Response to Review

Majumder, Cook, Hofmann

Thank you for the careful and positive reviews. *Our response to comments is written in italics.*

From the editor

Good set of comments. FYI, feel free to change the review title (if you wish) and add more material (as an on-line journal, page limits not really active). Look forward to seeing your revision. D Scott co-editor scottdw@rice.edu

Response: *Thank you*

I am writing in regards to Manuscript ID EOCS-756 entitled "Effect of Sociological Factors on Visual Statistical Inference," which you submitted to WIREs Computational Statistics. The reviewer(s) have suggested major revisions to improve the manuscript. Their comments are included at the bottom of this letter; I invite you to respond to them and revise your manuscript.

Please submit your revision within 6 weeks. If you need extra time, please let us know so that we can keep our systems updated for editorial planning purposes.

A revised version of your manuscript that takes into account the comments of the reviewer(s) will be reconsidered for publication. Please note that revising your manuscript does not guarantee eventual acceptance, and that your revision will be subject to re-review before a decision is rendered.

If you are updating your Main Document in Word, make sure that "Track Changes" is enabled so that the editors can easily find your amendments. If you are using LaTeX, make sure that revisions are made in red font.

You can upload your revised manuscript using this direct link: *** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://urldefense.com/v3/**https://mc.manuscriptcentral.com/compstats?URL_MASK=438fdf874e**
tU!GOtJWWHikD4k_JosJN5bkg2Kphby7E2KmkmqHJ4dZ95gyJOEY62FgPhrkDJhfcMHLuoagvllP182luVdjk
IXMjM$.

When you submit your revision, you must provide a point-by-point summary of how each of the issues raised by the reviewer(s) were addressed. Please note that the space provided is for plain text only, as the system will not accept any formatting. For each criticism raised, please provide: 1. The criticism from the reviewer. 2. The response and explanation of how the manuscript has been modified in response to the criticism.

IMPORTANT: We have your original files; therefore, please delete the file(s) that you wish to replace and then upload the revised file(s). You will be able to upload both marked and clean versions of your Main Document. Please note that figures must be in PRODUCTION-READY format.

Please do not hesitate to get in touch with us if you have any queries or concerns.

Once again, thank you for submitting your article to WIREs Computational Statistics. I look forward to receiving your revision.

Sincerely,

Hashini Selvam On behalf of Dr. David Scott Co-Editor-in-Chief and Review Editor, WIREs Computational Statistics scottdw@rice.edu

Reviewer: 1

Comments to the Author(s)

This is an interesting study where the authors attempt to examine the impact of demographics on evaluating a "line up" of statistical graphics to visually pick out a particular data set compared to several null graphs. The main results seem to be that while demographics do seem to vary the results from a statistical perspective (i.e. the p-values are significant, given the large sample size), the impact from a scientifically meaningful perspective is minimal. I have a several specific comments to highlight below.

Author response: *The authors agree with this comment. The results support our conclusion that the visual test is robust to variations in human factors.*

Specific comments:

1. The authors choose to examine two outcomes, log time (how long it takes the user to complete the task) and detection rate. It is not clear to me why a joint model is not shown. It seems relevant, particularly because there are clear demographic differences seen (albeit quantitatively small) that it would make sense to examine detection rate with the demographics *and* time in model A.1 to see whether, for example, the demographic differences seen in A.1 are fully explained by time as there are definitely impacts of demographics on time and it seems plausible that time would impact accuracy. Likewise, it is possible that after adjusting for time a learning trend may be more visible.

Author response: *A joint model is fitted to simultaneously analyze detection rate and log time taken. The estimated covariate effects in the joint model are largely consistent with those obtained from the separate models, with one notable difference: the estimated variance of the lineup-level random intercept for detection rate (lieup difficulty) is slightly higher in the joint model (2.37) compared to the separate model (2.27), suggesting that modeling the outcomes jointly may better account for heterogeneity across lineups. Importantly, the joint model also provides an estimate of the latent correlation between the lineup-level random intercepts for time taken and detection rate. This correlation is estimated to be -0.55, indicating that lineups which take longer to evaluate tend to be associated with lower detection rates. This result aligns with intuition: more difficult lineups likely require more evaluation time, yet are also more prone to non-detection. These findings and interpretations have been incorporated into the revised manuscript at the end of section 3.2.*

2. The authors state that the largest effect is by far the lineup difficulty. This mixture of difficulties makes it hard to learn much from the overall results (e.g. it makes the bottom portion of Figure 1 very hard to parse). Figure 2 (B) seems to show a the impact of a variable (education) across three different classifications of "difficulty" – from this plot it seems that the easy line ups and difficult line ups are providing very little information, as the proportion classified correctly is very close to 1 or 0 (at least for the 18-25 year old females this plot was drawn for). It may be more compelling to show the results excluding these and focusing only on the medium lineups (at least as a subgroup analysis).

Author response: *Lineup difficulties were predetermined as part of the experimental design, and each participant was randomly assigned 10 lineups spanning a range of difficulty levels. The fitted model accounts for lineup-level variation through random effects, effectively adjusting for lineup difficulty when estimating the effects of covariates on detection rates. For this reason, separate analyses by lineup difficulty are not the focus of this study.*

3. Building on the previous comment, it is possible (likely?) that the reason there is no "learning trend" seen is due to these very easy or very hard lineups. I would like to see a model for this that only includes the medium lineups.

Author response: *The model accounts for lineup difficulty when estimating the learning trend. It would be interesting to examine learning trend more deeply, in particular in relation to using lineups for teaching people how to read data plots. However, learning trend in itself is not the focus of this paper. Here we are assessing whether the human evaluation of lineups is consistent across demographic factors and trials. Each subject evaluated 10 lineups, at least three of each difficulty level. It is insufficient to only use one of these difficulty levels to assess the learning trend.*

4. It is interesting that while there are many significant (from a statistical standpoint) results, both the abstract and conclusion state that "demographics have very little impact" based on the coefficient size, which makes me wonder how the sample size of 2,340 participants was arrived upon. If a lower "effect" would be deemed meaningful, why was there such a large sample? It is a bit unsatisfying to have the results show statistical significance and then post-hoc have the authors state that the point estimate was too small to be meaningful – perhaps choosing a test that takes into account both the scientific and statistical significance would make this more compelling. For example, something like using an equivalence zone that is deemed scientifically unimportant.

Author response: *Agreed, it could be done differently, but this approach is simple and familiar. The statistical significance is meaningful and it encourages a discussion of effect size and practical significance.*

5. The authors state "Notably, countries like Canada, Romania, the United Kingdom, and Macedonia also had notable participant representation." – it is not clear what this means, does this mean there were many individuals from these countries? If this is important information for the reader, it is not easy to glean it or read from Figure 5 as shown (or even just adding the N next to each of these countries in the initial sentence, it just was unclear why this was included).

Author response: *This is simply to illustrate examples of countries from which we did not expect to receive much participants via Mechanical Turk. The first paragraph of section 3.1 has been revised to address this concern and clarify the discussion.*

6. I think it is instructive to layout a specific example (the 18-25 year old female undergraduate vs someone with some graduate course work), although I'm not sure the basic equation for the difference is necessary to include in the paper.

Author response: *We have removed the detailed calculation on page 6 and report only the final number.*

7. The authors state that participants who performed poorly on an initial test were excluded from the analysis - it would be interesting to see a table of the demographics of this excluded population compared to the included population, particularly as it is possible that researchers using "visual inference" may not always include this step (and perhaps it is crucial for the assumption that the demographic impact is small?)

Author response: *Online experiments can sometimes attract insincere participants who do not engage seriously with the task. To address this, an initial screening test was used to identify and exclude such individuals from the analysis. There were insufficient numbers to warrant additional analysis of this group.*

Reviewer: 2

Comments to the Author(s)

Summary: The paper investigates the effect of demographic variables on visual statistical inference and finds that demographic variables do not explain the variability of the main outcome, detection rate, well. Moreover, learning over time increases speed but not accuracy. I really like the paper. It is well written, I love the visual abstract, and the findings are strong, and for some findings effect sizes are given. In my opinion, the paper could be improved by changing the title: "sociological factors" is not a term I have heard before and it paints an entire discipline with an overly broad brush. Likewise, I would not argue anything about "cognitive skills" (abstract) given that you have no measures of cognitive ability. I enjoyed reviewing this paper.

Individual points:

1)Title: The title strikes me as odd. What are sociological factors? As far as I can tell none of the authors is a sociologist. The word "sociological" only appears in the abstract. I argue for a different title.

Author response: *The title has been changed in the revised manuscript. The new title is 'Effect of Human Factors on Visual Statistical Inference,' where the word 'Sociological' has been replaced with 'Human.'*

2) Abstract: "The effectiveness of this method, measured by power, depends on combining evaluations from multiple observers." I do not understand what this means. My association with "power" is sample size calculations; I am not sure what this means here. I see you are alluding to a definition of power in line 37 in a different paper (without explanation) and then define it on page 2 as the detection rate. I think the abstract would be easier to understand if you used "detection rate" or you explained what power means.

Author response: *The abstract has been revised to explain the concept of power and its relationship to the detection or identification of the observed plot in a lineup. Additionally, the introduction now includes further discussion to clarify why detection rate is used to assess the impact of human factors on visual statistical inference.*

3) Abstract: "Factors influencing power include observer demographics, visual skills, experience, the sample of null plots, and signal strength." This sentence does not move the narrative much. Also, it is not clear to me whether "visual skill "and "individual skill" in the next sentence are the same. The number of factors you list only partially matches the factors listed in section 2.

Author response: *The abstract has been revised to focus on the human factors examined in this paper.*

4) There were a high number of female participants (42%) across most countries This sounds a bit funny – do you mean in the sample there were at least 42% females for most countries?

Author response: *Overall, female participants make up 42%. The manuscript has been revised to clarify this.*

5) Notably, countries like Canada, Romania, the United Kingdom, and Macedonia also had notable participant representation I don't know what "notable participant representation" means.

Author response: *The manuscript has been revised to include a list of countries, other than the United States and India, where participants were located, highlighting the geographical diversity of the sample.*

6) Figure 1: I did not understand how you get a box plot out of an indicator variable for Detection (yes/no). May be it is not an indicator variable but an average of multiple detections? (there are 10 tasks listed in Table 6)

Author response: *The detection rate in Figure 1 refers to the proportion of times the observed plot is correctly identified in a lineup by each demographic factor group. It is a continuous variable ranging from 0 to 1. This measure is not derived from the 10 experiments shown in Table 6; rather, detection rates are calculated for each combination of demographic factors and lineups. For example, participants identified as male evaluate multiple lineups, resulting in multiple detection rate values corresponding to the lineups they assess.*

Figure 1 caption has been revised to clarify how the detection rate is calculated. The revised caption includes "proportion of data identifications (detection rate) across all lineups, aggregated for each level of the demographic factors."

7) Figure1: "Variability in detection rate indicates large variability in lineup difficulties." This is a very important finding. It took me a while to understand your argument: Maybe you argue because the variability is huge, it must be due to a factor other than the ones shown (demographics), and that factor must be lineup difficulty. I am not confident that this conclusion necessarily follows. Would it not be easier to extend the Figure by 3 columns with the 3 levels of difficulty? If you are right – and I believe

you are – then I would expect three much narrower box plots and the difference to the demographic variables would be more obvious.

Author response: *Lineup difficulties were predetermined as part of the experimental design, and each participant was randomly assigned 10 lineups spanning a range of difficulty levels. Figure 1 demonstrates that participants from each demographic group were exposed to lineups with varying difficulty levels. This is further illustrated in Figure 2(B), where three levels of lineup difficulty are shown more distinctly.*

The fitted model accounts for lineup-level variation through random effects, effectively adjusting for lineup difficulty when estimating the effects of covariates on detection rates. That is why In Table 1 we notice that lineup specific variability is estimated as 2.27 which is much higher than any of the other parameter estimates in model A.1.

8) Conclusion: page 10 line 7-10 You give significant effects. For India you mention the comparison group (USA), for age and for people with graduate degrees you do not mention it. It would also be helpful to give the percentage differences in parenthesis, e.g. "India shows significantly lower detection rate compared to United States (xx% vs xx%)".

Author response: *The Table 1 shows the baseline levels for each category and the other levels are compared with the baseline. Significance of detection rates are discussed in details in section 3.2 based on the estimated model parameters in Table 1. The conclusion has been revised as suggested and now clarifies the reported results by including detection rates for comparison.*

9) Discussion: I am in particularly intrigued by the finding that experience improves speed but not accuracy. I wonder whether other researchers have found something similar in a different context. This may or may not be easy to answer in a couple of sentences – so please use your judgment.

Author response: *The finding that experience improves speed but not accuracy in lineup evaluations is consistent with prior research in different contexts. For example, Hong and Williamson (2008) showed that participants trained on a simple psychomotor task exhibited significant improvements in all speed-related measures, while their error rates remained largely unchanged. This discussion has now been added to the conclusion section.*