Response to Reviewers

Editor Suggestions

- A more extensive literature review is needed; for instance, the authors claim on page 2 that "…the situation is different now than it was in 1984…it has even changed since 2014," but there are very few citations about this. I think even mentioning some application areas (with specific papers cited) where 3D charts and printed 3D charts are increasingly prevalent is important here to motivate why readers should be interested in this study.

  Fixed. It would be hard to cite printed 3D chart prevalence as there are very few applications of this at the moment (though there are a few US maps and things on Thingiverse). We expect they will be more common moving forward, though, which is why it's important to revisit the perception of these charts.

- Given that the authors claim this is a partial replication of Cleveland and McGill's study, I suggest creating a whole section devoted to more deeply explaining their work, strengths/limitations, and how the new study differs and improves on their work. This will help particularly for those not familiar with their study.

  We have added some limited additional clarity, but have also cited papers which go through the study in more depth, including the Annual Review paper "Testing Statistical Charts: What Makes A Good Graph" by Vanderplas et al.

- In its current state, the mixed model implemented in Section 3.2 needs to be vastly simplified; as both the authors and Reviewer 1 remark, this study simply has too few participants for this parameterized model to be sufficiently fit. At the very least, I would suggest fitting a simpler model and discussing this further (including assumptions), perhaps leaving a brief description of this mixed model as something of interest in a follow-up study or in an Appendix.

  The gam model we are currently using (after following up on Reviewer 2's questions) is a bit simpler but also much more interesting. We hope this model provides more useful results than the mixed-effects model in the previous version of this paper, which was pretty clearly not appropriate given the nonlinearity in the data.

Reviewer 1 Suggestions

- This experiment seems very far from a replication of the source material. This is somewhat softened in the paper by calling it a “partial” replication and acknowledging many of the shortcomings of the study, but even with this: the statement seems inaccurate. My reasons are as follows:
  - The participants were allowed to practice, with the true feedback given. Was there such a practice round in the original study? This seems like something that would improve later performance.

  Participants were practicing with the slider mechanism, but no feedback was provided as to the correctness of the slider position. During beta testing of the survey, we noticed that some participants were defaulting to estimating A/A+B instead of A/B, and so we added a clarifying demo page.

- A sliding error bar is a significantly different mechanism than simply stating a ratio. The experiment in this paper is, in a sense, asks participants to translate from a visual domain to a second visual domain. If the aim of the former visualization is to get to the latter visualization, why not present the latter visualization in the first place? What information would we like someone consuming a graph to get? This paper seems to want to comment on which medium best communicates the true numbers themselves, but what is tested is whether someone is able to translate one visual medium into another visual medium. It is the authors who then translate from this latter visual medium to the ratio.

  We agree that a sliding bar is different from estimating a ratio - the latter requires both the visual assessment of the ratio as well as translating this visual assessment into a number. We had hoped to reduce cognitive load by foregoing the numeric translation step; an added benefit to this is that it is typically very difficult to model numerical estimates from participants due to rounding effects that are censored but are not censored uniformly. We do have data from a different study suggesting that estimates using sliders are similar to numerical estimates (minus the rounding effects) - this paper will be submitted within the next 6 months or so but is not quite finished. At any rate, while we agree that the method for recording data is different (and that this might be an explanation for our failure to explicitly replicate the effects in the original paper), we are also working with very different technological constraints. The charts in Cleveland & McGill had axes which could be used to help participants with estimation; our 3D printed charts do not, which is one very real limitation of that medium. Because of this issue, as well as the cognitive load and technological considerations (e.g. we have the option of a slider, in 1984 that was much less feasible), we stand by our decision to deviate from the original protocol.

- Why are the data where a participant incorrectly identified the smaller bar removed from the study? Did they do this in the original study? Is this information irrelevant? How many observations were removed?

  The original study is not particularly specific about the methods employed. Attention check questions are fairly standard in online studies, and we have added some language clarifying the number of observations removed. Fundamentally, if participants don't identify the smaller bar, it is not clear what the slider ratio would even represent, as it is fundamentally confined to $[0, 1]$ and the participant isn't certain about what the denominator of the ratio is.

- The original experiment also uses stacked bar charts; this experiment only compares grouped bar charts. With these in mind, I recommend that this paper not argue that this experiment is in any way a replication, but a completely new experiment inspired by previous work. This will require that the authors acknowledge the poor sampling techniques used as their own.

  "Partial" is in this case a reference to the fact that Cleveland & McGill's 1984 study was an extensive assessment of many different simple graphical elements; we are specifically focusing in this study on the simple element of comparisons on an aligned scale (e.g. grouped, but not stacked, bar charts). When extending C&M to 3D, it is necessary to focus on one simple element at a time, if only because of technological constraints - the technology exists (but is extremely expensive) to print multiple color objects that would be needed for stacked bar charts, and it is entirely unclear how one would 3D print a scatterplot. While we agree that there are not very many 3D-printed charts at this point in time, it does seem to be a promising area for visualization and making data tactile, but we have to lay the groundwork first. We hope to follow up with a 3D heatmap style plot, where the 3rd dimension is not redundant, but comparing to existing studies seemed wise.

- There are details missing in the description of the experiment. Are the ratios participants are asked to estimate the same across the three plot types (2D, 3D, and 3D printed)? Are all plots compared either sepa- rated or not? How many times is a participant asked to estimate a ratio? There are a lot of missing details here. I would recommend an ordered list of tasks a participant is asked to do that emphasizes what variables (true ratio, separated/not separated, graph dimension) are controlled or varied at each step

  We have revised Figure 2 and the experiment description to clarify the design.

- The linear mixed effects model does not represent a thorough statistical analysis of the data. This is true for several reasons. The model seems very heavily parameterized for such few data. With my confusion described in the previous point, it is not immediately clear to me how many fixed effects there are, but it seems like there are 26 parameters (two s.d.s from the random effects, 7 ratios, 2 graph types nested in 7 ratios (7*2), and 3 graph types) to fit using a small dataset. This is too ambitious of a model. Furthermore, the normality assumptions on the error term and the random effect are not assessed in any way. I would recommend a simplier model with a proper assessment of the model fit. Examining p-values is meaningless unless the model accurately emulates the data generating mechanism.If a linear model seems inappropriate for these data (which seems likely without a random sample), the authors may want to forsake p-values and consider a more Machine Learning approach.

We have revised the model and are now using a generalized additive model with spline terms for ratio effects fit by comparison type and display mode and a random effect for participant. While this model has more flexibility, it does have fewer parameters and effective degrees of freedom than the previous model, and examination of random effect terms and model residuals suggest that the model fits reasonably well.

- Please clearly state the hypotheses before calculating p-values.

  We do not include model fit statistics for hypothesis testing purposes, but as a mechanism to graphically examine the components of the variance. While p-values are technically provided in the model output, we primarily include these tables in order to provide information with which to interpret the accompanying figures.

- The paper implies a consensus from existing literature, but the literature review presents several conflicting viewpoints. In the abstract, the paper claims that numerous studies advocate to avoid 3D visualizations when the third dimension does not convey any additional information to the user, yet it is never explicitly stated which studies claim this. The papers mentioned in the second paragraph of 1.1 comment on accuracy and encoding information (with conflicting conclusions) and no mention is given of any "consensus on 2D v. 3D." This paper would benefit greatly from a more focused literature review.

  We have improved the literature review; while the "3D is bad" advice is almost a trope in the graphics world, it is reasonable that people who are not visualization experts may not be familiar with this advice – it is so common that it is even addressed by authors of packages meant for 3D graphics, such as rayshader.

- Figure 6 seems to imply a significant confounding variable that is neither addressed nor investigated. It would appear that there is a drop in the midmean of the log error when the true proportion is around 50%. With this observation, I find the argument that certain proportions are easier to distinguish than others quite compelling. This should be investigated. Furthermore, if this is true, the graph that a participant randomly selects would have a strong effect on their ability to determine a correct ratio.

  When we switched to using the generalized additive model, we spent some time investigating this effect. This is a fairly common result in perceptual experiments resulting from anchoring at endpoints and the midpoint of the line; it does not seem to result from the default slider value of 0.5. This is an interesting finding that merits a follow-up study.

- Abstract: As mentioned earlier, the ’numerous studies’ are not explicitedly listed anywhere.

  This has been addressed

- The acronym EPTs is introduced, but never used. Perhaps most blaring: the paper fails to use it immediately after introducing it (at the beginning of the next sentence).

  This has been addressed

- What is Mechanical Turk? Please provide a short explanation.

  Added a short explanation

- Relative language such as “often” and “largely” detract from the point the paper is trying to make, especially when this point does not seem to be backed up in the literature review. For instance, each of these words appear in the first sentence of Section 1.2.

  Some of this relative language is because it is difficult to make absolute statements about charts - certainly some people create charts in 3D (as we have in this study), but the vast majority of visualizations are intended for display on flat, 2D screens or papers. Similarly, we cannot guarantee that perceptual heuristics will always be applied in certain cases - for instance, people who have extreme amblyopia or have been blind in one eye since birth are immune to the line-width illusion, while those with binocular vision cannot overcome it. There are very few absolutes in perceptual research. We have attempted to clarify the language, but it is very difficult to remove all relative language here.

- Also on the first sentence of 1.2: I have failed to understand this point. What is the visual system that is “assuming” something? Is it a computer? Our eyes? Is there evidence that chart perception is so affected by this assumption?

  We have cited papers indicating that chart perception is indeed affected by the human visual system’s heuristics. We have also added language clarifying that we are talking about the human perceptual system. In addition, addressing the line-width illusion using an example allows us to directly provide evidence to the reader that these effects are real.

- Please give a short explanation of the line-width illusion.

  Added.

Do the papers at the end of the first paragraph discuss the specific idea that 3D perceptions affect perceptions of 2D objects? Or do they just talk about these specific illusions? If it’s the former: move them up in this paragraph. The beginning of this paragraph feels like something that the authors are saying without evidence, as written.

> I believe this has been addressed because we re-wrote the paragraph explaining the connection between the line-width illusion and 2D/3D perception. Vanderplas & Hofmann (2015) in JCGS makes the case for the connection between 3D perception and the line width illusion in a compelling way; there is also an unpublished case study involving someone without binocular depth perception presented in Chapter 3.3 of https://github.com/srvanderplas/Dissertation/blob/master/thesis.pdf, if the reviewer is interested.

- second paragraph, first sentence: Did the subjects select these graphs for use, or did they better understand them? Are subjects here the people comprehending the graphs, or creating them? This is not clearly written.

  These short summaries of papers are all in the context of user studies, so these are designed experiments suggesting that participants prefer 2D charts for extracting information (that is, participants are presented with charts; they are not creating them).

- second paragraph: Why are the experienced subjects referred to as ’managers?’

  This study was evaluating usage of charts in a corporate setting - the experienced participants were literally managers within the company.

- second paragraph, second-to-last sentence, ’there are times where 3D graphs may better...’: Are these the same ’times’ when 2D graphs perform better? The two parts of this sentence could be interpreted as disparate ideas.

  As with most graphics research, there are always papers with contradictory results. We have tried to clarify the overall tone of this paragraph to assist with any confusion this may cause, but human subjects research is inherently messy.

- 1.2, second paragraph, last sentence: ’its’ should be ’their.’

  I believe this has been resolved

- 1.2, third paragraph, third sentence: ’has’ should be ’have’, unless ’Digital Graphics’ is the proper name for a field. In which case, capitalize ’graphics.’

  Fixed, thanks

- 1.2: Is this paper really ’reconsidering’ 3D charts? It does not sound like the literature really put them down.

We've done a better job at supporting the idea that 3D graphics have been denigrated in visualization literature for at least the last 70 years; they were blamed for misleading graphics in "How to Lie with Statistics" in 1954, and I would suspect there are older snide comments, but they get harder to find in the pre-digital era. While 3D charts may be acceptable in many fields, in graphics and visualization they are very frequently condemned, mocked, and denigrated, perhaps without sufficient evidence.

- 1.2, 'these charts provide the opportunity...': Make it clear that this is in reference to the 3D printed charts specifically.

  Fixed

- 1.3: Briefly discuss Cleveland and McGill's theory, or do not mention it.

  Cleveland & McGill's papers produced a hierarchy of simple graphical elements, and while they also produced various theories, in this case, we're not actually citing any of their theories. In any case, I believe the language that triggered this comment has been removed in the edits.

- Figure 2: What's the difference between Type 1 or Type 3?

  This has been clarified in the introduction, and we have used adjacent and separated as terms throughout this paper because they are a bit more intuitive.

- Discuss, or at least mention, Figures 2 and 3 in the text.

  Done

- 2.2, last paragraph, first sentence: What is covered in these papers? Are these two methods for digital rendering techniques? How are they different (and why does one require 'integration' into another)?

  These are software packages used in this work - Shiny creates the web applet used to collect data, WebGL provides the interactive 3D renderings to allow us to show 3D plots within the browser.

- 2.3: Is the ratio the same across graph type? How do we know that participants didn't use information from another plot type to drive their answers?

  Yes, the ratio is the same. As this was a relatively informal sample, we did check with participants verbally to see if they had any comments or had observed any interesting design quirks; most had no idea that the ratios were repeated. If statisticians didn't notice, we are fairly confident that their roommates who are not statisticians were also not aware. Examining the data, in 3/164 cases of (participant x ratio) trials participants had 0 variance on the provided estimates; in every case, these trials were ratios adjacent to 0.5 where participants did not adjust the slider. While this doesn't completely confirm that participants *couldnt* use information from other trials, it suggests that it was not a significant factor.

- 3.2, first paragraph, "(through given the midmean...": As written, this sounds like you are not expecting to find any log absolute error plots.

  This language is no longer present

- 3.2: Give more details on this Heer & Bostock model. Is there president for using a linear mixed effects model in this case.

  We've moved away from the LME towards a gam model (for which there is no precedent); if we could get the data from Heer & Bostock, we could examine differences between their model/data and ours, but that data is no longer available and we have reached out by email but have gotten no response.

- 4, first sentence: I don't think that your literature review really suggested this.

  Addressed earlier

- 4, first paragraph, last sentence: What discrepancy? The one between you in the literature? If so, you really need to argue a stronger consensus in the literature, if there is one.

  Addressed earlier

- 4: I wouldn't highlight that this paper does significance testing without more careful modeling.

  Addressed with better modeling and a bit more caution

- 4: "measurement method" should be "the measurement method."

  Fixed

Reviewer 2 Suggestions

Thank you so much for your careful comments and suggestions - they have greatly improved the paper and we appreciate the thought you put into this.

There are a few edits that I view as necessary to strengthen the paper. These edits will help clarify important issues for the reader, and also help avoid pitfalls.

- I recommend adding a brief explanation and refernce for the line-width illusion in section 1.2. This would help readers unfamiliar with graphics research.

We've added a figure demonstrating the sine illusion (as the simplest case of the line width illusion) and discussed in more depth the connection between this illusion and 3D graphics.

- At the end of section 1.2, it would be useful to add references, if there are any, for the comments of accessibility.

  We've added a paper by Chancey Fleet, who is blind and runs the NY Public Library lab for making visualizations available to the visually impaired. Hopefully that is sufficient.

- Section 2.3 could be clarified. Undertstanding the exact design with only the words in section 2.3 was difficult. Figure 2 helps, but was not cited in the text. Please clarify this a bit. This will also help the reader understand the model later, since the design is linked with the G(R) inclusion rather than having main effects and then an interaction.

  We've added a diagram which we hope will make the design much clearer. In addition, we've simplified the model so that hopefully it is easier to understand how the model works relative to the design.

- In Figure 6 the default smoother in ggplot2 was used. This accentuates the dip around x = 0.5, and allows the smallest and largest x values to have good bit of pull on the smoother. Since these features don't seem to be important to the narrative, I recommend an adjustment be made.

  We have adjusted the figure to use the underlying data instead of the midmeans; this ensures that the effect is not just due to the default smoother (though we understand where that perception came from). In the process, we decided to use a gam model instead of the linear mixed effects model to account for the clear nonlinearity of the ratio relationship to the error. This has made the paper much more interesting (in our opinion), so thank you for calling this out, because it led to a thorough dive into the data that was quite rewarding.

- Pages 7-8. I recommend avoiding the phrase "testing for significant differences", since this could be read incorrectly as not testing hypotheses set a priori (by someone rushing).

  I can't find that phrase anymore, but I hope that we've been sufficiently clear in our description that we're using the model in an informative/exploratory context more than as a means to do hypothesis testing.

- Page 8, bullet 1, I recommend acnowledging the full subscript in words

  I'm not sure what this is referring to, as there are no bullets on Page 8 that I can see.

- Details are needed as to what tests you ran from the LME. Only reporting p-values is not informative enough.

  Fair, but no longer relevant with the gam model. We've included model tables for informational purposes, but we're not really using them for significance testing (which is just as well since very few effects were significant, and even those effects need follow-up studies to ensure they aren't just an artifact of the experimental design and model-fitting process).

- In section 4, I found myself wondering if you would tell whether a participant interacted with a plot in the shiny app. For example, could you track whether they clicked or dragged?

  Yes, we can, and we've added a section about this. We were surprised by how few people did actually interact with the 3D rendering; this may be a case where we need to add a 3D render to the tutorial so that people can learn how to play with it before the study. Thanks for the suggestion! In more recent studies, we've actually taken to returning the rotation matrix on each interaction and we're looking forward to trying to analyze that data!

- Section 1.1, paragraph 2: "that 2D graphs" should read "than 2D graphs"

  Fixed

- Page 2, paragraph 2, "Digital graphics has" should be "Digital graphics have"

  Fixed

- In section 1.3, I would continue to use the Cleveland and McGill (1984) citation style. That's a stylistic preference, however.

  Fixed

- Page 4, paragraph 1. The citations Murdoch and Adler (2023) and Change at al. (2023) should be changed to rgl and shiny with parenthetical citations.

  Fixed

- Page 4, end of section 2.4. Change "3" to "Figure 3".

  Fixed

- Figures 4 and 5 were somewhat hard to read. Could these be clearer?

  If these are the app screenshots, I am not sure how to make them clearer while still meeting the image resolution requirements for the journal, but we have tried.

- Section 3, how many cases were discarded? A sentence or two would avoid readers asking this question and getting distracted.

  Fixed, thanks. Good catch.

- Page 9, first line “impact of measurement” could be “impact of the measurement”

  Fixed

- In the supplemental materials, please change the relative file paths in _code so the user could start in that main folder and not have to change the paths to rerun the analysis.

  Fixed, thanks.