Response to Reviewers, Revision 2

Grammatical fixes:

- Intro paragraph 2 line 1: "…almost a century, ["and" or replace comma with a semicolon] as new ways of representing information evolve"

  Replaced with semicolon

- Page 4 line 2 has an extra line break.

  fixed

- Section 2.3 (page 6): "comprised of 3 blocks (presentation mode), 7 bar ratios" -> replace comma with "and"

  fixed

In addition, clarifications to the following would be helpful:

- Section 1.2: "it is reasonable to reconsider the use of 3D charts, not only because of new technological developments, but also because 3D printed charts provide the opportunity to make data graphics accessible to those with limited or absent vision." A great point! However, this struck me as a bit out of place as justification given the paper focuses on comparison of visual perception, and a study of 3D chart accuracy for this population would require different methods.

  Added language to clarify that the use of 3D charts for those with limited/absent vision would be a possible future step

- All images of the Shiny app need to be enlarged. They are too small to interpret as-is, and too pixelated to read when the reader zooms in.

  fixed

- Please clarify in Section 2.4 how many total responses were collected with this experiment design, and then in Section 3 how many were left after removal of 33 responses and 1 trial. The AE expresses confusion about this, as well; there's lack of clarity around the difference between "response" and "trial" such that even with a few interpretations, I could not seem to identify the correct denominator to replicate the percentages cited in Section 3.2 (page 12).

  Added total number of responses in results section. "trial" is now replaced with "response" for consistency when referencing participant submitted an answer. Number of trials removed was changed from 33 to 18 due to a small code inconsistency with participants who completed the experiment after SDSS. The percentage on page 12 is given by 42/477, where the denometer is the number responses after removal of responses that were incorrect with the shape. In Results, cleared up language regarding total number of responses after invalid responses are removed.

- In data collection, I wondered if there was concern about participants entering the wrong kit ID or selecting the wrong 3D chart ID, and it appears this happened once with an invalid 3D printed chart identifier. How did that happen if the ID was from a drop-down menu (Section 2.5)? Might there be plans or strategy to mitigate this in future studies?

  This appears to be an isolated incident and anonymimity of participants makes it impossible to figure out exactly what happened in this case.

- Section 2.5 paragraph 3: "It should be noted that we chose to use a slider for entering ratio estimates - in part, this was intended to alleviate rounding effects with occur when participants enter numbers on their own or get numeric feedback from slider entries." What does "numeric feedback from slider entries" mean, and does that only apply when a slider has number ticks/scale?

  Added phrase to clarify that the slider we used did not have tick marks or numeric identifiers. We chose to remove these elements so that participants can't see the values they are inputting where they might attempt to round (e.g., seeing 41 on the slider might persuade the participant to move the slider to 40)

AE

- Just a clarification: The way Section 2.3 is written suggests that each participant in the study (48 total that completed it, from Section 2.4) was presented with 15 comparisons to make (3 blocks x 5 bar ratios), with an adjacent or separated task being presented to each one. Is this correct? So this would yield a total of 15 x 48 = 720 data points, for which 33 + 1 = 34 were removed for the reasons at the beginning of Section 3. Perhaps it may be beneficial to clarify this somewhere, since my initial thought was that removing the 34 responses was actually quite a substantial chunk of the collected data.

We added another layer of data cleaning to remove participants who completed the demographics screen but did not participate in the experiment, resulting in 38 participants. Participants were informed that they could stop at any time, so not everybody completed all trials (included this number in results section)

- Throughout Section 2, it appears the authors have done a good job of trying to replicate but also modernize Cleveland and McGill (1984); for instance, use of the slider as opposed to entering pure ratio estimates appears based on more recent results. However, the participant recruitment strategy described in Section 2.4 attempts to try to replicate that of the previous study, which almost seems like a bit of a flawed convenience sample (with some justification to include both people with more and less technical experience). Is there a specific reason the authors were motivated to match this recruitment strategy?

  This study was our first one in a series of graphics experiments. The pilot study gave us a first look at results and clear up any issues with experiment administration for future studies.

- I do think Figures 6 and 7 are still quite difficult to see, even when zoomed in.

  fixed

- Was there any exploration into the 33 responses that incorrectly identified the smaller bar? Even if this suggests a lack of attention, if there are any interesting patterns related to when these instances occurred, this could help suggest strategies for mitigating such errors in future studies.

  We did not explore the incorrect smaller bar responses for this study. This is a good suggestion moving forward!

- Did the authors test for interactions between display type and comparison type (both in the fixed effect terms and perhaps also the thin-plate spline specification) in the GAM in Section 3.2, or were these effects negligible?

  With the raw data responses, we did not see trends that would suggest interactions. We theorized that this particular study would be underpowered to detect any interactions, but this is something we plan to do in future studies with larger sample sizes.