Dear Seth Wiener,  
  
We write today with reference to your submission to LAB entitled "The Art of Wrangling: Best Practices for Reporting Web-based Eye-tracking Data in Language Research".  
  
Based on the recommendation of 3 reviewers (included herein, or, alternatively, available by logging in to the editorial manager, clicking on "action links", then on "view attachments") as well as our own reading of the manuscript, we would like to invite you to revise and resubmit the article. We find the topic extremely interesting, but we agree with the reviewers that there are some issues that we feel need to be addressed in a revision.  
  
As you will see from the reviews, all three reviewers had very positive comments about your paper, and think it will make an important methodological contribution to web-cam based eye-tracking research. Reviewer 1 suggested minor corrections and made a number of suggestions, while Reviewers 2 and 3 had some slightly more substantial comments. While we agree with the reviewers regarding the potential for your contribution, we also agree with the points made by Reviewers 2 and 3. In particular, the two points below are particularly important to address.  
  
(1)  
Reviewer 3 queries the paper's connection to the 'bilingualism' theme of both our special edition and the journal itself. While we appreciate that your contribution is methodological, and the advice you provide is applicable in various research contexts, it is also important for all contributions to the special issue to have a clear relation to bi-/multilingualism. We feel more could be done here. For the purposes of your paper, including discussion of how web-cam based eye-tracking can be useful for research in bi-/multilingualism broadly defined (e.g. perhaps accessing participant samples that would be difficult in 'in-lab' settings), and/or some discussion of any particular data wrangling issues that may be encountered in bi-/multilingualism research (e.g. the heterogeneity of participant samples), would suffice.  
  
(2)  
Reviewers 2 and 3 both query your description related to building an experiment in Gorilla, and worry that the discussion here is too restricted to this specific software. We agree here and believe it would be better if this section was written in a more general way, that would be applicable to researchers using a variety of different software environments to build their experiment. This is especially the case given how quickly web-cam based technology is developing. Indeed, as you yourselves note in Footnote 1, the terminology used in Section 2 refers to Gorilla Task Builder 1, while Gorilla now has a Task Builder 2. As you state, Gorilla's Task Builder 2 does not yet have eye-tracking functionality, but presumably at some point it will, and so it would be a shame for your advice here to eventually become out-of-date. If you reword this section in more generic terms, you will avoid this.  
  
As suggested by Reviewer 3, you can of course still link to your Gorilla experiments, as some readers are likely to find them useful. You could also include description of your experiment (as you currently do in Section 2 of the paper), as an online only supplement on the OSF, but we do think a more general discussion of web-cam based experiment building would be more appropriate for the paper itself.  
  
In addition to these main points, each reviewer makes various other comments that you should also address.  
  
Together with your revised manuscript, please send us a letter copying all the reviewers' comments, followed by an explanation of exactly how you have addressed each comment and suggestion. Please make sure that your manuscript follows the LAB stylesheet (<https://www.benjamins.com/series/lab/lab_style_sheet.pdf>). Manuscripts which do not adhere to these guidelines will not be accepted for publication.  
  
Please note that an invitation to revise and resubmit does not entail that the next version, or any subsequent version, will be accepted for publication. The revised version will be evaluated anew, and it is sometimes the case that manuscripts which are still judged not to be acceptable after revisions have been made are rejected. Additionally, the clarifications and/or changes in the revision may uncover new or additional issues that lead to rejection of the manuscript.  
  
To submit your revision, please go to <https://www.editorialmanager.com/lab/> and log in as an author. There, you can access the menu item 'Submission Needing Revision'.  
  
LAB requires you to make your materials and/or data freely available to enhance the visibility and reproducibility of your study. If you do this, your article will be awarded recognition from the Center for Open Science via an 'Open Materials' badge and/or an 'Open Data' badge which will be clearly displayed next to your article. This is a public kitemark, recognizing that you are practising open science. To qualify for a badge, your materials (and/or data) must be on an open repository that is recognised by the Center for Open Science, such as IRIS ([https://www.iris-database.org](https://www.iris-database.org/)).  
  
Thank you once again for your submission, and we are looking forward to reading your revised manuscript as soon as possible.  
  
Best wishes,  
  
The Editors  
  
Linguistic Approaches to Bilingualism  
  
-------------------  
  
Comments:  
  
Reviewer #1:  
The paper offers a comprehensive guide for web-based eye-tracking studies using Gorilla. As web-based eye-tracking is becoming more and more common, I believe a guide like this is very useful for researchers who are conducting web-based eye-tracking studies and researchers who are interested in it. Providing a guideline for reporting web-based eye-tracking results is also beneficial for the field. The paper is written in a clear and accessible manner, and the R codes provided in the main text offer a step-by-step guide for data processing. It includes basic R functions such as filtering, which should be helpful for R beginners. The differences between in-lab eye-tracking and web-based eye-tracking are summarised clearly. The provided replication data (e.g., proportion of data excluded) should give researchers a good idea of what to expect in web-based eye-tracking studies. The study used for the replication has a commonly used 2 by 2 design and seems to include a robust  
effect of predictability and a presumably less robust effect of talker identity and/or their interaction. This is good as researchers may be interested in knowing whether a relatively small effect or interaction found in in-lab studies can be replicated in web-based studies. The Shiny Apps nicely demonstrate how the data change depending on the analytic choices and are handy for researchers to decide their data filtering criteria before the data collection. Overall, I am very positive about this paper. I only have recommendations to further improve the paper.  
  
Note: I have never used Gorilla myself, so I cannot assess the experiment set-up part well (section 2).  
  
Major comments:  
  
- More detailed recommendations can be provided about the requirements for the participants. For example, the relative height of the participants' eyes and their webcam, the angle of the laptop/monitor, the angle of the light (front/back), the brightness of the light/room all seem to affect the quality of the data collected in web-based eye-tracking studies. It would be very helpful to add guidelines that researchers should provide to participants to maximise data quality.  
  
- The authors excluded data outside their window of interest at the last step. I suggest performing this data filtering at the earliest stage because this is useful for excluding subjects who have many missing data points, and subsequently, for reducing the size of the data set to handle. (For subject-based exclusion, we would like to check the proportion of missing data in the critical time window, not during the entire trial or recorded period.)  
  
- p.41. The authors recommend using Gorilla in-person to reduce potential participant loss. Although I agree this should maximise data quality, I believe the majority of the researchers using Gorilla are interested in online data collection. An alternative or additional recommendation would be to create a quick calibration test to filter out participants and only invite those who have passed this test for the main experiment.  
  
  
Minor comments:  
  
- p.6. "Visual stimuli are minimally made up of two types: targets and competitors". This is true in many studies, but there are also many studies that did not present any target (the target absent design, used in e.g., Huettig & McQueen, 2007, JML) or competitor (only the target plus distractors, e.g., in Altmann & Kamide, 1999, Cognition).  
  
- p.11. The x-axis and y-axis values in Figure 5 do not seem to correspond to possible values presented in Figure 2.  
  
- p.16. "…which participants should be excluded based on post-experiment questionnaire exclusion criteria (e.g., not an L1 English speaker and not between the ages of 18 and 40)?" This sentence should end with a period.  
  
- It would be helpful to provide package names for the functions used in the R code.  
  
- p.21. I find Figure 7 illustrating different types of data joining functions misleading because the left-join and right-join look like they are only selecting one of the data sets rather than joining two data sets.  
  
-p.25. "Whereas this remains true for competitor items…" There was no mention of competitors in the study description (section 2.1). I read the original study by Porretta et al. (2020), and their visual stimuli contained the subject (of the sentence), a target object and three distractors, i.e., no competitors.  
  
- p.31. Figure 10. The caption does not provide a full description of the figure. For example, what are high, medium, and low? What are the units for x- and y-axis?  
  
- p.38. Figure 11. Here, the captions also should be more detailed. I am not sure if the y-axis is really consistent between the two plots because the values are very different.  
  
- p.39. "Additionally, an interaction between speaker and verb type was found (β = -0.136, SE = 0.053, z = -2.554, p = 0.011), indicating less looks when listening to the accented speaker" This interpretation seems to suggest a main effect of speaker type rather than an interaction of speaker type and verb type.  
  
- Figures 12-13. The model outputs are better presented as tables.  
  
- The section 4.2. could be better organised by categorising the recommendations by different stages (e.g., set-up stage, filtering stage, data processing stage).  
  
- p.42. Regarding the OS and the browser, adding recommended OS and/or browsers would be helpful.  
  
- p.43. "The crucial decision comes down to understanding the area of interest." I believe the phrase "area of interest" should be rephrased to "time window of interest" or "period of interest", assuming that the authors are not talking about the critical object locations.   
  
Reviewer #3:  
Thank you for the opportunity to review this article, which presents a tutorial for wrangling webcam eye-tracking data produced by Gorilla. The tutorial uses a web-based replication of a previous in-house eye tracking experiment to guide the reader mainly through the pre-processing steps required to clean and analyse the data. The paper makes a huge contribution to the field by providing code and detailed explanations of preprocessing choices in a relatively new methodological area. The Shiny apps were extremely useful and impressive. My comments are mainly about framing and streamlining the article. One question I had was about the suitability of a primarily tutorial paper for a bilingualism journal; I make some suggestions below. Otherwise from my perspective, with some revision (probably classified as major), the paper would be suitable for publication and would make a high-quality contribution to online eye tracking research.  
  
Main comments  
  
1.      As mentioned above, the main focus of the paper is methodological, with some very loose connections to the topic of the journal. I'm not sure how crucial it is that the paper shows a strong bilingualism theme (that would be up to the editor I suppose), but the connection could be strengthened by highlighting why webcam research is particularly relevant for bilingualism research (as opposed to other kinds of linguistic research) and/or giving more weight to the experimental data presented in the paper. For example, the relevance of the findings of the replication are discussed solely in relation to the webcam methodology, but do they also have theoretical relevance?  
2.      Another main thing I stumbled on was that there seems to be a disconnect between how the paper starts and ends: The title and initial paragraphs suggest that it will be about best practices for reporting of web-based eye-tracking preprocessing and analysis choices, whereas the paper itself is mainly about how to do data wrangling. Then, the Discussion is about preprocessing choices and the Conclusion finishes with how the tutorial simplifies the data wrangling process. So is it about reporting, preprocessing choices, or simplifying data wrangling? Even the title of the paper is a bit contradictory, i.e. best reporting practices doesn't directly follow from the art of data wrangling (IMO). For me, the main contribution of the paper is the preprocessing tutorial for \*web-based eye-tracking data\*. For this reason, I would suggest the following (1) changing the title, e.g. to something like "The Art of Wrangling: Preprocessing web-based eye-tracking data…." (or even better,  
make The Art of Wrangling part specific to webcam data, e.g. The Art of Webcam Data Wrangling - data wrangling alone is not specific enough, and becomes a bit confusing when the paper is referred to later using only its short title because it sounds like the authors are referring to a separate textbook for learning R or something). (2) Then, swapping some information around in the Introduction, which starts out talking about data wrangling, which is too generic given the paper is specifically about webcam data. Instead, why not start with the text beginning "Web-based eye-tracking has become more accessible…." And then go on with "Data analysis is not only statistical analysis…". (3) The best practices for reporting are a (good) side effect of the tutorial that could be mentioned but not made part of the title.  
3.      P3, line 22 "This is especially true for methods that capture real-time language processing": Why would open reporting of pre-processing be especially relevant to timecourse data? Isn't it relevant for all data types? Needs clarification.  
4.      Top of page 5: The colour palette should be made colour-blind-friendly.  
5.      Section 2, building the experiment in Gorilla: This was the weakest section for me. Without a visual referent, the description provided in the text is hard to follow. Some suggestions: Begin with step-by-step instructions on how to reach the point where you would see e.g. the web audio zone so that the reader can follow along in Gorilla. And provide a screenshot of the Gorilla dashboard. The linked experiments are helpful to see the end result, but don't match the instructions in the text. Alternatively---and probably my preference---would be to leave this section out entirely. Since it doesn't really teach the reader how to build the experiment, and the focus of the paper is the data wrangling, a short description of Gorilla, an example experiment link, and a description of the output file would suffice. Another alternative would be provide a full tutorial on how to build the Gorilla experiment, but this might be too much for one paper.  
6.      Section 2.3: Some (most?) of the line numbers don't match the .Rmd script on OSF.  
7.      P39: Both statistical models are described as having "found null results" - this is not technically accurate I think (i.e. it implies there's evidence for the null hypothesis, which is not possible in a frequentist analysis). More precise would be to say that the model failed to reject the null hypothesis.  
8.      P41, Discussion: If I read these bolded text subheadings alone, I would not know what some of them meant. For example, "Identify a quadrant classification method" - could be made more self-contained by adding "… for the boundaries of the visual stimulus". Likewise, "Report all time adjustments" needs to be more specific as there are several time-based variables. Same with "Use a meaningful bin size", i.e. bin size of what? Also, if the authors want to stick with the "best reporting practices" framing, some of the subheadings could be rephrased to be more about reporting, e.g. "Ensure participant background info is accurate" is a preprocessing step; a reporting practice would be more like "report accuracy cut-offs for participant background info".  
9.      P42, line 2: "below 5 Hz seems to be unusable": Was it ever shown/discussed what was unusable about it? I thought it was just mentioned that 5 hZ was the cutoff used in previous studies. If not discussed, this should be added.  
10.     P42, line 56 "If eye-fixations are random": I think the authors are referring to eye-fixations \*not\* related to the experimental manipulation, right? This should be stated explicitly as currently it sounds like all eye fixations are random! I'm pretty sure there's a technical term for the non-related fixations which currently escapes me.  
11.     P43, line 22 "scenarios where the bin size is approaching the size of the area of interest": I think it's an issue of terminology but I couldn't understand this - what would the length of the time bin have to do with the area of interest (which usually refers to the square around the images on the presentation screen)? Do the authors maybe mean the time window of the target word in the auditory instruction? This should be made more precise.  
  
Minor comments  
  
1.      P4, line 15 "…given the lack of subscriber-based pre-processing software": The implication here (linguistically) seems to be that the lack of such software is the reason why preprocessing of webcam data needs open science practices. Instead I think there are two issues: 1. The lack of such software makes data wrangling difficult; 2. The complexity of webcam data (between-participant differences in camera/browser specifications) is what makes open science practices crucial.  
2.      Section 1.3 title: "core four" sounds grammatically incorrect to me, although I get the intention. Maybe "four cores" or "four cornerstones/building blocks"?  
3.      Section1.3 first sentence: Please add "insight into" between "provides" and "the time-course of processing" (the current formulation implies the data are cognition whereas we infer that they reflect something about cognition). Also add "cognitive" (or similar) before "processing" so that it's clear what kind of processing we're talking about here (i.e. not data preprocessing or computer processing time).  
4.      P13, line 49, "Prior to reaching 60 participants": this makes it sound like you stopped before 60 participants. Suggest replacing with "We rejected an additional 37 participants…".  
5.      P15, line 5, "appears to be lower than that of Porretta et al": it would be helpful to report Porretta et al's value for comparison; in fact a table comparing all the demographic variables vs. Porretta et al's would be helpful.  
6.      Section 2.3 Data Analysis: I went through the tutorial on OSF to see how it was to follow along, although note that I didn't check any code thoroughly. Everything worked great. One suggestion would be that it would be helpful to have the script and data in one folder that could be downloaded together - at first I didn't realise I had to download both the Github and the OSF folders. I can see why they are separate but from a user point of view, having one neat data/script package is nice.  
7.      P15, line 15: maybe give the OSF link again here, I had to hunt around for it.  
8.      P23, line 47, "removal results would be different": state briefly how they would be different.  
9.      P29, first paragraph: Clarification needed about the minimum frame rate - in the first sentence, a "mean" of 5Hz is mentioned. In the next sentence, a "median" of 5Hz is given. A couple of lines down, a "predetermined median" is mentioned - is this the 5Hz median? Is it the median or the mean? In what way is it predetermined (as in, just from the previous research mentioned? Or was it pre-registered? Or do you mean hypothetically it would predetermined?).  
10.     Section 3: Perhaps in the section title and throughout make explicit that you mean "statistical" models/modelling.  
11.     Section 4, Discussion: After the results section, I expected the Discussion to begin with a summary and interpretation of the results. Instead it begins by saying web-based eye-tracking is an excellent method, but then goes on to say there were differences between the current and previous results, which I found a bit confusing.  I think this is just a matter of moving the information around: I would start with the sentence "Our main models show that…", discuss the differences with Prystauka etc., say why these differences seem to be more to do with the sample than the method, then finish with "Therefore our replication results indicate that…".  
12.     Section 4.2: The first four sentences of this section are a bit clunky. Everyone has to write code to manage their data, so I think this could be left out to simply begin with "In what follows…".  
13.     Conclusions: Some of the information here is a bit vague, e.g. "researchers anywhere can design an experiment" - a web based eye tracking experiment? "For the first time" - because of web cams? Or because of the paper? Suggestion: The most relevant text is the part starting with "Web-based eye tracking is a powerful…". I would keep this, remove the text before it, and add something about how you were able to replicate an in-house eye tracking experiment - aside from the tutorial, this is the most crucial information!  
  
Typos, etc.  
  
-       P23, line 47 "behaviourial" (extra i)  
-       P40, line 38: Sentence beginning with "13" - thirteen should be spelled out. Likewise, single digits like in "0 Chinese experience" should be spelled out.  
-       P41, first line: "Exacting" - I think this should be "extracting"?  
-       P42, line 2: replaced "needed" with "required"

#review 2

The paper focuses on the web-based online eye-tracking method, introducing how to use task builder 1 (on Gorilla) and conduct data wrangling with the data output from Gorilla. To do so, the authors replicated an in-lab eye-tracking study. Overall, it fits the scope of the special issues. Additionally, the authors provided very detailed descriptions of the involved processes along with materials on OSF, making it a good guide to novice users of Gorilla for online eye-tracking. However, the applicability of the paper’s recommended steps/methods is restricted to only the specific method of building an eye-tracking task which is only one of the many ways. There are also some aspects and steps that can be important left undiscussed, making it less of the “best practices”. If it is indeed intended as a guide to novice specific to the methods/platform they used, I recommend a minor revision to address Points related to the replication itself and Other pointsbelow.

Otherwise, I recommend the authors to keep to “Best Practices for Reporting Web-based Eye-tracking Data in Language Research”—the key word here, for me, is reporting, instead of how to build an eye-tracking experiment on Gorilla (using a specific way) and how to specifically wrangle the data output from that particular platform. However, for me personally, it would be more interesting to have a guide on issues that are applicable to all ways of data collection using web-based eye-tracking and more focused on the importance of reporting decisions made during data wrangling (and points on where such reports are recommended). This would require a major revision where all below comments are suggested to be considered.

**Points related to the general content:**

The experiment building section focuses on how to use the interface functions provided by Gorilla. Our lab has been building eye-tracking experiments via Gorilla using code editor and sometimes directly host JavaScript-based eye-tracking experiments online. Not only would this make this specific section less useful for researchers that use other ways of building the experiment but also make the following sections for data wrangling less transferable because the data outputs would be completely different. This should be, at least, acknowledged as this might give rise to some differences and help reproducibility, something stressed by the authors.

Some of the recommended specific ways of dealing with data can be very idiosyncratic, e.g., dealing with time directly vs. changing the baseline. I recommend the authors to stress more on what and how to report—the specific ways data wrangling was done—or even delve deeper to show how and why it is important to report these—e.g., differential results as a function of different decisions on the data wrangling processes. In this way, the discussion/recommendation could also apply to other data outputs.

“The second Shiny app allows the exploration of outputs from statistical models as a function of decisions on bin sizes.” I think this can be dangerous as one could choose a bin size that gives a significant result otherwise unobtainable. Of course, for the present replication, any arbitrary sized bin captures the effect of verb\_type. But it might not be the case for other studies. I think deciding bin size based on sampling rates is more objective. But again, this highlights the importance of reporting the decisions made during data wrangling, something the authors should stress on.

There are also some points I am expecting but not included in the paper. For example, the authors mentioned about the variations of sampling rate across participants and even within participants as a function of experiment time. How should one deal with this, what to report, etc. are very welcomed to be included. This is especially true when different conditions have different starting point of the critical time window and when the same time bin for the same participant for the same trail has more than one looks that are distributed to different AOIs.

**Points related to the replication itself:**

Although it is not the central point of the paper and the paper is replicating a published work, I recommend the authors to include minimally sufficient information on the task/materials/design, such that readers do not need to refer back to the original study for more information. For example, what a visual scene would look like (maybe a figure). Similarly, to better understand the replicability of online Eyetracking of in-lab eye-tracking, a line or two of what the analyses and results were from the original study should be provided.

The differential results found between the current study and the original one was largely attributed to sampling differences. However, statistical power from both the current and the original studies were not provided such that a lack of an effect in one might be due to power differences.

**Other points:**

Page 4, line 39 onwards (para 3). The start of an example seems to come out of the blue, making it difficult to understand. I am wondering if making the example more specific can help readership, for example, specifying this is an example for word recognition experiment? Additionally, I think it would also make it better by starting this paragraph with “VWP experiments vary widely xxx”, then the example. But of course, this is my idiosyncratic preference.

Page 4, line 46-51. I would interpret some of the examples given for different linguistic processes under investigation as the same process, e.g., referent prediction and phonetic cue integration can be part of sentence processing. Maybe make them more specific or just say linguistic processing at different linguistic levels?

Page 6, line 17. In Figure 3 => in Figure 2?

Page 7, line 5. (’beaker’) => (‘beaker’).

Page 7, line 29. I am not sure why each trial “should” start with a fixation cross. Do the authors have a reason for this?

Page 14, line 7. Could you please specify what you mean by “centre images”?

Add figure for poretta et al. and show center image

Page 29, line 33-36. Could you please provide the reason and benefit for this method of binning?

Page 29, line 41-43. I am not sure why this should apply to all studies. For example, in one of our study the audio onset is always constant, as long as the critical time window is ensured to be the same across trials, this step seems unnecessary, unless the authors mean something else. Or to say, one could also do not deal with the timing of each data point directly but rather consider all relevant factors when selecting the time window for plotting and analyses.

Added: If time adjustments are made

Page 30, line 2. L379, is there a reason time\_binning is set to 50?

Page 30, line 49 onwards, maybe it is intended to mean what I understood it to be—the second app is depicted in the figure. But the figure is not about the app?

Page 31, Figure 10. The caption or description for this figure could benefit from more details. How much is high, medium, low (also please spell out this means the bin sizes in the legend)?

Page 36, line 7. GAMMS do better with coded variables. I struggle to understand “coded” here without looking at the code below. For me, dummy/treatment coded are also coded. Maybe specify here—sum coded.

Page 36, line 55-60. Looks to the target (L:603) as the independent variable with dependent variables including xxx. Independent and dependent are reversed—looks to the target should be the dependent variable.

Page 39, line 7-13. Won’t this interaction term mean less looks when listening to the accented speaker as a function of verb type—less looks is not obtained between accented and non-accented speech to the same degree for the two verb types?

Page 42, line 2. ’unusable’ => ‘unusable’.