# Response to Reviews: JCGS-23-139

Weihao Li, Emi Tanaka, Dianne Cook, Susan VanderPlas

2023-08-07

Thank you for the careful reviews. What follows is our point-by-point response to reviewer comments.

Note that *The original reviewer comments are in italic* and our response is normal text.

## Reviewer 1

*Comments to the Author*

*I did not need any convincing that residual diagnostic plots are superior to tests but this paper presents a narrower argument to support this claim than many analysts would employ. For example, diagnostic plots may reveal features which are not directly connected to a linear model assumption and yet are interesting and important to identify. Of course, it is difficult to justify a journal article on the basis of such observations. In this respect, I am impressed that you have succeeded in demonstrating the superiority within the more formal framework of the lineup protocol. You have restricted attention to the residual-fitted plot and a particular selection of formal tests. I accept that this was necessary to make definite progress. Nevertheless, it would be good to add some discussion of the wider picture.*

Thank for this feedback. We are happy to see that, like many people, you don't need convincing and that you are glad to see that we have succeeded in demonstrating the superiority of residual diagnostic plots.

We have added more explanation about the use of residual plots in the Introduction (highlighted in red) and we have added a new section "Limitations and practicality" that provides some insight into the wider picture.

*1. "The ubiquity of this advice is curious: investigating why, is the subject of this paper." I don't find this advice curious - it seems almost universal. I have taught this topic for many years and give the same advice. I explain why I give this advice and I believe several other authors do the same. Their reasons are wider in scope than you express in this article.*

We have changed the language.

*2. All your examples are simulated data and very obviously so. Real data shows much greater irregularity - the points will usually not be uniformly spaced along the x-axis. The response may be somewhat discrete or have other features that will distinguish it from bland randomness. This would seem to offer greater challenges to your lineup idea as generating data under the null will be much more challenging. If we simulate from the null model, the real data will be very obvious. You have done an experiment but the reality would be more like an observational study. This limits the scope of your conclusions which you should clearly acknowledge.*

The work very specifically uses simulated data because we need to study a problem in controlled conditions. You are right, though, that data will pose more interesting challenges, and this can be in two directions. One, being that the lineup protocol will allow interesting and unexpected patterns in the residuals to be detected (see Buja et al and Wickham et al for examples). The other is that the current process for generating null data uses residual rotation and will likely maintain irrelevant quirks of the data in the null plots. These should not inadvertently allow the reader to detect the data plot based on quirks.

We have addressed this criticism in the broader discussion of the use of residual plots and lineups for residual diagnostics added to the new Section "Limitations and practicality".

An observational study involving actual residual plots will hopefully be possible once a computer vision model is deployed that will allow analysts to upload their residuals for automated testing.

*3. It is true that formal diagnostic tests tend to be too sensitive and yet we don't have a definite sense of how much divergence from the assumed model would cause a significant problem. How much non-normality in the residuals is acceptable? We don't really know. Formal tests give us definite answers about violations of assumptions which we might find hard to ignore but diagnostic plots give us license to sweep inconvenient features under the carpet. Is there really an advantage to the diagnostic plot in this case other than the opportunity to hand-wave about our answer?*

When the diagnostic plots are embedded in a lineup of nulls, it does have the advantage of providing a calibrated answer, and protects against the subjectiveness of hand-waving. We are advocating that you should always calibrate the diagnostic plot with a comparison with null plots. We have strengthened this point of view by adding more explanation in the Introduction, in addition to what is provided in the Conclusions.

Having a definitive (numerical) answer to "how much is too much" would require more and different work. For our purposes, it was that the true residual plot could be detected from the null plots when the lineup was evaluated by K people. We've added a sentence to the Conclusion mentioning this.

*4. I use something like the lineup protocol in teaching to help students calibrate their assessment of diagnostic plots. But I doubt very much experienced analysts use them in practice. I do not think it is realistic or really that helpful to recommend their everyday use.*

It can be argued that it is being used by practitioners based on download rates of the nullabor package, and applications articles that report usage. Using it formally, we agree is cumbersome and would require engaging a crowd-sourcing service, that involves a small expense. Science labs typically have a budget for equipment and experimental devices and we would advocate this is a useful addition. We have definitely used it ourselves in 10.1007/s00180-014-0534-x for ecological data and 10.4172/2153-0602.1000139 for gene expression analysis. We have added some details in the "Limitations and practicality" section.

# Reviewer 2

*Summary*

*This article uses the lineup protocol in a visual inference experiment, which puts a residual plot in the context of null plots. Focus is mostly on detecting nonlinearities and heteroskedasticity. Comparisons with some classic tests are given.*

*Overall Comments*

*Overall, this paper is very well written and interesting. However, I have some serious reservations about the practicality of the procedure, which I articulate in greater detail below. I also think the paper is currently too long, and that a significant chunk of material should be shifted to the Appendix.*

Thanks for the positive feedback. We have shortened the paper by XXX pages. We have expanded the explanations on the practicality in the Section "Limitations and practicality".

*1. As mentioned above, the paper is too long. While the onus should be on the authors for getting the page number down, some possible suggestions could be combining Sections 1 and 2 while moving some of the historical context (which I fully acknowledge is quite interesting) to the Appendix, and moving some of the technical details of Section 3.2 to the Appendix.*

Thanks for your suggestions. This is how we have shortened the paper:

- Details of Section 3.2 Statistical significance moved to the Appendix.

- Sections 1 and 2 have been made more precise, while still covering the requested additional topics.
- The explanation of power calculation was left in the main paper because this is new and of primary importance for understanding the results.
- Part of Experimental design -> appendix

*2. The practical implementation of this as a procedure is, in my opinion, the biggest weakness. The authors acknowledge the costliness of human evaluation of residuals, but not until Page 33, Line 47. I find this a tad disingenuous as this will likely jump out to the reader as being a problem right from the beginning of reading this paper (it did for me). So I think being upfront about this limitation, and providing some practical considerations to address this concern will assuage my concerns. There might be some concepts from human-computer interaction that could be effective for operationalizing and improving the efficiency of the present work.*

The new Section "Limitations and practicality" addresses these concerns and we have added two sentences to the Introduction.

*3. I think a better connection needs to be made with where this computationally expensive process could potentially pay off. For example, applying the lineup protocol to, say, any of the data examples in the texts cited by the authors would amount to nothing more than an academic exercise. But, what sort of data problems can the authors point to and definitively claim that the lineup protocol will be invaluable at ensuring we visually capture possible nonlinearities or heteroskedasticity? I am not doubting the efficacy of the approach, but I'd like to see a stronger case made with its practical application.*

See the new Section "Limitations and practicality". Some additional text is provided in various places. It's very practical to employ the lineup protocol using the `nullabor` package, and the (growing) number of citations from publications in a variety of areas show that it is being used.

*4. What should be claimed about the sample size n as a limitation? Take in point the comments at the bottom of page 19. Am I to understand that the lineup protocol will not be appropriate for big data problems?*

The lineup protocol is very suitable for big data. A plot other than a scatterplot is probably needed to better display a large number of observations. We have changed the language so that this is clearer.

*5. At the beginning of Section 4.2.3, what cultural biases do the authors anticipate by restricting the participants to be "fluent in English"? Granted this is likely done as a matter of convenience, but it would be good to articulate how cultural biases may be present in interpreting a lineup protocol.*

The instructions for this study were in English, which motivated the constraint applied in the crowd-sourcing service. However many participants were likely multi-lingual, and they joined from a variety of countries. We don't anticipate that the results would be different with a different sample.

*6. I would like the authors to be a little clearer about their specific contributions. I felt as if the biggest contribution in this work was the experiments that they did. A lot of references are included for work that is applied in this manuscript, so at times I got lost as to what already existed in the literature and what "new" material the authors were giving the reader.*

We have clarified the language of the Introduction and the Conclusion.

*Minor Comments*

*1. Page 9, Line 13: Since the three references are by the same authors, use \cite{Ref1,Ref2,Ref3}.* DONE

*2. Page 12, Line 7:* $\alpha$. DONE

*3. Why not write the power formula in terms of a generic significance level? Granted, one should be mindful of an abuse of notation if using* $\alpha$ *as it has been introduced in the Rorschach setup of the previous subsection.* We have kept the current explanation of power.

*4. Page 13, Line 14: "...the scenario..."* DONE

*5. Page 20, Line 36: "...variety of difficulty..."* DONE

*6. In Figures 11 and 12, I am not sure the example residual plots beneath the x-axis are all that terribly helpful. In fact, I think they all look mostly similar, and am not entirely convinced that they add any value. Can the authors convince me otherwise?*

The small residual plots help to understand the change in effect size represented by the numerical values on the x-axis. They are indispensable, as reported by numerous readers during the proofing stage of the paper.