Round 1: Response to Reviewers

Reviewer 1

- **The introduction doesn't provide a clear rationale for the questions' importance.**
- The entire article reads more like a project report rather than a scientific paper, with an excess of questions and bullet-pointed sentences.
  - *To address this, we have removed two of the instances of bulleted lists: one which listed out the three variations in testing, each described in a bullet, and a second which demonstrated the options participants have to select from, each option shown in a bullet. The third list, which we have kept intact, is a list of the three research questions we pose to answer through our study. The authors could not identify a reason to remove the list structure for our research questions, as it clearly separates each of the distinct questions for the reader. There are four questions in the text of the paper: one at the very beginning, posing the question 'What do viewers see when we show them a data chart?', and three associated with the research questions. We have not removed any of these, as a question mark at the end of a research question is appropriate, and posing an opening question in the introduction is also appropriate.*
- **There is an overabundance of figures without clear emphasis.**
- I have doubts about whether the paper aligns with the Journal's scope.
  - *Per AE and Editor, we are not responding to this comment.*

Reviewer 2

- **There are a couple of assumptions that assume certain factors (order of question, rounds) are negligible to the outcome. While I think this is reasonable, for a higher confidence in results it might be a good idea to test all visual**

stimulus within each round and block (if it is hard to be individually randomised) so that some participants get different question order.

- Major comments
  - **The title, abstract and conclusion suggests a wider context but the content is actually more limited in scope. I suggest changing the title to emphasise the main points in the article: 1) nationally representative sample and 2) testing stacked barplots. I think statements like "Our study results provide actionable insights that data visualization practitioners can utilize in the design of data visualizations." extrapolates too much. It should be more specific in that the actionable insights is *for stacked barplots* and also write concretely what the recommendation is (i.e., vertical display, align groups that most important to compare, etc.).**
    - *We have updated the statement in the conclusions per a specific recommendation in the minor comments. In addition, we have updated the paper title and some of the language in the abstract to reflect that the scope is limited to testing stacked bar charts.* **We will also update the recommendation language to be more concrete**
  - Do the authors have recommendations for (non-representative) visualisation perception surveys given their experience on nationally representative sample? E.g. given that the surveys conducted via Mturks etc are biased toward highly educated young males and authors found differences between the income group, what do these non-representative survey inform us?
    - *We have added the following text to our conclusions to address this point: "If studies using non-representative or crowd-sourced samples find non-significant results when studying elements of design, those results may not actually be non-significant for other populations, particularly lower education or lower income populations. Researchers should not disregard non-significant findings in those survey studies but rather ask whether the lack of significance would hold across population subgroups."*
- Minor comments
  - **I think the paper can benefit in the intro to include a stronger emphasis why we want to achieve a representative sample. The authors do address this in part but it reads like an aside ("..., a common target audience for data visualisation"). Maybe a paragraph on about how data visualisation is used by general public (e.g. results on polls in the news – this can also lead to why the authors chose to focus on barplots too as polls are often count data). Authors can also discuss how it is *imperative to be inclusive* as these data visualisation can be informative and the results need to be**

accessible by the *whole* population and not just the elite/highly educated population (which seem to be the bias towards at lot of these surveys have). I think also a stronger emphasis on the importance of the authors' works is needed, e.g., "this is the first study to ...". Also in conclusion, the authors state "The size of our sample aids in identifying this signal.", but the authors can instead state e.g. "our study is orders of magnitude larger than other studies with similar scope, demonstrating...".

  * *There are multiple points here; we will address each.*
  * **Stronger emphasis on why we want to achieve a representative sample**
  * **Stronger emphasis on the importance of the authors' work is needed**

- **While I think the content is fine, I think the grammar in the paper can be better polished. (It is not that the current grammar is bad, but I just think it can be improved to make this a better paper). I have outlined some of the suggested corrections below.**

  * page 2, second paragraph: "The viewer's employment of comparisons of the components" -> "The comparison of components"

    · *This change has been made in the text.*

  * page 2, third paragraph: "to be biased towards more male, younger and relatively higher education" -> "to be biased towards more male, younger and relatively well-educated individuals"

    · *This change has been made in the text.*

  * **page 2, third paragraph: "Further, the populations' emphasis..." -> This sentence was confusing. I kind of get the gist what the authors are trying to say but I wonder if it can be paraphrased better? E.g. "Arguably, we could limit the population of interest to highly educated individuals, however, this group is also likely to have prior exposure to data analysis in higher education, which may not be the case for other groups in the wider population."**

  * page 2, fourth paragraph: I"m not sure if you should start the paragraph with "Here".

    · *We have replaced 'Here' with 'In this work'*

- **Figure 1, the text within the visualisation are too small to see. Perhaps lay it out as a 3 x 2 instead?** - HH: *Figure 1 was not meant to have readable text within the stimuli visualizations but only give an overview of the different stimuli and the testing situations. We have increased the font size for the overview, expanded the discussion of the figure to reflect its purpose and added image (a) to the supplement in a larger format.* **create supplement and add stimulus**

- **page 8, first paragraph: "$ 2$" is a typo?**
  - *This has been resolved in the source code.* **Heike, please confirm in the rendered paper.** HH: yes, it's a typo, it was supposed to be the expression l <= 2 and the latex expression $\ell \geq 2$ did not render properly
- **Table 1, effective sample size adds up to 1520.3 but the total says 1520.8?** *HH: What is shown is the effective sample size of the combined survey rounds, not the total of each round's effective sample size. The fact that these numbers are close is due to similar weighting strategies being used in each of the individual rounds. We have now renamed the line "Combined Rounds".* **and added a clarifying footnote in the table**
- **Figure 3: I take it that the American Community Survey 2021 is considered the Census estimates? I think it makes more sense to swap the bar and point so that the bar is the Census estimate. This way you can use the errorbar plot. The margin of error around a barplot is reminiscent of a dynamite plot, which I don't think is well thought of in the community (at least not me). I think you can also color the bars to indicate whether your survey estimate covers the census estimate so it is easier to see which one covers it or not. In this plot, you've facetted by the stimulus type but I think it makes sense to facet by the category level (the comparison of distribution is done with the census estimate directly, and I think the interest is to compare the distribution across stimulus so you want the primary factor of interest to compare to be close in proximity). Within a category level, the census estimate is constant across stimulus so you could substitute is with a horizontal line (instead of a bar or point) as well.**
- **Figure 4: I think (b) alone is sufficient as you can gain the same information as (a) (although perhaps it is less familiar for wider audience). You could arguably put figure (a) information in a table instead.**.

XXX HH: What should we do here? I'd be OK with putting all of this into a table.

- Rephrase needed on page 18 third paragraph "… asking directly asking …".
  - *We have removed the extraneous 'asking'.*
- "Our study results provide actionable insights … in the design of data visualisations." -> This suggests wider scope than what the results suggest. I think the authors should say "design of stacked barplots" instead.
  - *We have updated the language here to say 'design of stacked bar charts'.*
- In the source code of the paper, authors use 1.96 and 1.645 for 0.875 and 0.95 quantiles but I think the authors should use `qnorm(0.975)` and `qnorm(0.95)` (I realize it doesn't make much difference in the end but I think you should avoid rounding unnecessary until the presentation of the results).

    - *This change has been made in the source code.*
  - **I do appreciate that the authors present the results using vertical stacked barplot, reflected from their results. A minor point is that I actually don't like the angled display of the x-axis text and I would have preferred to present this in a horizontal display so the text stays horizontal. I realize this sacrifices partially the accuracy of the results for something more aesthetically pleasing for peripheral information. But this opens up another question: is communicating information about always presenting the most accurate results? Communication won't happen if the recipient isn't taking in that information in the first place and while it may be superficial, aesthetically pleasing display may perhaps drive *better engagement*. In that sense, horizontal display where axis text can be better presented may be a better alternative? Authors could perhaps include a discussion along this line and I'd be interested in knowing what the authors think/recommend.** HH: figure 6: changed labels to horizontal I'm not sure how much we can really go into liking something and engaging with charts. Signals from aesthetic changes are usually quite small and people like different things. Finding aesthetic changes that suit everybody is rare.

Editors

- The AE and I do believe this is within the scope for this special issue, contrary to the final comment from Referee #1, so no need to address that comment.
- **Please add an unnumbered section Supplementary Material to list its content (e.g., data, code, additional simulation results, technical proofs, etc.). See recently published papers for examples. Reproducibility is held high at the JDS; the reproducibility team will check for reproducibility when the paper is accepted.**
- When addressing the comments, please keep in mind that there is a 20-page limit (including everything) in the production style.
- **In general, both reviewers express some concerns regarding some of the language and description regarding overall impact of the work. I would encourage the authors to make it more clear that their primary contribution is in using a representative probability sample, and that their conclusions are focused on stacked barplot displays, as Reviewer 2 specifically remarks, and highlight why conclusions from a non-representative sample may be problematic within the data visualization world.**
- **I believe the figures are fairly clear (per Reviewer 1's comment), but the authors might also be inclined to be more selective with what is shown in Figures 3, 6-8, and maybe move some of the others to an appendix.**

- **Can the authors comment further on modeling assumptions, particularly in relation to presenting all participants with the unaligned plot first followed by the aligned plot second (as opposed to randomizing this)? For instance, perhaps some participants notice that the plots are the same, with just the bars stacked in a different order, and their response to the second task is biased to be equivalent to that of the first task. This seems like it would be more of a problem if the plots were presented in the reverse order, but I'm curious if the authors looked into this.**
  - *We acknowledge that this is a limitation. There is a chance that someone recognizes they are completing the same task with the same bars when we present the aligned task before the unaligned task; we made the assumption that the chance of recognizing the tasks is lower when seeing unaligned prior to aligned. Participants are not able to return to the prior question once they move on to the second question, which does make it more difficult for participants to compare the tasks to one another.* **We will also add some text to the paper to reflect this limitation a little more clearly**
- **Is there a reason why the meaning/values of i from 1 to 3 are reversed between the two models specified in Equations 4 and 7?**
- **It appears the cumulative logistic regression model with respect to participant demographic covariates just considers first-order effects, but interactions would also be pretty interesting here. Are there any notable interaction effects?**
- **The letter notation used to denote significance is a bit confusing to me (e.g., Figure 5, Table 3), can the authors clarify this within the text or modify accordingly? What do a, b, c, etc. actually mean? If this is convention from Piepho (2004), it would be useful for this to be explained more or changed for a more general audience.**