak Omnibus survey, which runs biweekly and samples around 1,000 U.S. adults



(a) **Tall horizontal bars**, unaligned

(b) **Wide horizontal bars**, unaligned

Figure 5: Changes to to the vertical design are made in two steps. First, the original chart is rotated. The areas of all tiles are kept the same. In a second step, the aspect ratio of tiles is changed while keeping the areas of tiles the same.

to answer questions on a variety of topics.

XXX I think what you are trying to say is that we do not know, if participants have seen the stimuli multiple times, but if they did, there was at least a month in between? Yes that's what I was trying to say here!!

Given the nature of pulling a sample from the panel for each Omnibus round, there is a possibility that some participants may have been included in multiple rounds; however, our collected data is not a longitudinal or panel study, so we do not have repeated responses from the same participants across each of our tests. Each of the tests was run at least a month apart, so if respondents participated in multiple rounds, there was at least a one month gap between viewing each visual stimulus.

Study design – survey weighting

Paragraph on how survey weights are derived? – **Question for Ed – is this something AmeriSpeak can provide boilerplate language for?**

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

When combining responses that were gathered during distinct rounds of the Omnibus survey, we make weighting adjustments to ensure that weights in each sample are properly calibrated to the population total across both rounds in the resulting model. We *combine* (rather than cumulate) surveys S_1 and S_2, as described in O'Muircheartaigh and Pedlow (2002), by multiplying weights in S_1 and S_2 by λ and $1 - \lambda$ respectively.

$$\lambda = \frac{n_1/d_1}{n_1/d_1 + n_2/d_2},$$

where n_1 and n_2 are the nominal sample sizes and d_1 and d_2 are the design effects for the estimators. Here, d_1 and d_2 are estimated as

$$d_1 = 1 + CV(w_i \in S_1)^2 \quad \text{and} \quad d_2 = 1 + CV(w_i \in S_2)^2$$

where CV is the coefficient of variation of the weights within each sample, and is estimated as in Kish (1965) :

$$CV(w \in S) = \frac{\widehat{Var(w)}}{\bar{w}^2}.$$

Note that O'Muircheartaigh and Pedlow (2002) estimate λ separately for any combination of race/ethnicity by sex. We employ that strategy whenever we include demographic variables in the analysis, otherwise we will use a single adjustment for the weights.

Results

Respondents

Description of respondents

- # of respondents in each round
- broad overview of demo characteristics(?)
- raw sample size and effective sample size that we are analyzing **in each test**

Test 1

- Alignment in stacked bars
  - Binary accuracy in responses
    - Figure: binary correctness and binary ggpcp diagram
    - Table:(?) T-test results
    - Table: results of cell means model on certainty
  - Ordinal accuracy in responses
    - Figure: ordinal correctness and ordinal ggpcp diagram
    - Table: Fitted model for ordinal response and alignment
  - Ordinal accuracy in responses by demographics
    - Figure: ordinal correctness by demographics, split by aligned and unaligned
    - Table: Fitted model for ordinal response and alignment by demos
      - Maybe this table should go in the appendix?

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 6(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: $< 2.2\text{e-}16$).

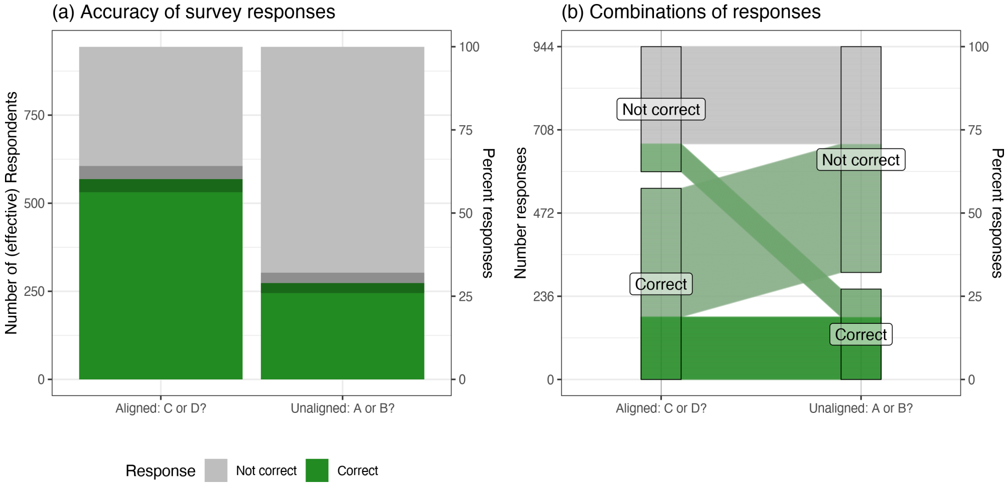


Figure 6: On the left (a), a stacked barchart shows the number of respondents with correct (green) and incorrect (grey) responses to the two comparison questions. When tiles are aligned along the same axis, more than twice the number of responses is accurate. The shaded area along the top of the green 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.

Test 2

- Structure of stacked bars – vertical, horizontal, horizontal wide
  - Accuracy and responses
  -

Test 3

- Facetted bars and pies
  - Accuracy and responses
  - Timing of responses

Results to discuss:

- Differences in accuracy between aligned and unaligned bar
- Differences in accuracy between framed, floating, and general pie
- Differences in timing and zooming behavior across all designs

Conclusion

Key findings to discuss:

- Surveys can be used to ask these types of questions – in the midst of other topics and with a limited number of questions, we are able to ask perception questions and produce results consistent with prior studies
- Reproduced (some) prior convenience sample results using a large, nationally-representative survey population:
  - Cleveland: position (aligned) versus unaligned
  - unaligned tiles within stacked bar are a hybrid between framed bar and floats
  - Talbot: pie charts - individual wedges vs piecharts
- Design choices impact viewer accuracy:
  - new finding: pie charts accuracy in work better when framed (wedges)
- Other important measures beyond accuracy: time to completion (automatic on web), certainty in response (asked).
- Studying how viewers interact with charts can be done in a survey, and more expansive work on this topic should be completed to understand how the general public interacts with and understands charts.
  - Design choices impact viewer behavior: zooming
- (some) Demographics matter: age and gender does not seem to matter for perception economic background and education does matter for perception: we need to know, we

do not want to create a hurdle in communication

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.

Dennis, J Michael. 2019. "Technical Overview of the AmeriSpeak Panel NORC's Probability-Based Household Panel." *NORC at the University of Chicago.*

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.

Kish, Leslie. 1965. *Survey Sampling.* Wiley.

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."

Mackinlay, Jock. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–41. https://doi.org/10.1145/22949.22950.

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/.