**Peer Review and Communication History**

**MS Title**: Time stand still: Effects of temporal window selection on eye tracking analysis

**Author Names**: Jonathan E Peelle and Kristin J Van Engen

**Submitted**: Oct 21, 2020

**Editor First Decision—Revise & Resubmit**

Dec 21, 2020

Dear Jonathan E Peelle,

I have now received all reviews of your manuscript, “Time stand still: Effects of temporal window selection on eye tracking analysis” from qualified researchers. I also independently read the manuscript before consulting these reviews. I agree that your manuscript has important strengths and also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology. Papers like this are a huge service to the field.

We’ve received two reviews regarding your submission and they are both highly positive although the second has gone into a deeper discussion on the specific methods and highlighted a potential issue in the way that you’re interpreting poly fits in R. I will highlight issues I think are particularly salient here. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

I’d like you to consider the concerns brought up by R2 and either address them or explain why they are not a particular concern. In some cases you may decide to just simplify the analysis, for example by spending less space discussing the indivdiual components of the polynomial.

One suggestion from me on presentation is that the color mapping on the p value scale makes it somewhat difficult to discern the fluctuations that occur near the 0 end, which is probably the most relevant. I wondered if either a different color map, which provides the eye with chromatic variation, or a log transform of the p values before color mapping would help to better visualize the dynamics at the low end of the scale. This is just a suggestion.

Also, another approach that may be useful for dealing with window selection is blind analysis (also referred to as “orthogonal contrast”) which you can read about here. You may find this useful as an addition, but it’s just a suggestion.

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. Annu. Rev. Nucl. Part. Sci., 55, 141-163.

Brooks, J. L., Zoumpoulaki, A., & Bowman, H. (2017). Data‐driven region‐of‐interest selection without inflating Type I error rate. Psychophysiology, 54(1), 100-113.

It would also be worth mentioning that this general class of problem is the crucial difficulty in FMRI analysis as well, where the space of analysis options is even larger. This has led to the development of techniques such as Statistical Parametric Mapping (i.e. SPM software). This approach has been repackaged in a format more specifically suited for time series data (Maris & Oostenveld, 2007) and then adapted to eye tracking:
<http://www.eyetracking-r.com/docs/make_time_cluster_data>

It might be worth tying the reader into thinking about this larger class of analysis problems that exist across methods.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Brad Wyble

# Reviewer 1

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

This manuscript reports a sensitivity/multiverse analysis examining the effect of time window on the statistical results from a standard “visual world paradigm” (VWP) experiment. The key finding is that the time window matters, even the “reasonable” range of time windows is quite broad and can produce different results. This result is perhaps not surprising, but it is important – VWP is a very widely used method (and many time course methods would have the same problem) and it is standard practice to report a single time window with minimal justification. The manuscript is very well-written and I completely agree with the authors’ recommendations: (1) pre-registration, with pilot data to define the full analysis pipeline (including time window); (2) sensitivity or multiverse analysis. Their Fig 5 is a very nice example of something that other VWP studies could do. Note that, for example, the linear:age interaction is not just arbitrarily significant in some time windows and not significant in others; it shows a consistent pattern (stronger effect in longer time windows) that provides additional insight into the time course of that effect. I think this further demonstrates the value of systematic sensitivity/multiverse analyses – in addition to testing the robustness of the effect, they provide additional insight into the data that can inform the inferences researchers draw from those data.

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

# Reviewer 2

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The authors conduct and report a sensitivity analysis for an eyetracking-based word recognition experiment. In these studies, an array of images are displayed onscreen followed by a prompt to find or look at or click on the name image (“find the dog”). This paradigm is widely used, usually as the “Visual World” paradigm when four images are displayed or as the “Looking While Listening” or “Intermodal Preferential Looking” paradigms when two images are displayed. Given its widespread use, methodological investigations like the current article are valuable in teaching us about the method and its limitations.

The authors focus on a specific type of analytic decision for eyetracking-based word recognition experiments: the analysis window. In these studies, we are interested in how word recognition, as measured by the probability of looking to a named image, changes over time and under different experiment conditions. In order to model this change, we need to identify a window of time to use for the analysis. The authors’ key result and message for practitioners is that small differences in this window selection can change the magnitude, direction and significance of statistical effects.

Using data from an earlier experiment, the authors consider 8000+ analysis windows by systematically varying the window start time and window duration. They then use R and fit the same model to the dataset made for each analysis window. (This part was too computationally intensive for me to reproduce.) The authors use a particular analysis approach: Growth curve analysis with orthogonal polynomials. This technique has been a conventional approach going back to at least 2010, and it accessible description in Mirman (2014) has led to its wide adoption. This is not an uncommon or bespoke analysis technique, so the findings here apply to many studies. Looking at the analysis code, the authors carry out the analysis steps correctly.

There are some subtleties to the orthogonal polynomials used in the manuscript that interfere with some of the results. In R, we use the poly() function to create a basis matrix so that poly(x, 3) would decompose x into orthogonal linear, quadratic and cubic features. These features are orthogonal in that they are uncorrelated with each other: Normally, x and x ^ 2 and x ^ 3 should highly correlated which makes it difficult to compare models with different sets of polynomials. Orthogonal polynomials allow one to add additional polynomials without affecting lower order polynomials because, by definition, the new predictors are uncorrelated with the lower order ones. This statistical convenience comes with two drawbacks: Drawback 1: the values of the polynomials cannot be directly compared with each other without some extra care. Drawback 2: polynomials above the linear term tend to be uninterpretable or only interpretable with extensive exploration and comparison of curves with and without the polynomial. (The linear time term tends to allows be interpretable because it estimates the overall slope of the growth curve.) I should also mention that the all of the parameter estimates, time or otherwise, tend to be difficult to interpret in logistic regression models because they work on the nonlinear log-odds scale.

Drawback 1 affects the comparison between different lengths for the time windows. The windows of time range from 17 to 108 frames. For shorter windows, poly() uses more extreme values. For example, the following code produces a cubic polynomial basis for 17 frames and then takes the range of each column. We do the same for 108 frames.

library(magrittr)

poly(1:17, 3) %>% apply(2, range)

#> 1 2 3

#> [1,] -0.396059 -0.2725865 -0.4497448

#> [2,] 0.396059 0.4543109 0.4497448

poly(1:108, 3) %>% apply(2, range)

#> 1 2 3

#> [1,] -0.1651305 -0.1075690 -0.2408272

#> [2,] 0.1651305 0.2092707 0.2408272

Neither of these two bases are comparable in terms of the linear time. The shorter window ranges from -.396 to .396, and the latter ranges from -.165 to .165. The linear effect of the model will estimate how the looking average looking probability (in logits) changes for a unit change in x, but these units are drastically different. A unit change in the first case describes a change of around 378 ms. A unit change in the latter case is a change from of around 5450 ms.

The point here is that the magnitude differences in Figure 2 and Figure 4 are amplified by artefactual differences in the basis function. The magnitudes of coefficients are not comparable unless the they are all normalized to represent the same unit change in time. Thus, I would ask the authors to reconsider the importance of magnitude differences because the magnitude reflect differences in window duration and not necessarily differences in looking behavior within a window.

Drawback 2 (uninterpretability of high order polynomials) affects the broader interpretation of much of the sensitivity analysis. Let’s contrast two different ways to interpret the polynomial features. In Effect Mindset, time works through different trends: there is a linear effect of time, quadratic effect of time and cubic effect of time, and these trends are important and interpretable on their own. In Function Mindset, there is some effect of time where the average probability of looking at the target changes as a function of time. We approximate this function by decomposing time into different wiggly lines, weighting those lines with regression coefficients, and summing them together. This is the kind of approach underlying generalized additive models and smoothing splines. In this approach, I don’t necessarily care about sign, direction, or significance of any particular polynomial because their job is sum together to approximate the effect of time. It’s the batch of them working together in tandem that’s important, not any particular polynomial.

The authors generally work under the Regression Mindset, which is fine for the age, noise, and frequency effects, and for the linear x age, linear x noise, and linear x frequency (if the linear terms are normalized) effects. But the authors wave at the Spline Mindset in the following note:

For GCA, a change in the sign or magnitude of a parameter estimate is not necessarily surprising, nor indicative of a problem. Rather, polynomial terms are just adjusting to fit the overall shape, and inference is frequently done by comparing models (deemphasizing the importance of a particular component of the model fit). However, the change in sign illustrates the important principle that models will necessarily have to change to accommodate different time windows of data. If our models are important, these changes in parameter estimates are also important.

This is an important point and it bears emphasis in the discussion. We should not worry about the individual contributions of higher order polynomials. Also, I don’t think that the final point about changes in parameter estimates is correct (about time parameters). The changes in time parameter estimates do not matter at all unless we are narrowly comparing datasets on the same window of time.

My broader point here is that Figures 3, 4, and especially 5, are a lot more useful than Figure 2 because the individual time parameters are individually meaningless (at least for the quadratic and cubic). Put differently, there is nothing to be scared of in Figure 2, but Figure 5 should be what keeps researchers up at night. This figure shows how just a few changes in window duration for a given starting point will push around parameter estimates in magnitude and significance. In my experience, it’s not usually difficult to a priori determine a general region for the start time of the analysis window. If we want to know how looks change in response to information, we start the window around when that information comes online: Somewhere around the start of the target word. Determining the endpoint is much more difficult because you need some idea of when listeners are done processing the speech and we are performing the experiment precisely to study when processing occurs. Reasonable people will come up with slightly different start times and different window durations, but I think the range of variability will be more like the variability in Figure 5 (instead of 8000+ options), and those people will come to different conclusions about statistical significance. Scary stuff. 👻 I applaud the authors for this example.

My final comments concern practical recommendations. The authors frequently mention a multiverse analysis. How hard would it be for them to do such an analysis as an illustrative example? They walk down every analysis path but stop short of integrating them together to describe the data. In order to do this multiverse analysis, I think they would need to weight the analysis windows because some of them seem worse or less plausible than others.

A final way to deal with analysis windows is hinted at in the discussion as BDOTS and GAMs are mentioned but I think it bears noting. Both of these approaches allow one to estimate when curves are significantly different from each other. In the GAM approach, we would choose the widest window and fit a smoothing spline over it and then we would make interferences about when the curves differ with a difference smooth. In the BDOTS approach, we still have to choose some analysis window, but eventually we use bootstrapping to determine when curves differ from each other. This approach is available in the polynomial growth curve models too. If we can estimate the average curve for the high frequency condition and the low frequency condition, we can estimate the difference between the two conditions over time. Whenever the 95% interval for this interval excludes 0 (however we choose to get that interval), we might say that the means are different. Then we can find windows of time where the conditions or participant groups are different from each other. With this approach, we would choose a wide window and eliminate many research degrees of freedom from having to choose the right analysis window. We would also gain some degrees of freedom in how to estimate these windows or how to interpret them.

One final note: There seems to be a small problem in how the subsets of time filter rows of data, as the nominal start time and nominal duration files might be off by a frame. For instance, I get the following the start and end times for the following from a sample of files:

#> # A tibble: 249 x 7

#> file nom\_window\_start nom\_window\_leng~ min\_time max\_time duration n\_frames

#> <chr> <chr> <chr> <dbl> <dbl> <dbl> <int>

#> 1 1000\_1~ 1000 1000 1017. 1983. 967. 59

#> 2 1000\_1~ 1000 1016.67 1017. 2000 983. 60

#> 3 1000\_1~ 1000 1033.33 1017. 2033. 1017. 62

#> 4 1016.6~ 1016.67 1150 1017. 2150 1133. 69

#> 5 1016.6~ 1016.67 1166.67 1017. 2183. 1167. 71

#> 6 1016.6~ 1016.67 1183.33 1017. 2183. 1167. 71

#> 7 1033.3~ 1033.33 1300 1050 2333. 1283. 78

#> 8 1033.3~ 1033.33 1316.67 1050 2333. 1283. 78

#> 9 1033.3~ 1033.33 1333.33 1050 2350 1300 79

#> 10 1050\_1~ 1050 1450 1067. 2483. 1417. 86

#> # ... with 239 more rows

Tristan Mahr, Ph.D.
Waisman Center
University of Wisconsin-Madison

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

**Author Response**

May 21, 2021

Thank you for the constructive feedback on our manuscript. We have now responded to the comments. We have pasted in the comments verbatim below in bold, followed by our response to each.

**Editor comments**

**One suggestion from me on presentation is that the color mapping on the p value scale makes it somewhat difficult to discern the fluctuations that occur near the 0 end, which is probably the most relevant. I wondered if either a different color map, which provides the eye with chromatic variation, or a log transform of the p values before color mapping would help to better visualize the dynamics at the low end of the scale. This is just a suggestion.**

Thanks for this suggestion. We’ve had that suggestion from several readers and have now included a different color scale for the p values. We hope this makes it easier to visualize the variability.

**Also, another approach that may be useful for dealing with window selection is blind analysis (also referred to as "orthogonal contrast") which you can read about here. You may find this useful as an addition, but it's just a suggestion.**

**Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. Annu. Rev. Nucl. Part. Sci., 55, 141-163.**

**Brooks, J. L., Zoumpoulaki, A., & Bowman, H. (2017). Data‐driven region‐of‐interest selection without inflating Type I error rate. Psychophysiology, 54(1), 100-113.**

Thank you for pointing us towards this literature, which we have now referenced. One of the fun (and sometimes challenging) areas of this manuscript has been the wide range of fields with which it makes contact!

**It would also be worth mentioning that this general class of problem is the crucial difficulty in FMRI analysis as well, where the space of analysis options is even larger. This has led to the development of techniques such as Statistical Parametric Mapping (i.e. SPM software). This approach has been repackaged in a format more specifically suited for time series data (Maris & Oostenveld, 2007) and then adapted to eye tracking:**<http://www.eyetracking-r.com/docs/make>***time*cluster\_data**

**It might be worth tying the reader into thinking about this larger class of analysis problems that exist across methods.**

Thank you for this suggestion. We have restructured our cursory mention of these issues and added a paragraph to the introduction to more clearly make this point, including that the specific dimensions causing the most “problems” will vary over method (p. 3):

There are, of course, many types of data involving large parameter spaces over which to search. In psychology and cognitive neuroscience, examples of common time series signals include EEG, single unit recordings from neurons, and data from eye tracking and pupillometry. Functional MRI has faced this challenge both in the context of extended detection of three-dimensional signals (Worsley et al., 1992), and in the large number of potential analysis pipelines (Carp, 2012). Permutation testing has proven useful in detecting a variety of signals extended in both space (Nichols and Holmes, 2001; Smith and Nichols, 2009) and time (Maris and Oostenveld, 2007), but the challenges of large data sets with many possible analyses remain a contemporary issue. Thus, although in the current paper we focus on eye tracking data as an example, the issues relating to data selection and model testing are broadly reflected in many areas of science.

**Reviewer 1**

**This manuscript reports a sensitivity/multiverse analysis examining the effect of time window on the statistical results from a standard “visual world paradigm” (VWP) experiment. The key finding is that the time window matters, even the “reasonable” range of time windows is quite broad and can produce different results. This result is perhaps not surprising, but it is important – VWP is a very widely used method (and many time course methods would have the same problem) and it is standard practice to report a single time window with minimal justification. The manuscript is very well-written and I completely agree with the authors’ recommendations: (1) pre-registration, with pilot data to define the full analysis pipeline (including time window); (2) sensitivity or multiverse analysis. Their Fig 5 is a very nice example of something that other VWP studies could do. Note that, for example, the linear:age interaction is not just arbitrarily significant in some time windows and not significant in others; it shows a consistent pattern (stronger effect in longer time windows) that provides additional insight into the time course of that effect. I think this further demonstrates the value of systematic sensitivity/multiverse analyses – in addition to testing the robustness of the effect, they provide additional insight into the data that can inform the inferences researchers draw from those data.**

Thank you for the careful reading of our paper.

**Reviewer 2**

**There are some subtleties to the orthogonal polynomials used in the manuscript that interfere with some of the results. In R, we use the poly() function to create a basis matrix so that poly(x, 3) would decompose x into orthogonal linear, quadratic and cubic features. These features are orthogonal in that they are uncorrelated with each other: Normally, x and x ^ 2 and x ^ 3 should highly correlated which makes it difficult to compare models with different sets of polynomials. Orthogonal polynomials allow one to add additional polynomials without affecting lower order polynomials because, by definition, the new predictors are uncorrelated with the lower order ones. This statistical convenience comes with two drawbacks: Drawback 1: the values of the polynomials cannot be directly compared with each other without some extra care. Drawback 2: polynomials above the linear term tend to be uninterpretable or only interpretable with extensive exploration and comparison of curves with and without the polynomial. (The linear time term tends to allows be interpretable because it estimates the overall slope of the growth curve.) I should also mention that the all of the parameter estimates, time or otherwise, tend to be difficult to interpret in logistic regression models because they work on the nonlinear log-odds scale.**

**Drawback 1 affects the comparison between different lengths for the time windows. The windows of time range from 17 to 108 frames. For shorter windows, poly() uses more extreme values. For example, the following code produces a cubic polynomial basis for 17 frames and then takes the range of each column. We do the same for 108 frames.**

**library(magrittr)**

**poly(1:17, 3) %>% apply(2, range)**

**#> 1 2 3**

**#> [1,] -0.396059 -0.2725865 -0.4497448**

**#> [2,] 0.396059 0.4543109 0.4497448**

**poly(1:108, 3) %>% apply(2, range)**

**#> 1 2 3**

**#> [1,] -0.1651305 -0.1075690 -0.2408272**

**#> [2,] 0.1651305 0.2092707 0.2408272**

**Neither of these two bases are comparable in terms of the linear time. The shorter window ranges from -.396 to .396, and the latter ranges from -.165 to .165. The linear effect of the model will estimate how the looking average looking probability (in logits) changes for a unit change in x, but these units are drastically different. A unit change in the first case describes a change of around 378 ms. A unit change in the latter case is a change from of around 5450 ms.**

**The point here is that the magnitude differences in Figure 2 and Figure 4 are amplified by artefactual differences in the basis function. The magnitudes of coefficients are not comparable unless the they are all normalized to represent the same unit change in time. Thus, I would ask the authors to reconsider the importance of magnitude differences because the magnitude reflect differences in window duration and not necessarily differences in looking behavior within a window.**

Thank you for highlighting the important details in interpreting model fits, here and in your other comments. In response to this, and other, comments, we have gone through and reduced the weight we place on interpreting specific parameter estimates. Really, our point is not to have a quantitative comparison across models, but to highlight the differences time window selection will have on any single model selected by researchers, which changing model values help to illustrate. We hope the revised manuscript is more accurate on these points. As one example, we have added the following explanation to the manuscript near the discussion of the relevant figures (p. 7):

At first glance, it may also seem troubling that the magnitudes and even signs of parameter estimates are also changing as a function of time window. For example, the quadratic and cubic terms (**Figure 2**) both transition from positive to negative and back again (a red sector, a blue sector, another red sector). These changes correspond to inflection points in the curves: as the start time and/or window length are varied, the inflection points move around within the time window of the modeled data, such that over all models (as plotted here) “stripes” occur. Although considering the variability in parameter estimates is useful for illustrative purposes, in the context of GCA, a change in the sign or magnitude of a parameter estimate is not necessarily surprising or indicative of a problem. Rather, polynomial terms are simply adjusting to fit the overall shape, and inference is typically done by comparing models (deemphasizing the importance of a particular component of the model fit). It is also important to consider that polynomial basis functions will differ based on the number of time points of data being modeled (i.e., shorter time windows get assigned more extreme values). Thus, changes in magnitude of parameter estimates are not directly comparable across time windows of different lengths. Nevertheless, from a big-picture perspective, it is important to realize that different time window selections will influence model behavior, and, in some cases, study outcomes. Particularly, if authors have specific hypotheses about how fixed effects of interest (e.g., age) will interact with parameter estimates (e.g., linear), then time window selection will impact the magnitude and direction of these interactions.

**Drawback 2 (uninterpretability of high order polynomials) affects the broader interpretation of much of the sensitivity analysis. Let’s contrast two different ways to interpret the polynomial features. In Effect Mindset, time works through different trends: there is a linear effect of time, quadratic effect of time and cubic effect of time, and these trends are important and interpretable on their own. In Function Mindset, there is some effect of time where the average probability of looking at the target changes as a function of time. We approximate this function by decomposing time into different wiggly lines, weighting those lines with regression coefficients, and summing them together. This is the kind of approach underlying generalized additive models and smoothing splines. In this approach, I don’t necessarily care about sign, direction, or significance of any particular polynomial because their job is sum together to approximate the effect of time. It’s the batch of them working together in tandem that’s important, not any particular polynomial.**

**The authors generally work under the Regression Mindset, which is fine for the age, noise, and frequency effects, and for the linear x age, linear x noise, and linear x frequency (if the linear terms are normalized) effects. But the authors wave at the Spline Mindset in the following note:**

**For GCA, a change in the sign or magnitude of a parameter estimate is not necessarily surprising, nor indicative of a problem. Rather, polynomial terms are just adjusting to fit the overall shape, and inference is frequently done by comparing models (deemphasizing the importance of a particular component of the model fit). However, the change in sign illustrates the important principle that models will necessarily have to change to accommodate different time windows of data. If our models are important, these changes in parameter estimates are also important.**

**This is an important point and it bears emphasis in the discussion. We should not worry about the individual contributions of higher order polynomials. Also, I don’t think that the final point about changes in parameter estimates is correct (about time parameters). The changes in time parameter estimates do not matter at all unless we are narrowly comparing datasets on the same window of time.**

Thank you for these helpful comments and the nice summary of the different mindsets. In the revised manuscript, we have backed off from the wording on specific polynomials (including deleting the problematic sentence) and hopefully now have a more accurate framing of the issues, including a more explicit inclusion of what you call the Function mindset. We also now include an expanded portion of the Discussion (p. 10) where we emphasize these points:

It is worth emphasizing that in the specific context of GCA, we should probably not worry about individual contributions of higher-order polynomials, losing the forest from the trees. That is, it is the overall fit of the function that is of primary interest (the combination of different polynomial basis functions). Statistical evaluations are frequently performed by comparing models (deemphasizing the importance of any particular component of the model fit). We have included plots of individual parameter estimates to help illustrate how the models fit the data, but these should be interpreted in the context of the overall analytic framework being adopted.

**My broader point here is that Figures 3, 4, and especially 5, are a lot more useful than Figure 2 because the individual time parameters are individually meaningless (at least for the quadratic and cubic). Put differently, there is nothing to be scared of in Figure 2, but Figure 5 should be what keeps researchers up at night. This figure shows how just a few changes in window duration for a given starting point will push around parameter estimates in magnitude and significance. In my experience, it’s not usually difficult to a priori determine a general region for the start time of the analysis window. If we want to know how looks change in response to information, we start the window around when that information comes online: Somewhere around the start of the target word. Determining the endpoint is much more difficult because you need some idea of when listeners are done processing the speech and we are performing the experiment precisely to study when processing occurs. Reasonable people will come up with slightly different start times and different window durations, but I think the range of variability will be more like the variability in Figure 5 (instead of 8000+ options), and those people will come to different conclusions about statistical significance. Scary stuff. 👻 I applaud the authors for this example.**

Thank you for these comments. We agree Figure 5 is probably a more plausible set of alternatives than the other figures, and are glad it is still scary.

**My final comments concern practical recommendations. The authors frequently mention a multiverse analysis. How hard would it be for them to do such an analysis as an illustrative example? They walk down every analysis path but stop short of integrating them together to describe the data. In order to do this multiverse analysis, I think they would need to weight the analysis windows because some of them seem worse or less plausible than others.**

Thank you for this suggestion. We seriously considered a more formal multiverse and/or sensitivity analysis. Due to practical constraints we don’t feel able to expand to this at the moment, but hope that in the future it’s something we can explore. More generally, we are optimistic that increased use of sensitivity analyses will be beneficial to these types of data sets more broadly.

**A final way to deal with analysis windows is hinted at in the discussion as BDOTS and GAMs are mentioned but I think it bears noting. Both of these approaches allow one to estimate *when* curves are significantly different from each other. In the GAM approach, we would choose the widest window and fit a smoothing spline over it and then we would make interferences about when the curves differ with a difference smooth. In the BDOTS approach, we still have to choose some analysis window, but eventually we use bootstrapping to determine when curves differ from each other. This approach is available in the polynomial growth curve models too. If we can estimate the average curve for the high frequency condition and the low frequency condition, we can estimate the difference between the two conditions over time. Whenever the 95% interval for this interval excludes 0 (however we choose to get that interval), we might say that the means are different. Then we can find windows of time where the conditions or participant groups are different from each other. With this approach, we would choose a wide window and eliminate many research degrees of freedom from having to choose the right analysis window. We would also gain some degrees of freedom in how to estimate these windows or how to interpret them.**

We agree that both of these approaches have advantages, and have expanded our treatment of them in the Discussion (p. 12):

Finally, it is worth noting there are other ways of approaching time series data that may reduce reliance on a specific analysis window. Seedorff and colleagues (2018), for example, suggest a bootstrapping approach, BDOTS, to estimate differences in time series data in a time-by-time basis, and would thus have different constraints than our function-based modeling approach. General additive models (van Rij et al., 2019) are able to fit a broader set of shapes than the polynomial-based GCA approach we use here, and as a result may be less sensitive to the specific shape of the data within a time window. Both approaches are probably less sensitive to specific time windows chosen, and would likely work better with larger time windows. For example, in the GAM approach, researchers could choose the widest possible window and fit a smoothing spline, making inferences about when curves diverge with a difference smooth. In the BDOTS approach, bootstrapping can be used to determine when curves differ from each other. In fact, it may be possible to implement some aspects of these analyses in GCA using a difference estimate between conditions over time. Such an approach would reduce or remove the influence of time window selection on the process.

**One final note: There seems to be a small problem in how the subsets of time filter rows of data, as the nominal start time and nominal duration files might be off by a frame. For instance, I get the following the start and end times for the following from a sample of files:**

**#> # A tibble: 249 x 7**

**#> file nom\_window\_start nom\_window\_leng~ min\_time max\_time duration n\_frames**

**#> <chr> <chr> <chr> <dbl> <dbl> <dbl> <int>**

**#> 1 1000\_1~ 1000 1000 1017. 1983. 967. 59**

**#> 2 1000\_1~ 1000 1016.67 1017. 2000 983. 60**

**#> 3 1000\_1~ 1000 1033.33 1017. 2033. 1017. 62**

**#> 4 1016.6~ 1016.67 1150 1017. 2150 1133. 69**

**#> 5 1016.6~ 1016.67 1166.67 1017. 2183. 1167. 71**

**#> 6 1016.6~ 1016.67 1183.33 1017. 2183. 1167. 71**

**#> 7 1033.3~ 1033.33 1300 1050 2333. 1283. 78**

**#> 8 1033.3~ 1033.33 1316.67 1050 2333. 1283. 78**

**#> 9 1033.3~ 1033.33 1333.33 1050 2350 1300 79**

**#> 10 1050\_1~ 1050 1450 1067. 2483. 1417. 86**

**#> # ... with 239 more rows**

Thank you for catching this! Now amended (short version: it was an oversight using > rather than >=). (The code has also been updated to run in R 4.0.3.)

**Editor Final Decision—Accept**

July 19, 2021

Dear Jonathan E Peelle,

I have now had a chance to read over your manuscript “Time stand still: Effects of temporal window selection on eye tracking analysis”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point in the paper itself. However, I would point out that there are several R scripts in the OSF repository that seem to extend beyond the paper itself. That is great and it would be helpful to document briefly what those extra R files do in the readme file.

Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,
Brad Wyble

# Reviewer 1

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I thought this was a good manuscript on the original submission and had no substantive critiques at that stage. I think the authors have made improvements to an already-strong piece of work.

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

# Reviewer 2

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I appreciate that the authors have acknowledged the fussiness of coefficient magnitudes with the poly() function. I also appreciate that they have addressed that the individual significance of polynomial terms is less important the summed of the basis functions for fitting the data. This manuscript, especially Figure 5, is an important note/disclaimer for researchers of word recognition time series.

Tristan Mahr

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |