Journal of Data Science xx (xx), 1–17
July 2020
DOI: xx.xxxx/xxxx_xxxx

Testing Charts: viewer's perceptual accuracy in surveys

Kiegan Rice*[1], Heike Hofmann†[2], Nola du Toit‡[1], Edward Mulrow§[1], and [2]

[1]*NORC at the University of Chicago*

[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? I don't think it has to

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; 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.

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

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. I think this is going to hold - the demographics are pretty significant

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?
3. Are there differences in perception and interaction with the tasks across demographic groups?

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., and

this allows us to test whether prior results from convenience samples hold with a nationally representative sample and whether there are differences in those results across demographic subgroups.

We present viewers with structural variations of bar 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 stimuli 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.

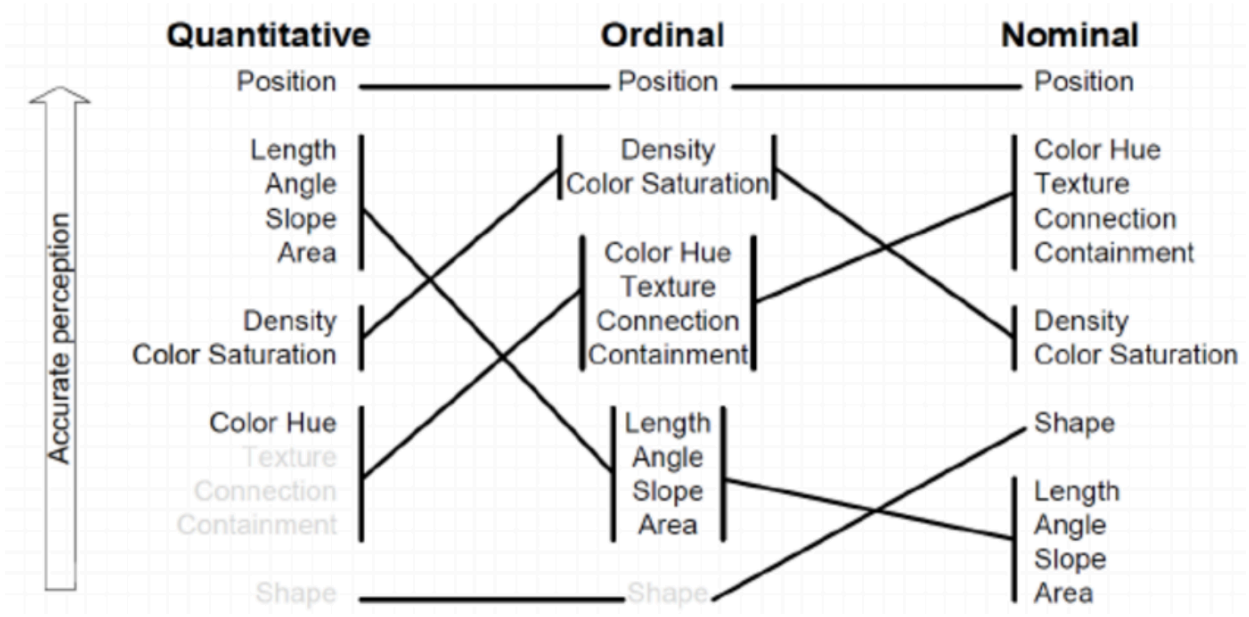


Figure 1: Ranking of perceptual tasks, as given by Mackinlay (1986). The ranking of tasks on the quantitative scale are empirically verified by Cleveland and McGill (1984).

Study design – stimulus

Each task is made up of two elements: a visual stimulus and a question about that stimulus 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 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 – we rotate the vertical stacked bar chart, and present a horizontally oriented version of the same chart, with identically sized marked pieces. Finally, we change the aspect ratio of the chart and present a wider version of the horizontally oriented chart which has longer, but thinner, marked pieces.

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 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 2 gives an overview of the comparisons of tasks 1 through 3 and the closest corresponding tasks in Cleveland and McGill.

I think we should include the floating bars here to demonstrate that they are worse than the wide horizontal bars.

expectations - end

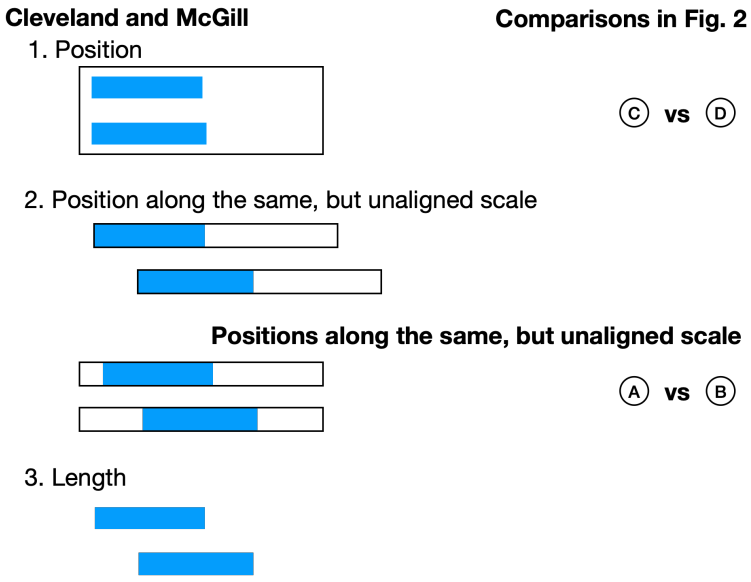


Figure 2: Comparisons made in charts within the Cleveland and McGill ranking

How do we present the task? Our use of a survey format 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. Second, we want to utilize content within the data visualization image which is not socially or politically charged for the average U.S. adult; this risks participants reacting to the subject matter within the chart rather than focusing on the presented task. For this reason, we utilize data on living arrangements of older U.S. adults – a topic which most U.S. adults will have some familiarity with, but is not inherently polarizing. 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.

When viewing each chart, participants were asked to compare the relative sizes of marked elements within the chart:

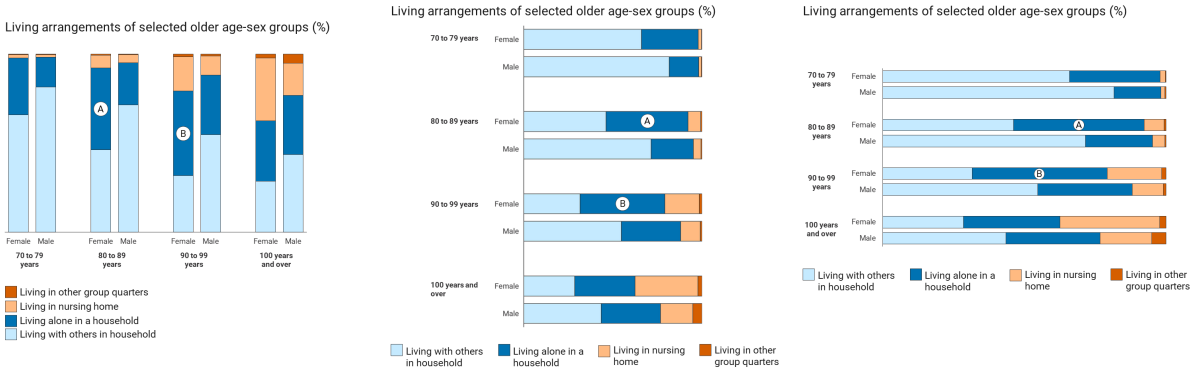
"There are many charts used in the news media to portray data visually. Looking at the chart below, which of the marked dark blue pieces is bigger, A or B? Just your best guess is fine.

- A is bigger
- B is bigger
- They are the same"

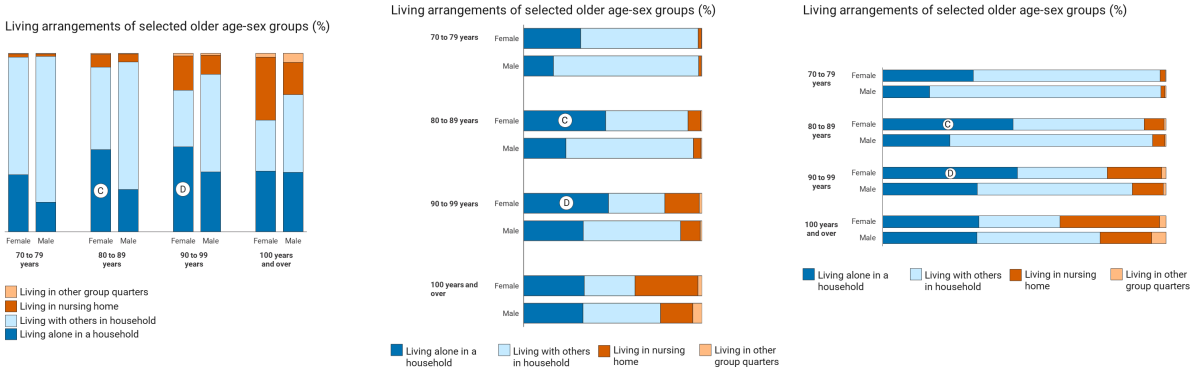
When presented with the aligned version of each chart, pieces were marked with a C and D and the question text is updated accordingly. In all scenarios, the second option (B [D] is

bigger) is the correct response.

The time in seconds that each respondent spent on each task was recorded, as well as whether the participant zoomed in on each chart while answering the question. Respondents were also asked to rate their certainty in their response to each question on a five-point scale.



(a) Vertical stacked bar, unaligned (b) Horizontal stacked bar, unaligned (c) Wide horizontal stacked bar, unaligned



(d) Vertical stacked bar, aligned (e) Horizontal stacked bar, aligned (f) Wide horizontal stacked bar, aligned

Figure 3: Each of the visual stimuli presented to viewers. In each bar chart, two pieces are marked. The larger piece in each chart are pieces B and D.

Task 1: Vertical stacked bar

Figure 3(a,d) shows the two stacked bar charts shown to participants in the first task. 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 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. 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.

Task 1 was presented to viewers in one of two rounds (Rounds 1-2), with four color scheme variations each presented to 50% of respondents in each round. There were no significant differences in respondent accuracy or behavior across color schemes, and we combine them here into one set of stimulus responses to compare structural differences. Should we put the color variations and a test for significance in accuracy/a visual in the appendix/supplementary materials?

Task 2: Horizontal stacked bar

The visual mappings in the second task, shown in **??**, are identical to the first task, but the axes of the chart are rotated so that the stacked bar is represented in a horizontal format. This represents a structural change in how the data are presented to the viewer while preserving the pixel size of the elements viewers must compare. Tasks 2 and 3 were asked within the same survey round (Round 3) and the full sample was randomly split among respondents, with 50% of participants seeing the Task 2 stimuli and 50% of participants seeing the Task 3 stimuli.

Task 3: Wide horizontal stacked bar

The images utilized in the third task again represent the same data, and the overall image has dimensions which are identical to the images in the first two tasks. However, the aspect ratio of the plotting area is adjusted to increase the length in pixels of the stacked bar elements, and decrease the relative thickness wording? of each bar.

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 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 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 a weighting adjustment to ensure that weights in each sample are properly calibrated to the population total across all rounds in the resulting model.

Do we need to update this description to just say how we did it for combining across all the rounds represented?

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.

The data for this paper were collected in several rounds as part of the NORC Omnibus.

we might be able to include the lambda values for combining the tables here as well

Results

There are differences by alignment There are not huge differences across structures However, we see a difference in how people react across different structures (zooming behavior, certainty? time?)

- We can reproduce prior findings with nat rep sample
- We can rank
- We can rank better because we have a bigger sample size
- We can find changes in viewer behavior
- Demographics matter to accuracy and choices

Respondents

A total of 2807 respondents participated across the three rounds. The number of responses and corresponding effective sample sizes in each round are shown in Table 1.

We need to be careful about not making it seem like the rounds are the same as the tasks, since we have 3 rounds and 3 tasks. We describe it, but I feel like this table kind of muddies the waters. Could we instead do a table of respondents by task? does this comment still stand after the re-organization? ... I would suggest to move it up to the survey methods and add the lambdas

Table 1: Survey rounds: dates, number of participants (nominal sample size), effective sample size, sum of weights, and factors for the convex combination as discussed above.

Name	Date	# Participants	effective sample size	Sum of weights $\sum_i w_i$	λ_i
Round 1	April 2022	933	521.1	934.9	$\lambda_1 = 0.343$
Round 2	May 2022	953	485.7	953.4	$\lambda_2 = 0.320$
Round 3	Jun 2022	921	513.5	923.1	$\lambda_3 = 0.338$
Total	—	2807	1520.8		$\lambda_1 w_1 + \lambda_2 w_2 + \lambda_3 w_3$

All responses were combined into one set of survey responses, with adjusted combined sample weights and indicators for which task each respondent was exposed to. Across the distinct stimuli, we model the resulting accuracy of responses, metrics on viewer interaction with the chart, and differences across demographic groups.

Insert figure here of demographic breakdown across tasks or rounds?

```
[1] 1520.786
```

Accuracy of responses

We first investigate respondent accuracy in selecting the correct response. As participants were able to select the option 'They are the same', there are several ways to model accuracy and response. The argument could be made that for the purposes of practical interpretation, 'they are the same' is a correct choice, as the options are visually very similar and not substantially different values within the context of the data shown in the chart. However, we are interested in understanding whether viewers *can* perceive the difference and correctly identify which piece is larger. We consider multiple ways of modeling response to investigate participant response patterns. First, we define a measure of binary 'correctness', for which all answers that are not the correct option (B, D is bigger) are 'incorrect', including the selection of 'they are the same'.

??(a) displays all responses along the binary correctness measure, separated by whether the stimulus was an aligned task or unaligned task. We can see that levels of accuracy for all responses are significantly higher for the 'easier' (aligned) task, with about twice as many respondents correctly selecting the larger of the two marked elements.

Can we update the figure and t-test with combined data across all tasks/all rounds 1-3?done

Can we update this to include the values for ROund 3 as well?done

Figure 4(a) shows that more than twice the number of responses are correct, 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: 21.2, df: 2265, p-value: $<$ 2.2e-16). how about: p-value: $<<$ 0.0001 – the e-16 looks very computery

We need to update this chart to include the data from all 3 rounds instead of just the first 2 done

Next, we consider an ordinal model to investigate response behavior across all three options – 'A (C) is bigger', 'B (D) is bigger', and 'they are the same', and consider these response patterns across each of the stimuli.

Figure 5 shows the results of a cell-means model with ordinal response Y_k, where Y_k is the kth participant's response, $Y_k \in \{1, 2, 3\}$, where 'correct' is encoded as 1, 'they are the same' is encoded as 2, and 'wrong' is encoded as 3:

$$\text{logit } P(Y_k \leq \ell) = \mu_{ij\ell(k)},$$

where $\ell \in \{1, 2\}$; $i \in \{1, 2\}$ is the comparison type (1 = Aligned, 2 = Unaligned), and $j \in \{1, 2, 3\}$ is the chart design, with 1 = Vertical, 2 = Horizontal, and 3 = Horizontal wide. The estimated values and 95% confidence intervals are shown in Table 2

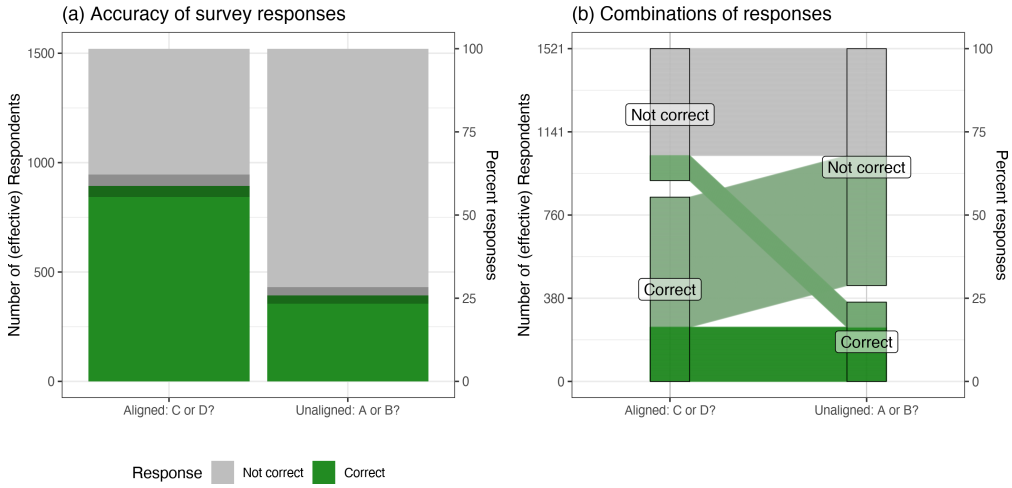


Figure 4: On the left (a), a stacked barchart shows the number of respondents with correct (green) and incorrect (grey) responses to the two comparison questions. When tiles are aligned along the same axis, more than twice the number of responses are correct. The shaded area along the top of the green 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.

Table 2: Log-odds for the cell-means model, letters behind numbers indicate pairwise significances. Within the same **column** values are significantly different (at 5%) if they do not share the same letter.

Log odds of accuracy by task and chart type

	correct | same or wrong	correct or same | wrong
Unaligned		
Horizontal	0.22 [0.15, 0.32] a	5.54 [4.04, 7.59] a
Horizontal wide	0.26 [0.19, 0.37] ab	6.20 [4.46, 8.61] a
Vertical	0.41 [0.36, 0.47] b	6.90 [5.75, 8.28] a
Aligned		
Horizontal	0.63 [0.49, 0.81] c	14.02 [9.23, 21.30] b
Horizontal wide	2.67 [2.04, 3.48] d	17.12 [10.49, 27.95] b
Vertical	1.51 [1.33, 1.71] e	11.33 [8.99, 14.28] b

The same pattern in accuracy holds across each of the three structural variations; the aligned task has a higher level of accuracy than its unaligned counterpart. Interestingly, while we expect an improvement in accuracy when shifting from the horizontal to the horizontal wide design given the larger difference in pixel length between the two pieces, the resulting effects

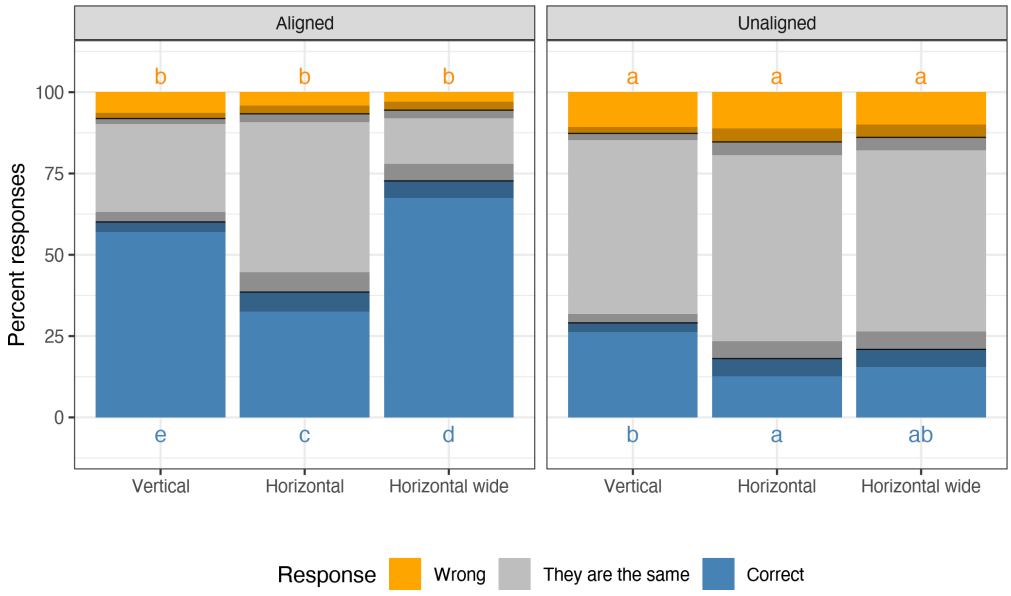


Figure 5: Responses for accuracy in the three designs. Responses to the same task are shown side-by-side for the three designs. The overlaid rectangles represent 95% confidence intervals. The letters in blue and orange encode significances between pairwise proportions: two bars have a significantly different proportion (at a 5% significance level) if they do not share a letter. There is no significant difference between the three designs for wrong responses. When tiles are unaligned, the horizontal wide barchart is showing the highest accuracy. For aligned tile, the horizontal wide design and the vertical design do not show a significant difference in accuracy.

on the accuracy of the responses are not completely straightforward: The shift from a vertical to the (tall) horizontal design is detrimental to an accurate perception for both aligned and unaligned comparisons. The re-scaled design of the wide horizontal bars reclaims some of the loss for unaligned bars and outperforms the vertical design by a similar margin in aligned bars, but does not out-perform the vertical design when comparing unaligned tiles.

Respondent behavior

A contributing factor to the observed patterns in response accuracy might be the way that participants interact with the different designs. Generally, about half of all participants make use of the option to zoom into charts - while zooming does help with the overall accuracy (which is in agreement with the findings by Lu et al. (2022) about the physical size of stimuli), the increase is not significant. However, different designs lead to different rates of zooming: when dealing with the vertical design, the rate of zooming is significantly higher than for the two horizontal designs.

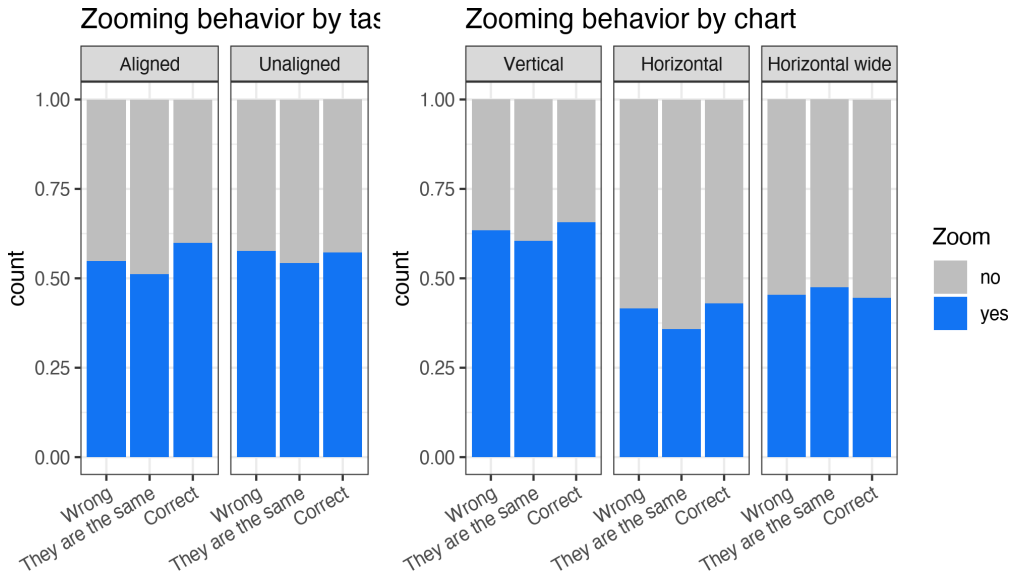


Figure 6: Zooming - not significant for accuracy or task, but changes by the type of chart.

Let Y_{jk} describe the zooming behavior of panelist k on task j. We model zooming behavior (no = 0, yes = 1) as a logistic regression by correctness of response (ρ), task (τ), and design (δ) of the chart:

$$\text{logit } P(Y_k \leq 1) = \mu + \rho_{i(k)} + \tau_{j(k)} + \delta_{\ell(k)},$$

I'll pretty up Table 3 if we are going to keep it. I think we should! it's nice to have a model for this

Table 3: Coefficients for logistic regression of zooming by task

Estimates for logistic regression on zooming behavior

term	estimate	SE	t-statistic	p-value
Intercept	−0.35	0.15	−2.4	0.0155
response3They are the same	−0.17	0.09	−1.9	0.0578
response3Wrong	−0.06	0.13	−0.5	0.6466
taskcd	0.00	0.06	−0.1	0.9435
chartHorizontal wide	0.27	0.15	1.8	0.0752
chartVertical	0.98	0.13	7.7	< 0.0001

However, zooming behavior *only* differs significantly by structural design – rates do not differ significantly between the aligned and unaligned tasks, nor do they differ significantly by the chosen response.

Do we have anything about total time spent on tasks across the 3 structures? We have time on each question for rounds 2-3. And can we pull in some of the certainty stuff?

Differences across demographic groups

Let Y_k be the response of participant k, on a scale from 1 = 'wrong', 2 = 'they are the same' to 3 = 'correct'. We use a generalized cumulative logistic regression, where μ_i are intercepts $1 \leq i < 3$, X_k are demographics of the kth participant (in form of the model matrix), and β_i are the coefficients.

$$\text{logit } P(Y_k \leq i \mid X_k) = \mu_i + X_k' \beta_i$$

Table 4 doesn't survive the formatting to pdf very well - need to shorten the row names or put into multiple lines

Table 4: Demographics matter for perception, particularly when the tasks get harder.

Log odds of accuracy by task and demographics of respondents

	Aligned tiles						Unaligned tile			
	correct | same or wrong			correct or same | wrong			correct | same or wrong			correc
term	Est.	[95% CI]		Est.	[95% CI]		Est.	[95% CI]		Est.
Intercept										
	1.51	[0.91, 2.50]		12.75	[5.59, 29.10]	***	0.60	[0.35, 1.03]	.	3.24
Gender										
Female	0.82	[0.67, 1.01]	.	1.19	[0.81, 1.74]		1.03	[0.82, 1.30]		1.42

Age									
30-44	1.08	[0.78, 1.49]		0.87	[0.43, 1.75]		0.79	[0.55, 1.13]	0.94
45-59	1.13	[0.80, 1.60]		0.95	[0.46, 1.95]		0.92	[0.62, 1.35]	0.76
60+	0.96	[0.69, 1.34]		0.70	[0.35, 1.40]		1.01	[0.70, 1.45]	0.73
Education									
HS or equivalent	0.77	[0.46, 1.30]		0.50	[0.20, 1.21]		0.81	[0.47, 1.40]	1.22
Some college	0.91	[0.56, 1.49]		0.74	[0.32, 1.71]		0.63	[0.38, 1.06] .	1.17
Bachelor	0.79	[0.48, 1.32]		1.12	[0.42, 2.97]		0.53	[0.31, 0.93] *	1.82
Post graduate	0.84	[0.49, 1.44]		1.16	[0.41, 3.30]		0.51	[0.28, 0.91] *	2.18
Income									
$30k - $60k	1.19	[0.87, 1.63]		1.25	[0.77, 2.05]		1.00	[0.71, 1.41]	1.22
$60k - $100k	1.23	[0.90, 1.68]		2.14	[1.18, 3.86]	*	0.87	[0.60, 1.27]	1.88
$100k plus	1.36	[0.99, 1.88]	.	1.57	[0.86, 2.86]		0.86	[0.59, 1.25]	2.18

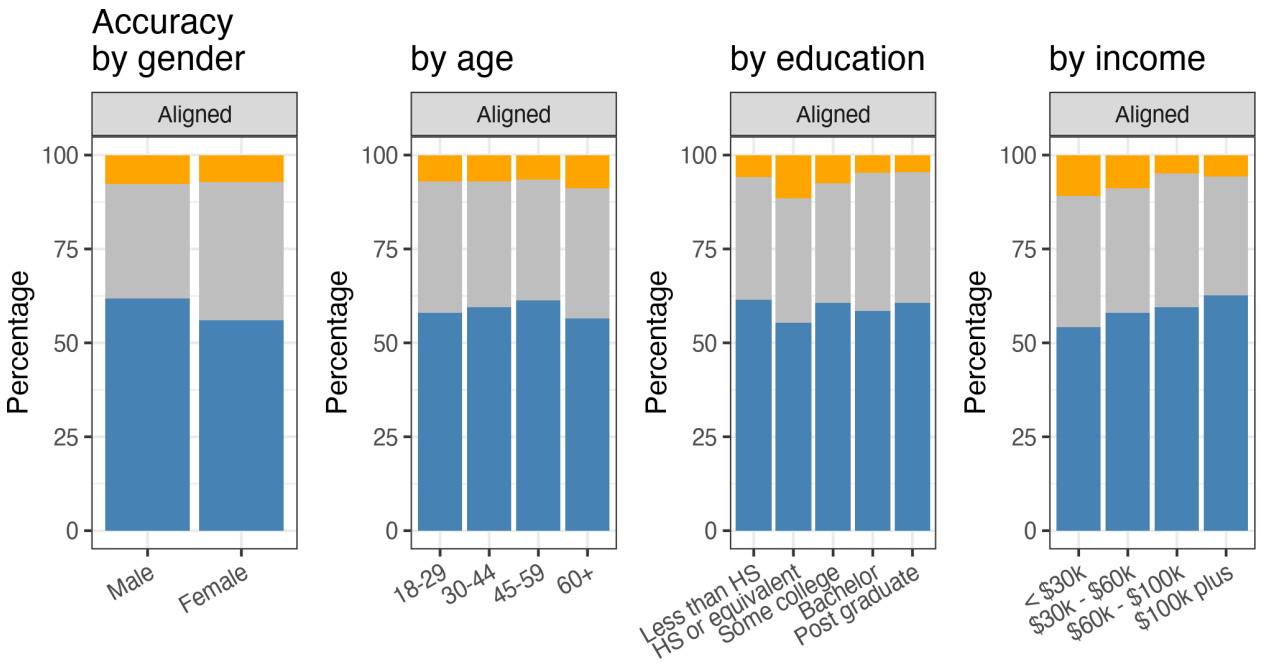
Results to discuss:

- Differences in accuracy between aligned and unaligned bar
- Differences across structures
- Differences in timing and zooming behavior across all designs
- Demographic differences and their interaction with alignment

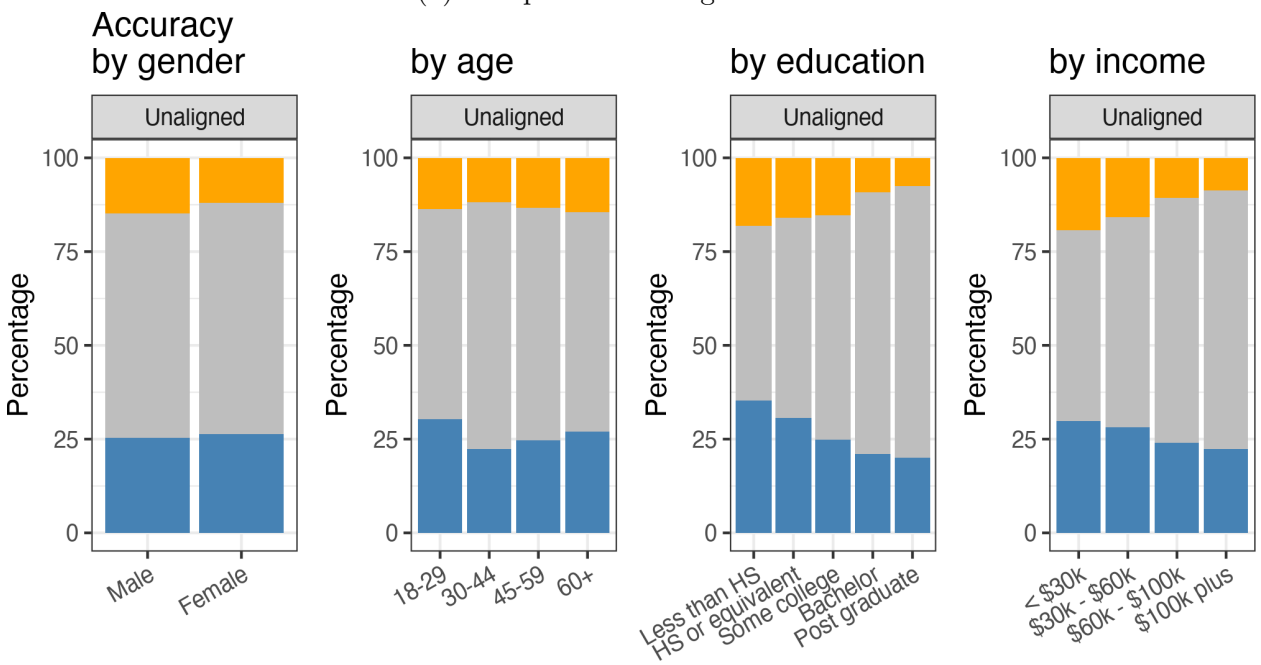
Conclusion

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



(a) Comparisons of aligned tiles



(b) Comparisons of unaligned tiles

Figure 7: Panelist's demographics matter, particularly, when the task difficulty increases. For aligned tiles, gender, age, and education are not significant factors. However, income levels do have - a small - effect. When income levels increase, the percentage of wrong answers (orange) decreases, while the percentage of correct answers (blue) increases slightly (significant at below 0.05). The more difficult task of comparing unaligned tiles, demographics are more significant. XXX With increasing levels in education, we see a small drop in correct answers, as well as a small drop in wrong answers, for a significant increase in 'they are the same'.

- (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/.