Response to Review

We thank the reviewer for their careful notes about our manuscript. We have addressed many of the issues identified; in some cases, we defer to Dr. Jurgen Symanczyk, who asked that we e.g. not focus as much on statistical lineups because there is at least one other paper slated for WIRE on that topic which will provide much more detail.

Below, please find the point-by-point responses.

- The title refers to "statistical graphics", however in text authors use "charts". In fact on Page 1, line 27: Authors write "charts and data visualization" but why the distinction?

  In our experience, graphics, charts, and visualizations are used nearly interchangeably, and we often rotate between them to ensure that no single word becomes semantically satiated.

- The authors call the visualization lineup protocol as "statistical lineups" but is this really a good name for it? It has not been called as such before and I don't think it is a good idea to give unintuitive names (it will confuse future research in the area). If you must give it a name, I would prefer "visualization lineups" but it should really be defined, e.g. "In this paper, visualization lineups refer to …".

  The term statistical lineups have been used throughout several papers, including Loy and Hofmann (2013),Vanderplas and Hofmann (2016),Loy, Hofmann, and Cook (2017), Vanderplas, Cook, and Hofmann (2020), Vanderplas et al. (2021), Li et al. (2024), Robinson, Howard, and Vanderplas (2025). We have added some text explaining the origin of the term.

- The tone is sometimes casual, e.g. Page 8, Line 33: "system is very, very good", Page 13, Line 5: "a combination of density and rug geoms". These sound colloquial and it is not of an academic convention to write things like "very, very" and "rug geoms".

  We have addressed these language issues, but note that 'rug geom' is a convention from the grammar of graphics as implemented in ggplot2, describing the type of chart. We've adjusted how we refer to this to clarify that it is a composition of two different plots which show both the aggregate density and the individual observations.

- Citation style issues. E.g., page 8, Line 3 "Lu et al (Lu e al, 2022)" should be just "Lu et al (2022)".

  We've fixed these (whoops) and also the citation issue noted further below (shah?).

- Being specific, e.g. Page 1 line 43: be explicit that "test charts *with human evaluation*".

  We've used slightly different language (test the visualization on humans), but have added this specificity.

- I would appreciate a more informative caption for Figure 1.

  The caption now reads: "Direct adjustment of a plot in a perceptual task. In this experiment designed to assess the strength of the sine illusion, the user adjusts the plot using - and + buttons, which control the strength of a transformation designed to correct the effect of the sine illusion. When the user is satisfied that the lines are of equal length, they select the 'Finished' button to move to the next task. The experiment used a psychophysics experimental design, the method of adjustment, but leveraged the interactive Shiny interface to record the entire sequence of adjustments made by the user for each trial. A demo version of this application can be found at https://shiny.srvanderplas.com/sine-illusion/."

- The description of the lineup itself and its role in visual testing is somewhat abstract. Figure 5 might need to be introduced earlier for better clarity. Similarly, no explanation to how the see-value is calculated – if this is expected knowledge, could authors give recommendation of what literature readers should read to know the appropriate background, otherwise I think it is useful to make this guide self-contained and describe it briefly.

  We specifically were asked to not go into depth about lineups because of a pair of other articles which are intended to be published alongside this article. We have added the appropriate references so that readers can find the calculation details there. We have also included some additional language (which can be removed easily if not desireable) referring readers to the companion article about lineups.

- Page 1 line 46: the concept of internal and external validity is not explained. This might be common knowledge in psychometrics or psychology but I don't think it is in statistics.

  Well, that is a pity, as the concepts are broadly applicable to many different areas of statistics and were part of my statistical education. In any case, the language has been rephrased to refer to generalizability (external validity) and experimental control (internal validity), which hopefully will capture not only statisticians but also the InfoVis crowd made up of e.g. computer scientists and psychologists.

- Is there a guideline or reference for "think-aloud" processes? It's a concept that not a lot of statisticians would be familiar with so it would be nice to know more about this and the benefits of this approach.

  We've added several additional citations for think-aloud processes and explained both the informal use in experiment testing and the formal use in data collection.

- Should you assume your audience knows what a "PHP" (page 11, line 18) is? Same goes for Qualtric surveys, Amazon Mechanical Turk, Prolific and Reddit. These have no reference/link and little to no explanation. Given technology platforms can evolve rapidly, perhaps authors would like to give more context so that their paper can be understood by future readers. E.g. hypothetically suppose if Reddit disappears today then researchers in 20 years may not know what Reddit is.

  I have attempted to explain these sites in context wherever possible and link to relevant information. I have emphasized that Reddit is a social media site alongside X, Mastodon, and Bluesky – while we have not yet attempted to pilot studies on Mastodon/Bluesky, those are the natural replacements for X now that it is becoming less popular. I have screenshots of a recruiting message on X as well as the reddit.com/r/samplesize homepage, which could be included in the paper but seem slightly distracting. I have linked to an Internet Archive of the samplesize homepage to preserve this information for the future, but X blocks archival sites, so a similar archive could not be created for that platform. If the goal is that the paper is still comprehensible in 20 years, I think this should be sufficient - MySpace and Blogger, which have been obsolete for 15+ years, still have wikipedia pages which could be used to understand the references.

- What is IRB? Page 10, Line 43: "make the argument to IRB that our research is exempt".

  We've added the acronym definition and some additional information clarifying the ethics review process for human subjects experiments, in response to an additional suggestion.

- Also NORC panel is not defined.

  This one is surprisingly tricky - NORC used to stand for National Opinion Research Center, but they got rid of the acronym and just named the organization NORC within the last few years. We've clarified that it's an organization like Gallup and that the basic goal is to have a panel survey rather than a sample recruited from a platform like Prolific/MTurk.

- Rasch models on Page 12, Line 39 not defined nor cited.

We've added additional citations for Rasch models, but as they're suboptimal relative to mixed models for these types of experiments, we hope that this will be sufficient, as we don't want to spend the space explaining a less powerful model.

- how many participants should be recruited?

  This is so heavily dependent on the experimental design and factors under investigation that it is difficult to offer guidance on sample size. In attempting to add information about sample size, we refactored the paper somewhat, differentiating the protocol development process from the experimental design calculations. Graphics studies are sometimes overwhelming because statisticians do not usually control both the subject matter expert decisions and the statistical design decisions, and separating these two out helps to ensure that power calculations and statistical design decisions are isolated (as much as possible) from protocol decisions like the number of trials per participant, which is typically determined based on fatigue.

- what number of plots should be shown to participants in a lineup?

  We have attempted to not make this paper specific to lineup studies, so we will defer this discussion to the visual inference paper that is also to be submitted to WIRE.

- mention of ethic approval? A reminder that people do need to have ethics approval lined up and linked in their study to use results for publication?

  A section on ethics approval has been added, though we have no experience with this process outside of the United States and so have tried to balance specificity (e.g. exempt research and expedited review) with the fact that many readers will be outside of the US and subject to different regulations.

- can authors comment on financial costs related to participant recruitment?

  Every platform is different and even within platforms, pricing schemes vary by the week or month. We've included some mention of considerations like participant completion time vs. number of participants with regard to platform fees, and have also included some general comments about the relative expense of panel surveys vs. online platforms vs. in-person studies vs. social media recruitment.

- any guide to ensure representative sample in participant recruitment? Do authors adopt particular method or strategies?

  We have added some additional references on the effects of participant sampling on results of studies, as well as the different characteristics of different platforms, though at least one of the studies we cite are published by those platforms, so there are conflict of interest concerns as well.

- what about filtering out obviously non-serious participants?

  A section on response validation and attention checks has been added

- Authors mention usage of Amazon Mechanical Turk, Prolific, Reddit and social media. Could they provide advice on which online recruitment platform to use?

  Additional discussions of the advantages of categories of recruitment platforms has been added, but we have tried to remain as factual as possible about different competitors within those categories, in part because this space is evolving rapidly and if mentioning sites like Reddit is a concern (as above), then recommendations for specific platforms would be similarly problematic. Some discussion of the evolution of platforms over time is included, though, to explain why Amazon Turk is less popular now than it was in the 2010s.

- any advice for the design of interface that participants use?

  This is heavily dependent on the choice of data collection software. Bespoke solutions like Shiny and PHP web applications can be infinitely customized, but platforms like Qualtrics and Google Forms are much more rigid in design. Any advice we could give about the interface would have to handle many different configurations.

- Any examples of qualitative questions asked?

  We have added a few suggestions for qualitative questions, and reorganized the qualitative analysis section to better emphasize the citation to Vanderplas and Hofmann (2017), which contains a qualitative analysis that complements the quantitative modeling.

- Page 1 line 36: grammar of graphics technically doesn't "classify" charts but just is a means to specify one.

  Yes, this is true, but the grammar also provides a convenient way to classify charts that separates the mappings from the geometric representation. This is particularly powerful in graphics studies, where different geometrical representations of the same mappings can be compared head-to-head. We've modified the language slightly in the hopes of clarifying what we mean here.

- Page 2 line 46: "target plot with signal" — but there can be a target plot with no signal right?

  We've amended the language to avoid this issue.

- "automatically transcribed using text-to-speech functions of large models" — why LLM particularly? What about the approaches that pre-dates LLM?

It is only with the advent of LLMs that these features have reached an accuracy where speech-to-text might be a viable option for think-aloud recording, but we've removed the mention of LLMs in any case.

- Page 6, line 40: what is "(shah?)" supposed to be?

  That's a failed markdown reference that we have fixed. Thanks for catching it!

References

Li, Weihao, Dianne Cook, Emi Tanaka, and Susan Vanderplas. 2024. "A Plot Is Worth a Thousand Tests: Assessing Residual Diagnostics with the Lineup Protocol." *Journal of Computational and Graphical Statistics*, May. https://www.tandfonline.com/doi/abs/10.1080/10618600.2024.2344612.

Loy, Adam, and Heike Hofmann. 2013. "Diagnostic Tools for Hierarchical Linear Models." *Wiley Interdisciplinary Reviews: Computational Statistics* 5 (1): 48–61. https://doi.org/https://doi.org/10.1002/wics.1238.

Loy, Adam, Heike Hofmann, and Dianne Cook. 2017. "Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners." *Journal of Computational and Graphical Statistics* 26 (3): 478–92. https://doi.org/10.1080/10618600.2017.1330207.

Robinson, Emily A., Reka Howard, and Susan Vanderplas. 2025. "Perception and Cognitive Implications of Logarithmic Scales for Exponentially Increasing Data: Perceptual Sensitivity Tested with Statistical Lineups." *Journal of Computational and Graphical Statistics* 0 (ja): 1–14. https://doi.org/10.1080/10618600.2025.2476097.

Vanderplas, Susan, Dianne Cook, and Heike Hofmann. 2020. "Testing Statistical Charts: What Makes a Good Graph?" *Annual Review of Statistics and Its Application* 7 (1). https://doi.org/https://doi.org/10.1146/annurev-statistics-031219-041252.

Vanderplas, Susan, and Heike Hofmann. 2016. "Spatial Reasoning and Data Displays." *IEEE Transactions on Visualization & Computer Graphics* 22 (1): 459–68. https://doi.org/10.1109/TVCG.2015.2469125.

———. 2017. "Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics." *Journal of Computational and Graphical Statistics* 26 (2): 231–42. https://doi.org/10.1080/10618600.2016.1209116.

Vanderplas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. "Statistical Significance Calculations for Scenarios in Visual Inference." *Stat* 10 (1). https://doi.org/https://doi.org/10.1002/sta4.337.