Analyzing the time course of pupillometric data

Journal Title
XX(X):1–18
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Jacolien van Rij[1], Petra Hendriks[1], Hedderik van Rijn[1], R. Harald Baayen[2], Simon N. Wood[3]

Abstract

This paper provides a tutorial for analyzing pupillometric data. Pupil dilation has become increasingly popular in psychological and psycholinguistic research as a measure to trace language processing. However, there is no general consensus about procedures to analyze the data, with most studies analyzing extracted features from the pupil dilation data instead of the analyzing the pupil dilation trajectories directly. Recent studies have started to apply nonlinear regression and other methods to analyze the pupil dilation trajectories directly, utilizing all available information in the continuously measured signal. This paper applies a nonlinear regression analysis, Generalized Additive Mixed Modeling (GAMM; Hastie and Tibshirani 1990; Wood 2017a), and illustrates how to analyze the full time course of the pupil dilation signal. The regression analysis is particularly suited for analyzing pupil dilation in the fields of psychological and psycholinguistic research because GAMMs can include complex nonlinear interactions for investigating the effects of properties of stimuli (e.g., formant frequency) or participants (e.g., working memory score) on the pupil dilation signal. To account for the variation due to participants and items nonlinear random effects can be included. However, one of the challenges for analyzing time series data is dealing with the autocorrelation in the residuals, which is rather extreme for the pupillary signal. On the basis of simulations we explain potential causes of this extreme autocorrelation, and on the basis of the experimental data we show how to reduce their adverse effects, allowing a much more coherent interpretation of pupillary data than possible with feature-based techniques.

Keywords

pupillometry, statistical analysis, autocorrelation, preprocessing, Generalized Additive Mixed Modeling

Introduction

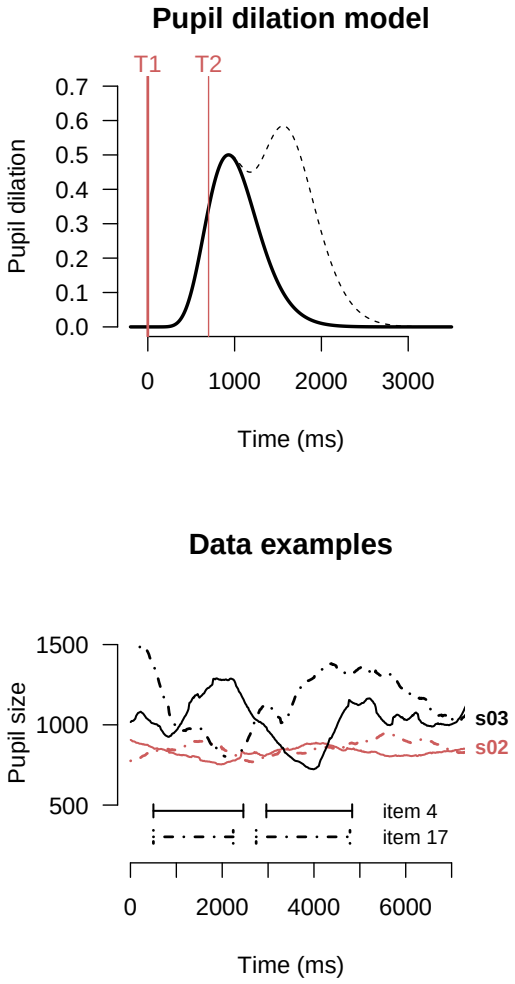
Pupil dilation is a well-established and highly sensitive measure of cognitive processing and resource allocation, which has been used in many research areas, ranging from cognitive psychology and psychophysics to language and speech processing. Under constant luminance, the pupil size is a relatively slow changing signal that is generally assumed to peak around one second after stimulus onset. However, the peak latency may vary considerably between tasks. For example, Hoeks and Levelt (1993) estimated the mean peak latency in a simple reaction task on 930 ms after the stimulus, whereas Just and Carpenter (1993) report peak latencies around 1.3 s for reading sentences (for reviews, see a.o., Beatty 1982; Beatty and Lucero-Wagoner 2000; Janisse 1977; Laeng et al. 2012). The pupillary response is most likely a combination of many different underlying cognitive processes, as illustrated in Figure 1, *Top*, which may cause the larger latencies for more complex tasks (e.g., Hoeks and Levelt 1993; Wierda et al. 2012).

The sensitivity of the pupil dilation signal is also a drawback of the measure: even when keeping the luminance levels constant, in addition to the experimental manipulation the signal is likely to be influenced by potentially confounding factors related to the mental state of the participant, or properties of the stimuli and the task (e.g. Beatty and Lucero-Wagoner 2000; Goldwater 1972). Pupil dilation experiments require specific design considerations to avoid that such factors confound with conditions in the experiment. These considerations are especially important when stimuli are used that can differ in many dimensions such as pictures of complex scenes (Goldwater 1972) or linguistic stimuli, which have many properties that could confound with pupil dilation such as emotional valence, word length, or frequency (e.g., Kuchinke et al. 2007), prosodic information (e.g., Engelhardt et al. 2010), speech intelligibility (e.g., Zekveld et al. 2011), discourse structure (e.g., van Rij 2012; Zellin et al. 2011; Vogelzang et al. 2016), grammatical complexity (e.g., Just and Carpenter 1993; Schluroff 1982; Schluroff et al. 1986), and other linguistic factors (e.g., Hyönä et al. 1995; Scheepers and Crocker 2004). In this manuscript, we will introduce an analysis method that allows for investigating potentially nonlinear effects of properties of stimuli. We will focus on pupillary data collected in a visual world paradigm experiment (i.e., sentences presented auditorily together with a visual context), but the presented techniques apply equally well to or any other experimental psychological paradigm.

[1]University of Groningen, The Netherlands
[2]Eberhard Karls Universität Tübingen, Germany
[3]University of Bristol, United Kingdom

Corresponding author:
Jacolien van Rij, University of Groningen / Institute of Artificial Intelligence & Cognitive Engineering, Nijenborgh 9, 9747 AG Groningen, The Netherlands
Email: j.c.van.rij@rug.nl

Figure 1. Properties of pupil dilation. *Top:* The effect of cognitive processing on pupil dilation, as described by the pupil dilation function from Hoeks and Levelt (1993, p.21). The pupillary response is scaled to 0.5 mm for comparison (cf. Beatty and Lucero-Wagoner 2000). The red vertical line T1 represents an event that triggers dilation (black solid line). The dashed line shows the adjusted dilation when a second event T2 shortly follows the first event. *Bottom:* Example of two actual recorded pupil dilation time series, recorded from two different participants (solid versus dashed lines) in two different trials (red versus black lines). The data is re-aligned on the onset of the pronoun. The horizontal bars indicate the duration of the auditory stimuli (two spoken sentences) of the two trials.



The variability of the pupil dilation signal poses a challenge for the analysis: the signal shows both large within and between subject variation (Winn et al. 1994; see also Figure 1 *Bottom*). When the analyses do not take this variability into account, then the conclusions are likely anti-conservative (as the observed effect might be due to the noise caused by the large variability).

Generally, the pupil dilation signal is not analyzed directly, but different features are extracted for further analysis. The most often-used measures are the *peak dilation*, the maximal dilation within a specified time window (often labeled the analysis window), and the *peak latency*, the time between a critical point in time of the task (typically the onset of a stimulus) and the peak dilation. However, other measures have also been proposed, such as a *mean dilation* measured over an analysis window, and the *dilation slope*, the steepness of the increase in pupil dilation in a particular time window.

Quantifying the pupil dilation data into a set of features reduces the complexity of the analysis, but the large amount of variation in the pupil dilation recordings causes a problem for this type of analysis: Regularly, trials do not show a peak in pupil dilation, and therefore the peak amplitude and peak latency cannot be determined. This may happen more often in tasks where trials do not have a clear start and end, such as when detecting specific words in a continuous stream of speech. Sometimes these trials are excluded from analysis, thereby reducing the size of the data. To avoid the problem of determining the peak in trajectories without a clear peak dilation, Verney et al. (2004) proposed a principle components analysis (PCA) to reduce the time course into three measures. However, a disadvantage of the PCA analysis, which also holds for the feature analysis, is that the interpretation of the results may be more difficult when the various features show different effects. Therefore, we would like to argue that analyzing the pupillary response signal as it develops over time yields a much more coherent interpretation of the data.

Different methods have been used to model the pupil dilation time course, including *Growth Curve Analysis* (GCA; Mirman et al. 2008), a mixed effect regression approach that represents the predictor *Time* with orthogonal polynomials (see Kuchinsky et al. 2013; Winn et al. 2015), *Functional Data Analysis* (FDA; Ramsay and Silverman 2002; Ramsay and Silverman 2005) (see Jackson and Sirois 2009), *Generalized Additive Mixed Modeling* (GAMMs Hastie and Tibshirani 1990; Wood 2017a, 2011), a nonlinear mixed-effects regression method (see Lõo et al. 2016; van Rij 2012; Vogelzang et al. 2016), and a *deconvolution approach* (Hoeks and Levelt 1993; Wierda et al. 2012). These analyses are more powerful than the traditional approaches, which analyze features of the pupil dilation signal separately. For example, they make it possible to investigate differences in pupil size that do not result in differences in peak dilation or peak latency. Another important advantage of time course analyses is that they allow for a systematic description of the data rather than focusing solely on the statistical significance of the differences between the experimental conditions. For investigating cognitive processing, it may be more informative at which moment in time the conditions start to differ, and whether this difference is found in all participants, and what the size is of the effect in comparison with other factors that influence pupil dilation. With the increase in computational speed and memory, this type of analysis has become easier to apply and therefore gains in popularity.

However, what many pupillometric studies fail to mention is that the direct analysis of time series, such as the pupil dilation signal, raises the problem of autocorrelated errors (e.g., Baayen et al. to appear)[1]. Autocorrelated errors increase the probability of Type I errors (i.e., false positives, for example finding a significant difference which does not actually exists), and lead to conclusions that cannot be replicated.

In this paper we use Generalized Additive Modeling (GAMM; Hastie and Tibshirani 1990; Lin and Zhang 1999;

Wood 2017a, 2011) to analyze the pupil dilation time course directly. GAMMs offer various features that are relevant for the analysis of pupillometric data: The method can handle the variability of the pupil dilation signal by means of nonlinear random effects. Moreover, GAMMs provide the option to include nonlinear interactions with two or more numeric predictors. These nonlinear interaction surfaces are particularly useful for studies in the domain of hearing research, because they allow to explore the potential nonlinear relations between the pupil dilation response and continuous properties of the presented stimuli, such as formant frequency (pitch), word length, or signal-to-noise ratio in the case of language and speech experiments. Finally, GAMMs offer the possibility to include an AR(1) error model for Gaussian models to deal with autocorrelational structure in the errors.

Aim of current paper

This paper provides a tutorial for analyzing the time course of pupillometric data, and explains the problems that arise when analyzing the time course directly. We also will discuss various options of how we could reduce the effect of autocorrelation. Although the paper uses GAMMs to illustrate problem of autocorrelational structure in the errors, the problem of autocorrelation is not limited to this method, but applies to other time course analyses as well.

Generalized Additive Mixed Modeling is implemented in the R package `"mgcv"`(version 1.8-23) (Wood 2017b,a). Additionally, we use the R package `"itsadug"`(version 2.3 van Rij et al. 2017) for interpretation and visualization of the statistical analyses. The code and the data are available online as Supplementary Materials[2].

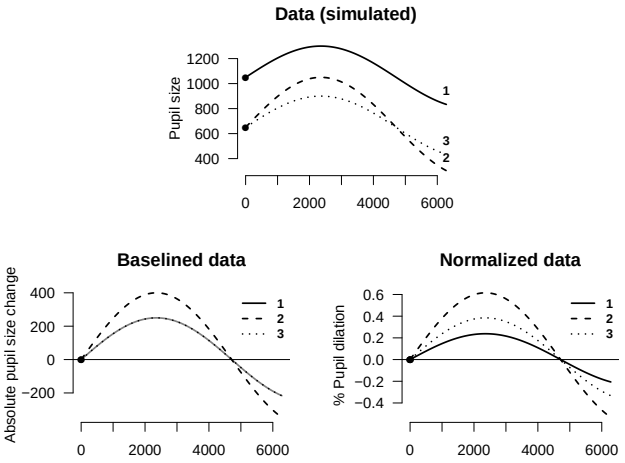
The paper is organized as follows. In the remainder of the introduction we will present considerations for the analysis of pupillometric data. These aspects are often not (explicitly) taken into account in pupil dilation studies, but might affect the studies' validity. Then we summarize an experimental data set that we use as a case study to introduce GAMMs. In the next section, we will introduce GAMMs by presenting an analysis of the experimental data. We assume the reader to be familiar with the basic concepts of regression analysis. On the basis of the presented GAMM analysis, we will explain the issues that arise with a time course analysis of pupillometric data, and provide solutions for these issues. In the discussion we will compare the GAMM-based methods to currently existing analysis techniques.

Considerations

Typical video-based eye trackers report for each sample the measured X and Y positions of the eye, and the measured pupil size. However, there are a number of considerations that have to be taken into account before this measured pupil size can be reliably used.

Normalization. To reduce the influence of the factors unrelated to the experimental design that may influence pupil size, such as the luminance of the room, the measured pupil dilation is usually normalized with respect to a baseline. Typically, the baseline is defined on a trial-by-trial basis as the average pupil dilation during a short time frame just before the presentation of the relevant stimulus. For

Figure 2. Baseline correction and normalization. *Top:* Three pupil dilation trials (simulated data) with Trial 1 and 3 differing in baseline, but showing the exact same pupillary response. Trials 2 and 3 share the same baseline, but Trial 2 shows a higher peak amplitude. *Left:* The same three trials after the baseline is subtracted. The baselined data for Trials 1 and 3 overlap. *Right:* The proportion pupil dilation change with respect to the baseline for same three trials. As the baseline of Trial 1 is much higher than of Trial 3, the pupil dilation change is much lower for Trial 1 than for Trial 3, although the measured pupillary response was exactly the same.



baselined pupil size, this baseline value is simply subtracted from the measured pupil dilation, because the changes in pupil size elicited by cognitive processes have been found to be similar for a wide range of baseline values (e.g., Bradshaw 1970; Hoeks and Levelt 1993). This is illustrated in Figure 2 (Bottom-left panel). Alternatively, many studies report the pupillary response value in units of percent dilation over the baseline pupil size following the early studies of Hess and Polt (1960, 1964): After subtracting the baseline value from the measured pupil dilation, the resulting value is divided by the baseline. This is called normalized pupil size. It is good to realize that with this method the same pupil response will result in higher percentage of change with smaller baseline values in comparison with larger baseline values, as illustrated in Figure 2 (Bottom-right panel).

Mixed-effects modeling techniques such as GAMMs do not necessarily require baseline correction of normalization, but allow for taking into account the effect of the baseline on the evoked response as (nonlinear) covariate (see Supplementary Materials for an example). This is an alternative to baseline corrections, especially useful in situations where a baseline correction is likely to introduce artifacts. For example, because pupil size is sensitive to factors as arousal and vigilance, the participant's pupil size *b*efore trial onset may affect the pupillary response *w*ithin the trial. A potential difficulty with this approach is that for pupillometric data from language and speech experiments, the pupil dilation response is rather small in comparison with the baseline values and the variation in baseline values. The statistical model will reflect this uncertainty, which may mask differences in the pupil dilation response. In this paper however we will follow the conventional approach and use baseline corrected data.

Measurement Scales. Although eye trackers report pupil dilation expressed on an absolute scale, modern eye-tracking systems, and especially the popular table-mounted cameras, often do not report pupil size in mm, but use different units. These values do not easily translate to length units (SR Research Ltd. 2005–2010, p.98), and are dependent on the experimental setup, such as the lens that is being used, or the distance between the recorded eye(s) and the camera (e.g., Hayes and Petrov 2016). [3] Given the different units of measurement, some have argued to express the measured pupil size in percent change from baseline to be able to compare the results of different studies. However, this derived measure has other issues, as described in the previous section.

Effect of Gaze Location. Another aspect that influences the measured pupil diameter is the gaze position. When the eyes look directly at the camera, the pupil will be observed as an (almost) perfect circle. However, if the eyes move to bring more eccentric positions in focus, the camera will not perceive the whole pupil, instead it will perceive a smaller, squashed image. As the camera in table-mounted setups is typically located below the screen, the measured pupil size will increase when looking to locations near the bottom of the screen, and decrease when looking to higher locations. Similar effects will be visible for movements on the horizontal axis. Note that these effects are not limited to table-mounted cameras, it is a physical constraint of recording a partial surface on a rotating sphere with a fixed positon camera. Recent studies have indeed shown that gaze position systematically affects the measured pupil size (Brisson et al. 2013; Gagl et al. 2011; Hayes and Petrov 2016), with similarly sized effect sizes as the typical evoked pupillary response. In addition, gaze position may indirectly influence the measured diameter, as recent findings suggest that saccade preparation may also elicit a pupillary response before the actual change of the gaze position (e.g., Jainta, Stephanie et al. 2011; Mathôt, Sebastiaan et al. 2015).

The preferred way to solve this problem is to only record pupillary data when information is presented at a fixed location on the screen (e.g., van Rijn et al. 2012), but in many paradigms participants needs to be able to shift their gaze during recording, for example, during sentence reading (e.g., Gagl et al. 2011; Just and Carpenter 1993) or in studies in which relevant information is distributed over the screen (e.g., Cooper 1974; Tanenhaus et al. 1995; Scheepers and Crocker 2004; Engelhardt et al. 2010, and the study discussed in this paper). When gaze shifting cannot be prevented, the observed pupillary response should be corrected for gaze-dependent fluctuations. However, it is difficult to correct for pupillary responses elicited by the preparation of eye movements, because they cannot be linked to gaze position directly.

Preprocessing data. Analyzing pupil dilation with a time course analysis also has implications for preprocessing of the pupil dilation data. An important step in the preprocessing of pupillometric data consist of removing artifacts, saccades, and blinks. Generally, blinks are interpolated with a linear or polynomial function to avoid missing data (e.g., Mathôt 2013). In a mixed-effects framework this process becomes optional: mixed-effect approaches can handle missing data. Pupil dilation data is often recorded with a higher sampling rate than necessary for analyzing the relatively slowly changing pupillary response. To reduce the autocorrelation in the residuals downsampling to at most 50Hz is recommended, as downsampling increases the distance between consequent data points and reduces the autocorrelation. An additional advantage of downsampling is that it reduces the size of the data, which speeds up the analyses. Before downsampling the data is normally filtered to avoid aliasing. Although this is a good practice, filtering also reduces the noise. Therefore we recommend to be careful with filtering, and to apply filtering only to avoid that aliasing would interfere with the measured signal.

These different considerations, namely the measurement scale, normalization, the effect of gaze location, and data preprocessing should be taken into account in the analysis.

Table 1. Example sentence materials (in Dutch, with their English translations given below):

Introduction sentence:	
A1: actor 1st	Hier zie je een pinguïn en een schaap. *'Here you see a penguin and a sheep'*
A2: actor 2nd	Hier zie je een schaap en een pinguïn. *'Here you see a sheep and a penguin'*
Test sentence:	
De pinguïn slaat hem met een pan. *'The penguin is hitting him with a pan'*	

Experimental data

The pupillary data analyzed in this paper was collected in an experiment in which native Dutch participants were looking at a picture depicting an action between two agents, see Figure 3. While looking to the picture, two short Dutch sentences were presented auditorily in which the two agents were introduced, see Table 1. The first sentence always introduced both agents (e.g., 'Here you see a penguin and a sheep.'), and the second sentence described an action (e.g., 'The penguin is hitting him with a pan.'). The picture on the screen was either *congruent* with the auditory presented second sentence or *incongruent* (i.e., a picture of a penguin hitting himself, rather than the sheep). Participants had to indicate whether the picture and the last sentence were in agreement. In addition to the picture/sentence congruency, a second manipulated factor was the order of introduction of both agents. The introduction sentence could first introduce the actor (as in the example above), or first introduce the non-acting agent (i.e., 'Here you see a sheep and a penguin'). Thus, the experiment followed a classical 2x2 factorial design, with picture-sentence congruency (henceforth *Congruency*) and introduction order (henceforth *Introduction Order*) as manipulations. The predictor *Condition* describes the four conditions of the 2x2 design.

The experiment was aimed to investigate the effects of linguistic and visual context on the processing of the third-person singular masculine pronoun ('him' in English) in object position. More details of the experimental design are reported in van Rij (2012, Chapter 6).

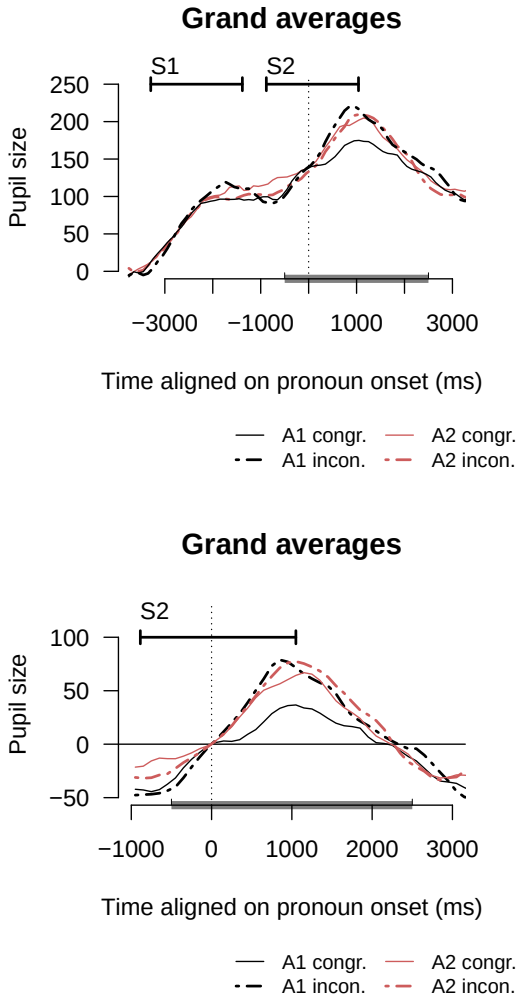
Figure 3. Example visual materials (see sentences in Table 1):

congruent with test sentence incongruent with test sentence

Figure 4. *Top:* Example of two pupil dilation time series, recorded from two different participants (solid versus dashed lines) in two different trials (red versus black lines). The data is aligned on the onset of the pronoun. The horizontal bars indicate the duration of the auditory stimuli of the two trials, which consisted of two sentences. The baseline for the averages in this graph was calculated from a 250 ms time window before sound onset. *Bottom:* The grand averages for the four conditions. In contrast with the plot above, the baseline window of this analysis data was at the pronoun onset, indicated with the vertical line.

Pupil dilation was measured with an EyeLink 1000 (SR research) eye tracker at 250 Hz. Before analysis, artifacts were automatically removed from the pupil dilation signal in R (R Core Team 2017). Blinks and saccades were detected automatically by using a velocity threshold (cf. Mathôt 2013). Artifacts were removed, without interpolation (with 100 ms and 20 ms padding for blinks and other artifacts respectively). Visual inspection of the data followed the automatic artifact rejection. Trials with more than 25% missing data were removed. The data was downsampled to 50Hz by taking the median per timebin. The baseline was calculated as the average pupil dilation in the time window of 250 ms at the beginning of each trial, more precisely 250-500ms after the picture appeared on the screen and immediately before the auditory stimuli started (i.e., 500ms after picture onset). The baseline was calculated per trial and subtracted from the pupil dilation measures, but the pupil dilation was *not* divided by the baseline, to avoid changing the dilation pattern. As the experiment investigated the processing of the pronoun ('him'), the data was aligned to the onset of the pronoun.

Figure 1 *Bottom* shows the data of two items recorded in two participants, after aligning the data on the onset of the pronoun. Figure 4 shows the grand averages of the data for each of the four conditions in the 2x2 design. Although the individual pupillary responses show considerable variation, the average curves show two clear peaks (Top panel): around 2000 ms before pronoun onset, which is after the introduction of the first agent, and around 1000 ms after pronoun onset. The pupil dilation between 500 ms before pronoun onset and 2500 ms after pronoun onset was analyzed.



Introduction GAMMs

In this paper we propose the use of Generalized Additive (Mixed) Modeling (GAMM; Lin and Zhang 1999; Wood 2006, 2011) for the analysis of pupillary data. GAMM is an extension of typical regression methods as it estimates the relation between a dependent variable and a number of given predictors. Instead of forcing the relation between dependent variable and predictor to be linear, as is the case in typical linear regression, this relation is modeled as a *smooth* function, which can, but does not need to be linear. A discussion of the technical aspects of smooth functions is beyond the scope of this paper (see Wood 2017a; or van Rij et al. to appear; Wieling submitted; Baayen et al. 2017, for an introduction), but in the context of this work it can be thought of as a continuous, potentially wiggly but not abruptly changing line that is expressed over time. GAMM approximates smooth functions as a weighted sum of a set of base functions that each have a different shape but that together form a smooth function that fits the (nonlinear) pattern of the data. It is possible to set these base functions to polynomials, but by default they are set to thin plate regression splines (tprs) for 1-dimensional smooth functions as these have more optimal properties for fitting unknown functions (for more information, see Wood 2017a, chapters 4 and 5). GAMM obtains the maximum likelihood estimates of the smooths using penalized regression methods (based on *penalized iteratively re-weighted least squares*). When multiple predictors are entered in the regression, GAMM will estimate the smoothing parameters for each smooth function using cross-validation (for details, see Wood 2006, chapters 3 and 4). The estimation procedures determining the smooth functions and parameters are designed to avoid

overgeneralization and overfitting of the data. GAMMs have been applied before to pupil dilation data (Lõo et al. 2016; van Rij 2012; Vogelzang et al. 2016), and to other measures in psychology, such as for the analysis of EEG/ERP data (e.g., Boehm et al. 2014; Nixon et al. 2015; Tremblay and Newman 2015; Hendrix et al. 2016), gaze data (e.g., Nixon et al. 2016; van Rij et al. 2016), articulography (e.g., Tomaschek et al. to appearb,t; Wieling submitted), reaction times (e.g., Baayen et al. 2017; Milin et al. 2017; Baayen 2010), or F0 contours (Kösling et al. 2013). We refer to these papers for a general overview of using GAMM for analyzing time course data (see, especially, van Rij et al. to appear; Wieling submitted).

We will first present an initial example of a GAMM model that predicts pupil dilation as a function of time in trial by condition and gaze position. To account for variation in participants and items, we include random effects for participants and items over time and gaze position. Although this model looks sensible at first, we will show that this model does not meet the assumptions of a regression model. As a result the model provides anti-conservative estimates, detecting effects that are actually not there. To clarify the structure of the models, we provide both a formal description and the R code to run the model.

Initial GAMM model

We start with a relatively simple model that estimates the effects of Introduction Order and Congruency on the pupil dilation trajectory. The model includes an interaction between the covariate *Time*, representing the time in the trial aligned with the onset of the pronoun (i.e., the word determining whether the sentence was congruent or incongruent with the picture), and the categorical predictor *Condition*, which is a four-level predictor that implements the interaction between the two manipulations in the 2x2 experimental design. This model will estimate four regression lines over time, one for each level of *Condition*.

Linear regression model. In a linear regression framework we could formalize such as model as follows: $p_i = \mu + \gamma_{c(i)} + \beta_{c(i)} t_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. In the model formalization, μ represents the intercept estimation, the baseline condition. The subscript $c(i)$ reflects the level of the categorical predictor Condition at observation i. The coefficient $\gamma_{c(i)}$ represents the intercept adjustment for the four levels of predictor Condition. t_i is the value of the continuous predictor Time of observation i, and $\beta_{c(i)}$ reflects the slopes for the four levels of Condition. In R, the following code is used for representing the linear model: `Pupil ~ Condition*Time`.

Nonlinear regression model. However, we know that pupillometric data does not show a linear trajectory over time. Generalized Additive Modeling allows fitting of nonlinear regression curves. The coefficients in a linear regression function that characterize the slope of the linear regression lines will be replaced with a smooth function f which now has to be estimated as part of the model fitting: $p_i = \mu + \gamma_{c(i)} + f_{c(i)}(t_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. The term $f_{c(i)}(t_i)$ indicates that for each level of Condition a different nonlinear regression line is fitted over Time.

Nonlinear interaction. To account for sudden drops and increases in pupil dilation due to changes in pupil position (e.g., Brisson et al. 2013; Gagl et al. 2011; Hayes and Petrov 2016), we included a nonlinear interaction between X and Y coordinates of the gaze positions, the predictors *Xgaze* and *Ygaze* at observation i: $p_i = \mu + \gamma_{c(i)} + f_{c(i)}(t_i) + f_2(x_i, y_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Implementation in R. The package `mgcv` implements GAMMs using the function `bam()`. The R code for this first preliminary statistical model, `model1`, is presented in R Screen 1. As first argument it takes the formula specifying the mathematical model. The function `s()` is used for fitting a 1-dimensional nonlinear regression line, and the argument `by` indicates that for each of the levels of Condition a nonlinear regression line has to be estimated. The shape of the nonlinear regression lines is not determined by the user, but estimated from the data using penalized regression methods. The argument `k` provides an upper bound to the order of basefunctions used to fit the regression lines. This argument is set to 10 by default, but here increased to 20 to allow to fit more wiggly patterns. The function `te()` normally fits a tensor product interaction to estimate a nonlinear interaction surface. However, as the X and Y position of the gaze are measured on the same scale, the function `s()` can be used to implement the nonlinear interaction that accounts for the changes in pupil size caused by gaze position.

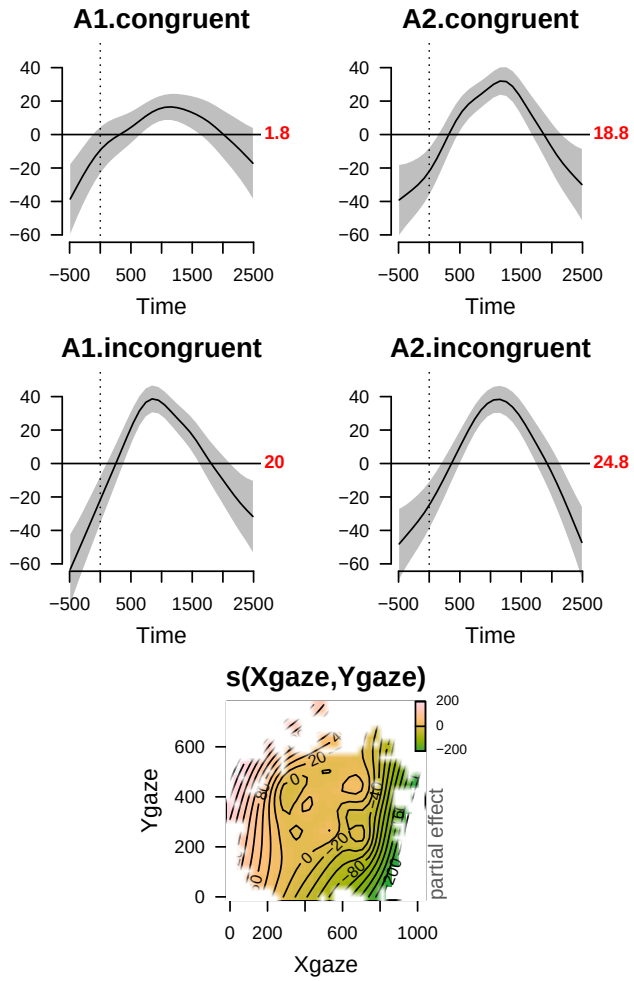
R Screen 1: Initial GAMM model. For presentation purposes only the fixed effects are included here. R Screen 2 shows the full model.

```
model1 <- bam( Pupil ~  Condition
+ s(Time, by=Condition, k=20)
+ s(Xgaze, Ygaze)
# ... random effects ...
, data=dat, discrete=TRUE)
```

Interpretation. The top four panels of Figure 5 show the estimated regression lines for each of the four conditions, as modeled with the smooth `s(Time)`. The x-axis of these graphs represents the time from the onset of the pronoun. The y-axis shows the pupil size estimations. However, the range of y-values is rather different than the data, see the Bottom panel of Figure 4. These plots only reflects the *partial* effect of the smooth over *Time*, and do not include the intercept or estimates for the other predictors in the model. The sum of the parametric effects (i.e., the intercept and intercept adjustment of Condition) that adjust the regression lines up or downward are specified on the right side of the plot. The condition 'A1.congruent' (topleft panel) shows the smallest peak amplitude in the regression line, and also the smallest intercept adjustment (1.8).

The estimated effect of the X and Y gaze positions on the pupil size are presented in the contour plot on the third row. The x-axis represents the X gaze position, the y-axis the Y-position. This graph reads like a map, with the contour lines

Figure 5. Partial effects (fixed effects only) of the initial GAMM model. The top four panels show the nonlinear regression lines for each of the four conditions with pointwise 95% confidence intervals, with the value of the parametric estimates for that condition on the right (red numbers). The bottom panel shows the interaction between Xgaze and Ygaze, and implements the effect of gaze position on the measured pupil size. Note that (0,0) represents the topleft corner of the screen.



and colors indicating the height of the pupil dilation. The screen areas without observations are empty (white). The contour plot indicates that the looks to the center values are represented by the average pupil size, because a value around zero needs to be added to the regression lines. When the gaze position moves to the right-top side of the screen (i.e., right-bottom side of the plot, as (0,0) represents the topleft corner of the screen), the pupil size measured is smaller, as indicated by the negative values of the contour lines and color coding. When looking more to the left-bottom side of the screen (i.e., left-top side of the plot), the pupil size measured is larger, as indicated by the positive values of the contour lines and color coding.

This initial GAMM model illustrates the advantage of including nonlinear regression lines and interactions: Instead of dichotomizing over continuous experimental factors such as frequency, working memory scores, or age, we include them as continuous predictors. This feature of GAMMs is especially useful for modeling pupillary response data. Control variables, such as gaze position, can be included as nonlinear smooths and interactions to take care of potential confounds. As such, GAMMs provide a good alternative for preprocessing procedures, such as corrections for baseline and gaze positions (e.g., Brisson et al. 2013; Gagl et al. 2011). The smooth functions in GAMM are driven by the data, and do not pose a priori assumptions on the shape and size of these effects.

Random effects. In the GAMM model presented in R Screen 1 we did not account for variation in participants and items. The pupillary response can vary significantly between participants, and it is particularly sensitive to factors that can differ between participants, such as fatigue, age, and medication, but also to learning, and fluctuations in attention (e.g., Winn et al. 1994; Watson and Yellott 2012; Beatty and Lucero-Wagoner 2000). In addition, items could modulate the pupillary response because of perceptual differences (e.g., louder speech signal, visual image with darker colors) or other stimulus properties, such as the linguistic content (e.g., high versus low frequency nouns, differences in grammatical complexity, etc). Regression models need to account for this variation, because otherwise the model residuals (i.e., the error, or unexplained part of the data) will also reflect systematic patterns, thereby violating the assumption of independent observations.

In a mixed-modeling framework, this type of variation could be modeled as random variation around a population mean (Pinheiro and Bates 2000). A random intercept for *Subject* would estimate a normal distribution with the variance based on the variation between participants. For each participant a value is selected from the distribution that models the difference between the mean pupil dilation and the participant's pupil dilation. Participants with extreme high or low pupil size measures are assigned values that are less extreme and closer to the mean, because these values are drawn from a normal distribution (shrinkage of the mean; e.g., Pinheiro and Bates 2000; Baayen et al. 2008, for introductions in random effects for linear regression mixed modeling). Because random effects only estimate the variance of the distribution, they use fewer parameters than fixed effects, and they allow for generalizing over participants and making new predictions (based on the fixed effects) for other people from the same population.

In GAMMs three types of random effects could be specified: *i)* random intercepts, which adjust the intercept of a (nonlinear) regression line or interaction surface, *ii)* random slopes, which adjust the slope of a (nonlinear) regression line or interaction surface, and *iii)* factor smooths, which adjust the shape of the regression line or interaction surface with a potentially nonlinear trend (see Figure 6). The parametric random effects, the random intercepts and random slopes are also available in linear mixed-effects models. As the factor smooths may also include adjustment of the intercept and slope of the regression line, these are generally not combined with random intercepts and slopes for the same predictors.

To capture the participant and item variation in pupil dilation trends, the initial GAMM model `model1` was extended with random smooths for participants and items. We could formalize this GAMM model with the following description of the pupil size p_{iab} at observation i, of

Figure 6. Schematic illustration of the three types of random effects. The y-axis respresents the measurement scale. The black thick line outlines the fixed effect estimate, whereas the dashed red lines illustrate how the random effects modulate the fixed effects. The bottom right panel separates the fixed effects from the random effects.

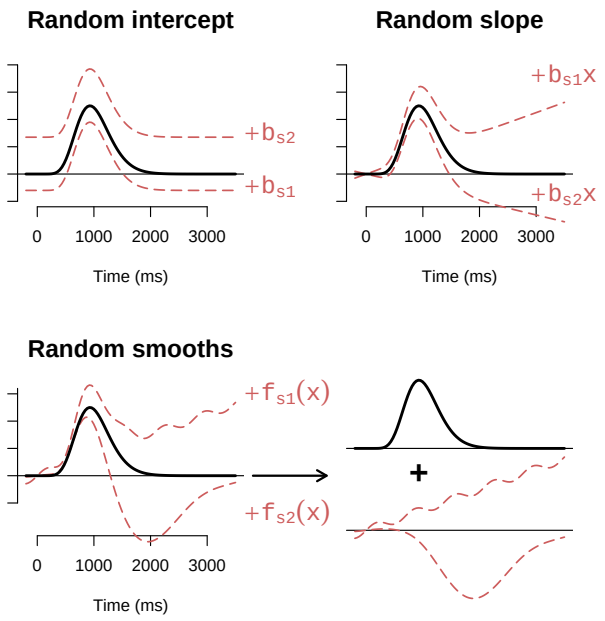
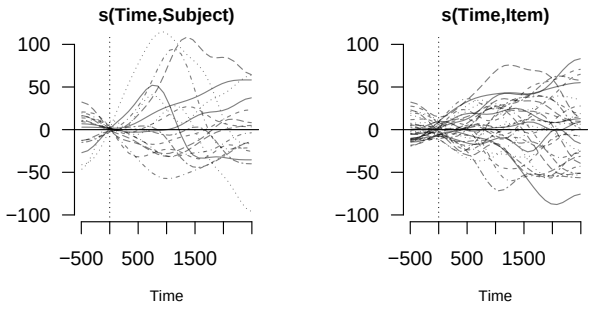


Figure 7. Random factor smooths for participants and items estimated by model `model1`.



participant a and with item b: $p_{iab} = \mu + \gamma_{c(i)} + f_{c(i)}(t_i) + f_2(x_i, y_i) + f_{s(a)}(t_i) + f_{i(b)}(t_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

R Screen 2: Initial GAMM model with nonlinear random effects.

```
model1 <- bam( Pupil ~  Condition
+ s(Time, by=Condition, k=20)
+ s(Xgaze, Ygaze)
+ s(Time, Subject, bs='fs', m=1)
+ s(Time, Item, bs='fs', m=1)
, data=dat, discrete=TRUE)
```

The complete code for the first preliminary statistical model, `model1`, is presented in R Screen 2. The package `mgcv` specifies factor smooths with the basis `bs='fs'`. Because we also included general smooths of *Time*, the random smooths of *Time* are actually random adjustments from these general smooths. Figure 7 shows the random adjustments for *Subjects* and *Items* estimated by model `model1`.

Testing for significance

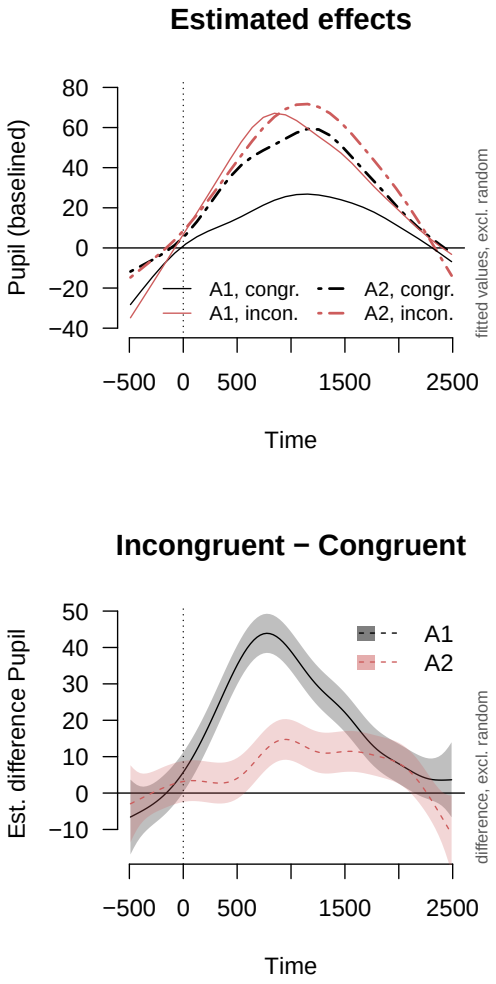
When using GAMMs there are various ways to determine whether the experimental manipulations influenced the pupil size. Here we will use *i)* visual inspection of the model's estimates of the differences between the conditions, *ii)* a model-comparison procedure, and *iii)* inspection of the model summary statistics to determine the differences between conditions. These methods are complementary, because they provide different types of information.

Visual inspection. The output of a GAMM does not present a description of the nonlinear regression lines, because the smooth functions often cannot be captured by a few coefficients. Instead the summary provides information on the wiggliness of the regression line, and whether the line is (somewhere) significantly different from zero. Visualization is necessary for interpreting the nonlinear terms. Figure 5 and Figure 8 show the estimated regressions lines in two different ways. As described earlier, Figure 5 shows the *partial* effects, which are the estimated smooth functions. Each plot represents one term in the model summary. In contrast, Figure 8 (Top panel) shows the *summed* effects (or fitted effects), which include the intercept and a value for every other predictor as well. The predictors that are not visualized are set to their median value or reference level. The summed effects in Figure 8 (Top panel) reveal that the pupil size is reduced when the sentence is congruent with the picture and the actor is introduced first (i.e., condition 'A1.congruent', the solid thin black line) in comparison with the other conditions.

Figure 8 (Bottom panel) shows the *differences* in congruency for the two types of introduction. The positive differences suggest that the sentences that are *incongruent* with the pictures elicit more pupil dilation than the sentence that are *congruent* with the pictures. However, the difference is modulated by the introduction order: the difference is considerably larger when the actor is introduced first (i.e., introduction order 'A1', black solid line) than when the actor is introduced as second referent (i.e., 'A2', red dashed line). In addition, when the actor is introduced first ('A1') the congruent and incongruent sentences elicit a difference in pupil size immediately at the onset of the pronoun, but when the actor is introduced second ('A2') the difference in pupil size arises only around 500 ms after pronoun onset. This pattern suggests an interaction between Introduction Order and Congruency (as implemented in the four-level factor Condition) in the pupil size trajectories, rather than two separate main effects of Introduction Order and Congruency.

Visualizing the summed effects and the estimated differences between conditions provide a fast way to inspect the model's predictions, as they do not require running alternative models. However, these difference estimates may provide a misleading picture when the model does not fit the data well, for example when the model fails to account for all the structure in the data. When model validation signals problems with the model fit, one should be careful with the

Figure 8. Estimates of the initial GAMM model `model1`. *Top:* Summed effects for all conditions, with the random effects set to zero. *Bottom:* Difference curves, derived from `model1`. The gray solid line represents the estimated difference (and pointwise 95% confidence intervals) between the incongruent and congruent items when the actor is introduced first ('A1'), and the dashed red line represents the estimated difference (and pointwise 95% confidence intervals) between the incongruent and congruent items when the actor is introduced second ('A2').



interpretation of the estimated differences. We will work this out further in the following sections on model criticism.

Model-comparison procedure. A model comparison procedure is a second method to assess whether the interaction between Introduction Order and Congruency is indeed contributing significantly to the model. We could compare a model that includes the interaction between these two predictors with a model that does not include the interaction between these predictors. The models could be compared using a generalized likelihood ratio test or with an AIC comparison. By default, the smoothing parameter selection score is set to fREML (fast restricted maximum likelihood) when using `bam()`. However, (f)REML scores are not comparable between models with a different fixed effects structure. Instead, the selection method should be set to ML (maximum likelihood) when comparing fixed effects (e.g., Wood 2017a, chapter 2). Therefore, we refitted model `model1` with method ML, and compared it with model `model2` as presented in R screen 3. This model captures the main pupil dilation trend with two regression lines, one for congruent items and one for incongruent items. The main effect of Introduction Order is captured by a binary predictor *IsA1*, which takes the value 1 when the actor is introduced first and the value 0 when the actor is introduced second. The regression line modeled by the smooth term `s(Time, by=IsA1)` hence fits the *difference* between the two types of introduction sentences.[4]

R Screen 3: GAMM model with separate terms for Congruency and Introduction Order.

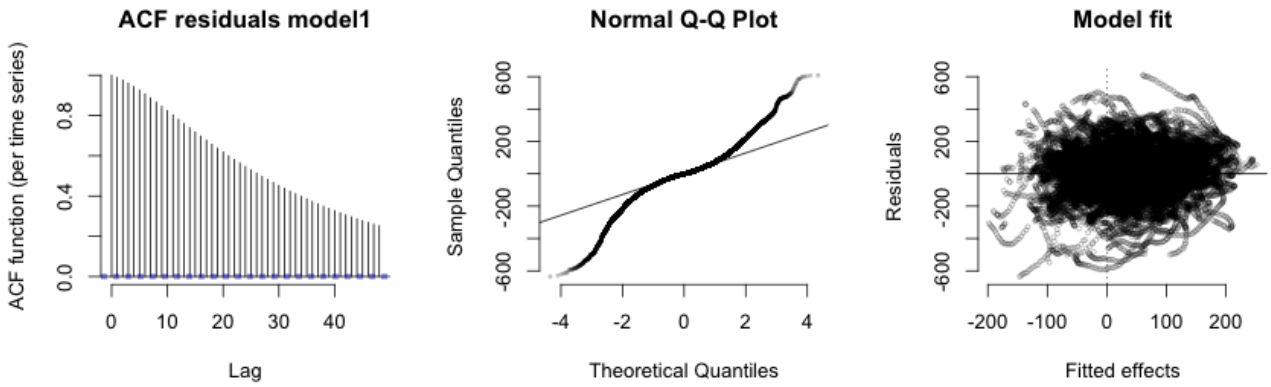
```
dat$IsA1 <- ifelse(
  dat$IntroductionOrder=="A1", 1, 0)

model2 <- bam(Pupil ~ Congruency
  + s(Time, by=Congruency, k=20)
  + s(Time, by=IsA1, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Subject, bs='fs', m=1)
  + s(Time, Item, bs='fs', m=1)
  , data=dat, method='ML')
```

Comparisons between `model1` and `model2` confirm that `model1`, which implements the interaction between Introduction Order and Congruency is preferred ($\chi^2(3) = 54.08, p < 0.001$; $\Delta\text{AIC} = 112.50$) over a model that separates the effects Introduction Order and Congruency.

A disadvantage of a model-comparison procedure is that multiple statistical models have to be estimated. As pupillometric data sets often consist of a large amount of measurements, this may take a long time. In addition, the ML estimation required for comparing fixed effects takes much more time than running a model with fREML. An efficient strategy is to start with the evaluation and optimization of the initial model based on visual inspection and the summary statistics, and verify the conclusions with a backward-fitting model comparison procedure.

Summary statistics. The summary of the smooth terms (nonlinear components of the GAMM model) shows for each nonlinear regression line whether this line is significantly different from zero and how wiggly the regression line is (i.e., by reporting the *edf*, the effective degrees of freedom). However, the summary does not show whether the regressions lines for each of the levels of the predictor Condition are different from each other. Only when we explicitly model the *differences* between conditions, the summary reports whether these difference curves are significantly different from zero. In `model3` we have replaced the four regression lines for each of the four conditions of `model1` by a reference curve and three binary difference curves implementing the effects of Introduction Order, Congruency and their interaction, see R screen 4. The summary statistics indicate that the three difference curves are significantly different from zero: *IsCongr* (F(4.69, 71628.77)=10.38; p<.001), which implements the difference between congruent and incongruent items, and

Figure 9. Residuals of the initial GAMM model `model1`.



the effect of *IsA1* (F(7.71, 71628.77)=9.19; p<.001), which implements the difference between the actor-first introduction and the actor-second introduction, and *IsA1Congruent* (F(7.48, 71628.77)=15.35; p<.001), which implements the interaction effect that is needed to model the difference between the conditions 'A1.congruent' and 'A1.incongruent' (in addition to the main effects of *IsA1* and *IsCongruent*).

Although the summary statistics are very useful for reducing the number of statistical models to run, they have their own limitations. For complex interactions (more than four conditions) binary curves are difficult to interpret. A second limitation is that the summary statistics do not tell *where* the difference curves are different from zero, nor the amplitude of the difference. Visualization is needed for interpreting the results. And similar to the other methods, the statistics should be treated with caution when the model does not fit the data well, as we will explain in the following section.

R Screen 4: GAMM model with a set of binary predictors modeling the four experimental conditions.

```
dat$IsCongr <- ifelse(
  dat$Congruency=="congr", 1, 0)
dat$IsA1 <- ifelse(
  dat$IntroductionOrder=="A1", 1, 0)
dat$IsA1Congr <- dat$IsA1 * dat$IsCongr

model3 <- bam(Pupil ~ s(Time, by=20)
  + s(Time, by=IsCongr, k=20)
  + s(Time, by=IsA1, k=20)
  + s(Time, by=IsA1Congr, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Subject, bs='fs', m=1)
  + s(Time, Item, bs='fs', m=1)
  , data=dat, method='ML')
```

Model criticism

To evaluate the model fit, we generally look at the residuals because they are the deviation between the observed values in the data and the estimated values of the model. Figure 9 shows different aspects of the residuals. The Left panel shows an ACF plot, which visualizes the correlation between the residuals and the lagged residuals, i.e., the residuals of earlier measurements. Correlations between temporally adjacent residuals indicates that there is structure in the residuals that is not captured by the model. The Center panel shows a QQ (quantile-quantile) plot, which compares the residuals with a normal distribution (indicated by the straight line). And the Right panel plots the residuals against the fitted values. The plots in Figure 9 reveal some problems with the fit of our initial GAMM model: *a*) the residuals are heavily autocorrelated (the *Left* panel shows high values for Lag 1 and following lags); *b*) the residuals are heavier tailed than the normal distribution (the *Center* panel shows that the residuals deviate from the straight line); and *c*) the residuals of the model are quite large in comparison with the effect sizes (the *Right* panel shows a larger y-range than x-range). In addition, the Right panel confirms the high autocorrelation, because the residuals show clearly detectable time series: the residuals look like threads, instead of a cloud of random dots.

In the next section we present a series of simulations to investigate the problems that we need to address when analyzing the time course of pupil dilation. We then present new analyses of the real pupil dilation data set that address these concerns.

Autocorrelation in residuals

Autocorrelation of errors violates the assumption of regression analyses that the errors are independent. Violating this assumption may underestimate the standard errors, and hence reduce the reliability of our GAMM analysis. To understand the potential source of the high autocorrelation in our pupil dilation data we ran simulations of autocorrelated data.

In each simulation, 250 randomly sine waves were generated with randomly modified amplitudes: $y = a \cdot \sin(x) + u$, $a \sim \mathcal{N}(1, .25)$, $u \sim \mathcal{N}(0, .25)$. Parameter a is the amplitude modification, parameter u is the random (independent!) noise added to the signal.

The first simulated data set consisted of 250 simulated trials without noise added ($u = 0$). A simple GAMM model was fitted to estimate the mean trend over x: $\texttt{y} \sim \texttt{s(x)}$.

Although no noise was added, the residuals of the model are highly correlated (see right panel of Figure 10), because the model does not capture the differences in amplitude between simulated trials.

When independent noise was added to the same simulation data ($u \sim \mathcal{N}(0, .25)$), the autocorrelation was reduced in a GAMM with exactly the same model specification. A similar GAMM model was fitted to estimate the mean trend over x: $y \sim s(x)$. When a new data set is created with less variation in amplitudes between the individual trials ($a \sim \mathcal{N}(1, .10)$), but the same noise distribution ($u \sim \mathcal{N}(0, .25)$), the smaller variation in amplitudes further reduces the autocorrelation (see Supplementary Materials).

These example simulations show that autocorrelation may reflect differences between the trials for which GAMM is fitting a mean trend. The residuals of the GAMM models are determined by the differences between each individual trial and the estimated trend over time. The residuals are autocorrelated, because the differences between a measurement of the trial and the estimated trend unfold relatively gradually over time. These differences are particularly clear when the amount of noise on the signal is relatively small, an inherent property of slow changing signals such as pupil dilation data. Measurements that contain a large amount of noise and change rapidly over time, are less likely to elicit autocorrelation, because these two factors reduce the correlation between the residuals.

Autocorrelation in the residuals will arise if a general non-linear trend is fitted to data with a large variation in individual trends and a small measurement noise. The most intuitive solution would be to include a random wiggly curve for each individual trial of each individual participant in the model as random effect, in addition to the smooth functions for the different experimental conditions as fixed or random effects. However, even including all these individual time series as random effects often does not remove the autocorrelation completely, because the fitted curves are smoothed and allow for differences with the general non-linear trend, which are a source of autocorrelation.

Correcting for autocorrelation by improving model fit

We applied this method to our pupillometry data using the model as specified in R Screen 5: Instead of random smooths of Time for Subjects and Items, we included a random smooth of Time for each individual time series *Event* (unique combination of Subject and Item). The random smooths for each time series inform the model that the measurements within a time series are not independent. As a result the confidence intervals around the model's estimates have increased in comparison with the estimate of our earlier model. Model comparisons and the statistical information from the summary both confirm that the interaction between *Time* and *Condition* should be included in the model, but not all levels are significantly different from each other anymore. There is still a significant difference between the two conditions with an actor-first ('A1') introduction order, but the difference between the two conditions with an actor-second ('A2') introduction order has disappeared.

R Screen 5: GAMM model with nonlinear random effects for individual time series.

```
# Create column Event:
dat$Event <- interaction(dat$Subject,
  dat$Item, drop=TRUE)

# New model:
model4 <- bam(Pupil ~ Condition
  + s(Time, by=Condition, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Event, bs='fs', m=1)
  , data=dat, discrete=TRUE)
```

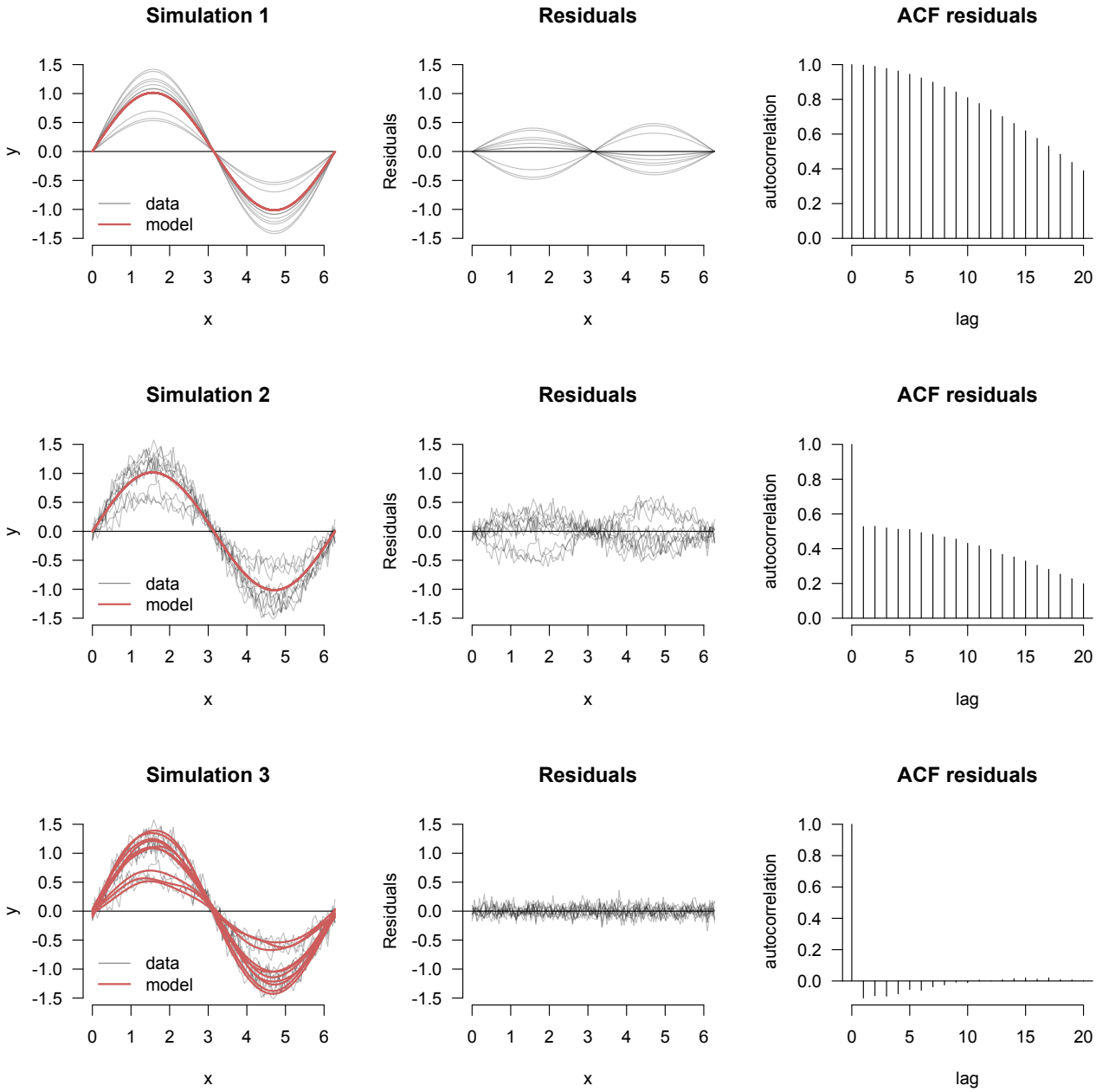
Inspection of the residuals indicates that the model fit has drastically improved in comparison with the first model. The fitted values (Figure 11, *Bottom* panel, thick red lines) follow roughly the same trajectory as the measured pupil dilation (Figure 11, *Bottom* panel, black lines). This conclusion is supported by the very high correlation of 0.996 (only 0.49 for the first model). The residuals are still highly autocorrelated, but the shape of the average autocorrelation graph is different (positive correlations for lower lags, and negative correlations for longer lags). More importantly, the residuals are smaller (median absolute residuals=4.68) than the residuals of the first model (median absolute residuals=43.34, compare also the ACF plots of Figures 9 and 11). With smaller residuals, i.e., a better fit of the data, the correlation in the residuals is less likely to affect the confidence of the model.

However, the proposed analysis with random factor smooth for each individual time series (currently) only works for relatively small experiments. The predictor *Event* codes the unique time series as a combination of participants and items. In our experiment we analyzed the data of 17 participants and at most 32 items, resulting in only 507 unique time series. However, many psycholinguistic experiments will result in more than 1000 time series. Estimating random factor smooths for all these time series is computationally very demanding, and (currently) only possible with a powerful server (even when setting the argument `discrete` to true, for more efficient storage and processing).

When including a random factor smooth for each unique time series is not possible, a sensible compromise is to include a random intercept and a random slope for each time series in addition to random factor smooths for Subjects and Items. This allows random variation in the intercept and slope of each time series separately, whereas the random smooths for Subject and Item specify random adjustments in the shape of the smooth (see the Supplementary Materials for the effect of different random effects structures in nonlinear regression models). R Screen 6 shows the the model with random intercepts and slopes included.

The model's fit of each time series is less precise than that of the model with random effects smooths for each time series, which is visible in the slightly lower correlation of 0.9. However, the model's fit is still somewhat better than the fit of the first model that only included random smooths for

Figure 10. Autocorrelation in simulation data (n=250). For each simulation (Simulation 1 in the top row, Smulation 2 in the center row, and Simulation 3 in the bottom row) the same three plots are provided. *Left*: 10 randomly selected modified sine waves (of the 250 in total), and the model fit for the sine waves (red thick solid line); *Center*: residuals of the model for the same 10 sine waves; *Right*: autocorrelation of the model's residuals for each lag (x-axis).

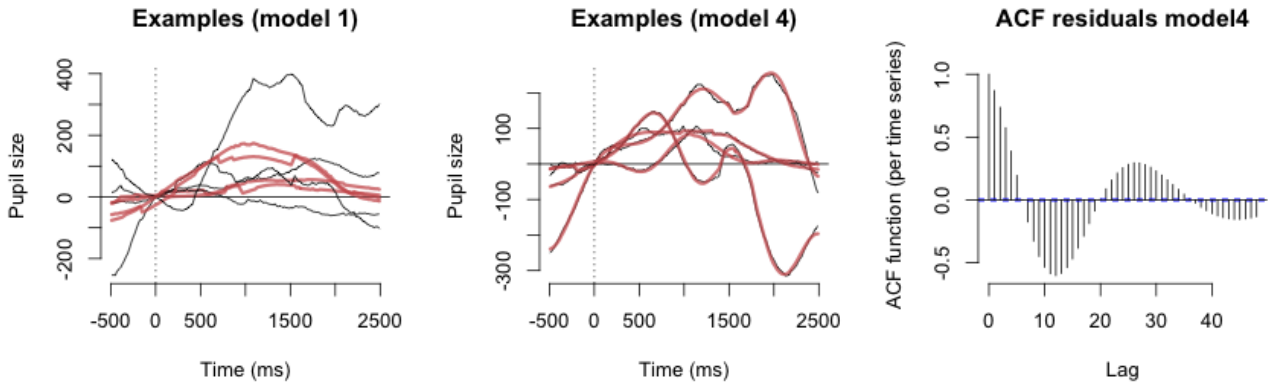


Subjects and Items (i.e., 0.49). The residuals also show that the current model is a compromise between the two previous models: the median absolute residuals of 24.42 is smaller than the residuals of the first model (43.34), but larger than that of the model with random factor smooths for each time series (4.68).

Correcting for autocorrelation by including AR1 model

An alternative way to account for autocorrelation in the residuals is to include an AR1 model within a GAMM model. Our simulations suggested that the autocorrelation in the errors is caused by the large variation in individual time series, and the high signal-to-noise ratio in the pupil dilation data. However, we cannot exclude the possibility that autocorrelation is (partly) due to an autoregressive (AR) process in the data that is not captured by the model. Note that GAMMs cannot distinguish between these sources of autocorrelation. To remove autocorrelation effects that are potentially caused by AR processes, we included an AR1 error model for the residuals in GAMM (Wood et al. 2014; Wood 2017a). An AR1 model is a linear model that estimates influence of the immediately preceding measurement on the current measurement in a time series: $X_t = \rho X_{t-1} + \epsilon_t, \epsilon \sim \mathcal{N}(0, \sigma_X)$. Wood (2017a) recommends to find the optimal value of ρ by comparing the fREML scores of the models that include different values of ρ. Figure 12 (Top panel)

Figure 11. Improvement in model fit by adding random smooths for unique time series. *Left:* Data (black thick lines) and the model fit of the initial GAMM model (thick red lines) for three random three events in the experiment. *Center:* Data (black thick lines) and the model fit of the improved GAMM model (thick red lines) for the same three time series. *Right:* ACF for the improved GAMM model, `model4`. (The ACF of `model1` is presented in Figure 9).



R Screen 6: GAMM model with random intercept and slope for individual time series.

```
model5 <- bam(Pupil ~ Condition
  + s(Time, by=Condition, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Subject, bs='fs', m=1)
  + s(Time, Item, bs='fs', m=1)
  + s(Event, bs='re')
  + s(Time, Event, bs='re')
  , data=dat, discrete=TRUE)
```

R Screen 7: GAMM model including AR1 model.

```
# mark start of time series:
dat <- start_event(dat, column="Time",
    event="Event")
# order data:
dat <- droplevels(dat[order(dat$Subject,
    dat$Trial, dat$Time),])
model4 <- bam(Pupil ~ Condition
  + s(Time, by=Condition, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Event, bs='fs', m=1)
  , data=dat, discrete=TRUE
  , AR.start=dat$start.event, rho=0.87)
```

shows the fREML scores for a range of ρ values in our pupil size data. For the initial model (i.e., `model1`) with only random factors smooths for participants and items there does not seem to be an optimal value for ρ, as the fREML scores keep improving with higher values. However, when Event is included as random effect, the improvement in fREML scores seem to decrease for higher values of ρ. The ACF Lag 1 score of the models (indicated with a dot in the plot) seems to be a reasonable estimate for the value for ρ.

Figure 12 (Bottom panel) shows the correlation between the data and the model fit for the same range of ρ values. This plot highlights the fact that including an AR1 model generally does not improve the model fit, but mainly increases the uncertainty over the estimates. As the residuals in `model1` are much larger than in the models `model4` and `model5` (as the model does not explain the data very well), high values of ρ affect the model fit much more as it does affect the other two models. The estimates for models with a poor fit could dramatically change when including an AR1 model, leading to different conclusions.

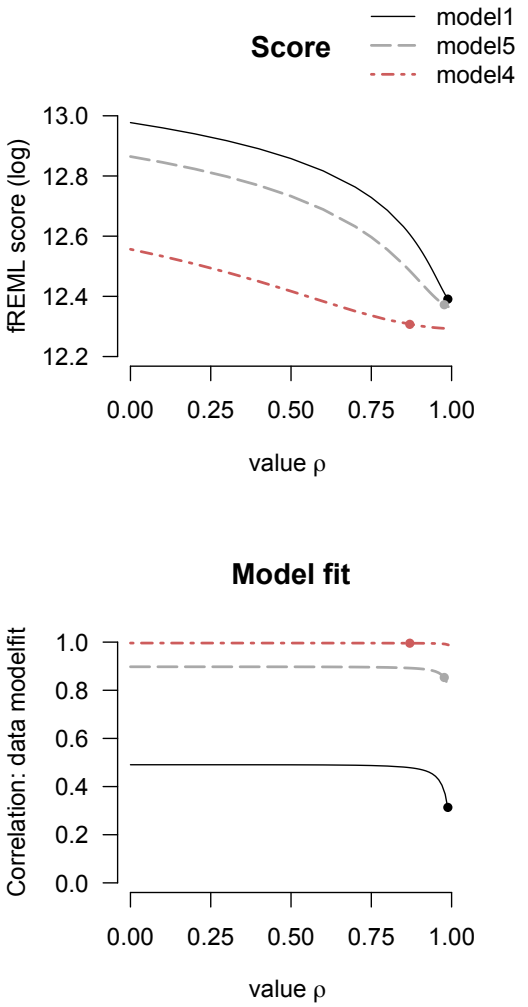
We set ρ to a very high value of 0.87 (based on the ACF Lag 1 score of model `model4`, but we also used a model comparison procedure to verify the optimal value).

Including an AR1 model increases the uncertainty in the predictions of a GAMM. The correction for autocorrelation takes place within the fitting of the model, and influences the estimations of the predictors. The differences between the conditions may become stronger, although the confidence intervals increase, but we have also seen examples where the correction for autocorrelation reduced all effects. In the current data only the difference between the congruent and incongruent items with an actor-first introduction sentence ('A1') is found to be significant, but not with an actor-second introduction sentence ('A2').

Although this method reduces the effect of the autocorrelation in the residuals, it also has some limitations. In the current version of the statistical software (mgcv 1.8-23), a single AR1 process is applied to all trials by setting a single value for ρ, the AR1 correlation parameter. However, inspection of the residuals of separate participants shows large differences in correlation structure. Therefore, this method does not completely remove the autocorrelation in the residuals and for trials with little autocorrelation, it may artificially induce autocorrelational structure (Baayen et al. to appear). Further, only first-order AR processes are currently implemented. Higher-level autocorrelation structure cannot be removed.

Thus, we have explained that autocorrelation in residuals will arise in time course analysis of pupil size data when the model does not fit the data well. Therefore, improving the model fit is the most important solution, for example by including a random effects structure that captures each time series. However, if improving the model fit is

Figure 12. Determining the optimal value of ρ. *Top:* Decrease in fREML scores with increasing values for ρ for different models discussed in this paper. *Bottom:* Correlation between model fit and data as indication for the precision of model fit plotted against the values of ρ. In both panels, the dots show the selected start values for ρ, based on the ACF Lag 1 value.



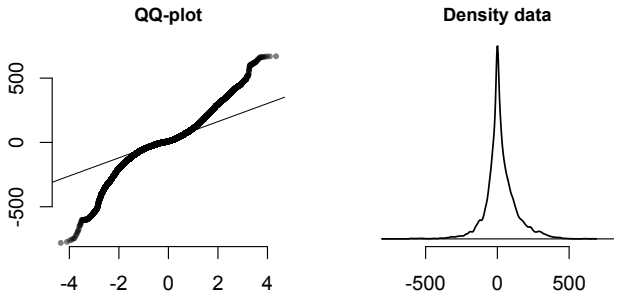
(computationally) not possible, including an AR1 model may provide an alternative solution.

Besides the autocorrelation in the residuals our initial model also did not show normally distributed residuals, which is assumed when running a regression model for Gaussian data. Although this issue is not likely to affect the model's estimates as severely as autocorrelation in the residuals, nevertheless the model's estimates are less reliable when the assumptions are not met. The following section addresses the issue of not normally distributed residuals.

Distribution of residuals

Generalized regression models allow for modeling data that is not normally distributed. For example, logistic regression models are frequently used in our field to model binomial data such as answer accuracy (correct/incorrect). Figure 13 visualizes the distribution of the pupil size data (after baseline subtraction) with a QQ-plot and a density plot. The plots clearly shows that the measurements are not normally

Figure 13. Distribution of the data.



distributed: the lower extreme values are much lower than would be expected with a normal distribution, and the higher extreme values are much higher than would be expected with a normal distribution (i.e., the distribution has heavier tails). It is difficult to correct this symmetrical pattern using transformations, such as the log or an exponential function. Instead, the package `"mgcv"`allows to model this type of data as a scaled-t distribution for heavy tailed response variables (Wood et al. 2016).

We re-run our best-fitting model with the scaled-t distribution specified. On the basis of this new model, we determined a new value for the ρ-parameter in the AR1 model. The code for our final model, with `family='scat'` and `rho=0.92` and random factor smooths for each time series (specified by the predictor Event), is presented in R Screen 8.

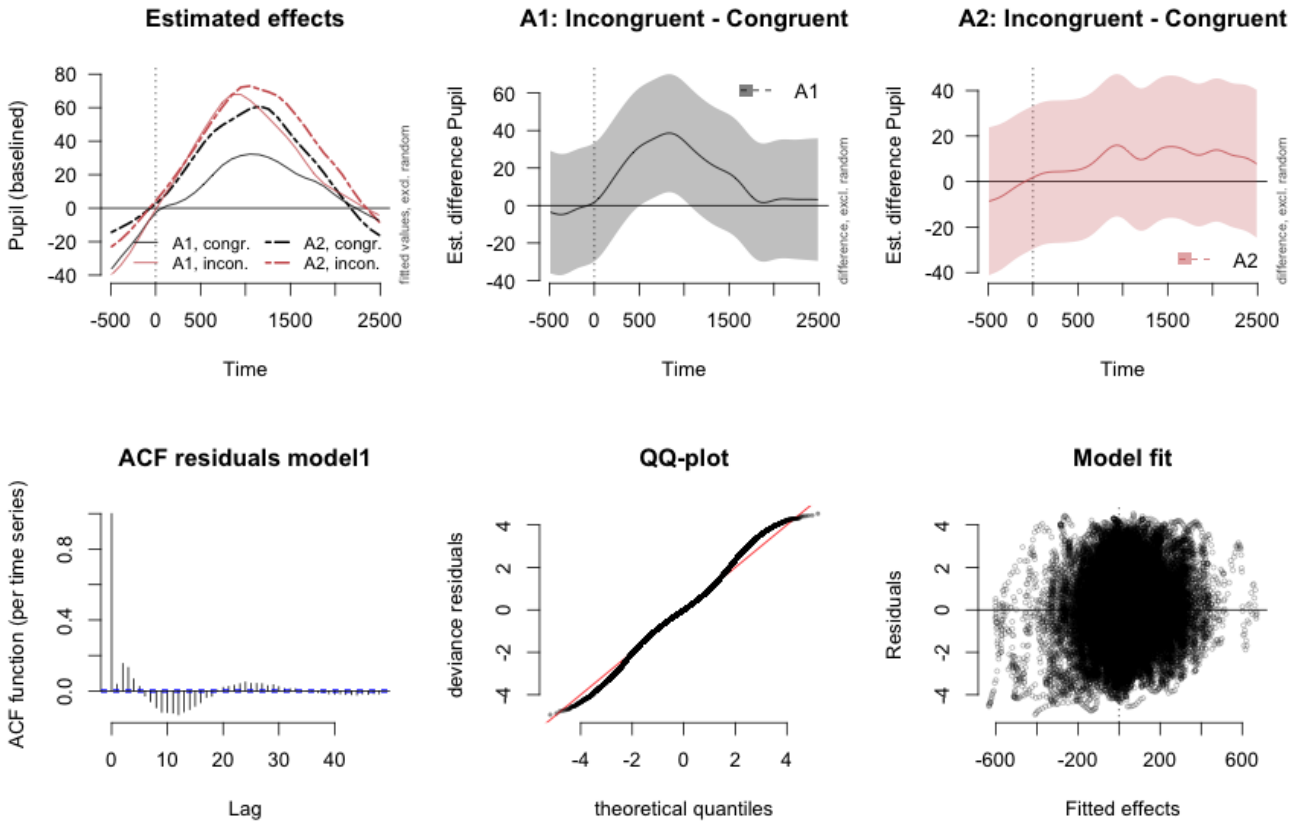
R Screen 8: Scaled-t GAMM model.

```
model6 <- bam(Pupil ~ Condition
  + s(Time, by=Condition, k=20)
  + s(Xgaze, Ygaze)
  + s(Time, Event, bs='fs', m=1)
  , data=dat, discrete=TRUE
  , AR.start=dat$start.event, rho=0.92
  , family='scat')
```

Figure 14 evaluates our final model. The top row shows the model's estimates for the four conditions and for the differences between the incongruent and congruent items for each Introduction Order. Interestingly, the model's estimates have not changed much in comparison with `model5`: with an actor-first introduction sentence ('A1'; Center panel), there is a significant difference between the incongruent and congruent items from 500-1000 ms after pronoun onset. The pupil dilation is lower for the congruent items than for the incongruent items. This difference is not found for the actor-second introduction sentence ('A2'; Right panel). Actually, only the condition 'A1.congruent' shows a significantly lower peak dilation than the three other conditions (Left panel).

The bottom row visualizes the residuals, which look rather different from the residuals of our initial GAMM model (see Figure 9). The autocorrelation is reduced by including the individual time series as random effects, and by including an AR1 model ($\rho = .92$) to account for the

Figure 14. Evaluation of the scaled-*t* model `model6`. *Top row:* Estimated effects (Left panel) and estimated differences with pointwise 95% confidence intervals (Center and Right panels). *Bottom row:* Residuals of model `model6`.



remaining autoregressive processes in the data (Left panel). To check whether the scaled-*t* distribution did capture the distribution of the data, the Central panel shows a QQ-plot of the residuals. Note that for a generalized regression model it is preferred to plot the standardized residuals, because the raw residuals of a generalized model do not behave like normally distributed residuals (Wood 2017a, Chapter 3). Although the standardized residuals do not completely fit the expected distribution, their distribution has improved considerably.

Discussion

In this paper we have analyzed a pupil dilation experiment (van Rij 2012) that investigated the effects of visual context and the introduction order (i.e., the order in which the two characters are introduced) on object pronoun processing in Dutch. Instead of analyzing various features that describe the pupil dilation trajectory, such as the peak amplitude and peak latency, we have examined the time course directly using Generalized Additive (Mixed) Models (GAMMs; Wood 2017a). This nonlinear regression method is interesting for analyzing pupil dilation data in hearing research, for several reasons: *a*) GAMMs allows us to quantify differences in the time course of pupil dilation directly, without the need to make any assumptions with respect to the shape of the trajectory on beforehand; *b*) GAMMs can model complex nonlinear interactions, such as the effect of gaze position on the pupil dilation curve. *c*) GAMMs can include nonlinear random effects. Using nonlinear random effects, the regression models can account for the large variation in pupil dilation time series.

An important advantage of time course analyses over traditional methods is that time course analyses allow for asking different questions, such as at which moment in time conditions start to differ and whether this difference changes over time. Another major advantage of GAMMs over traditional methods is the combination of (nonlinear) random effects and complex nonlinear interactions. This combination provides new possibilities for experimental designs, because dichotomization or discritization of a numeric predictor into a factor with two or more levels is not necessary anymore. The possibility to estimate complex nonlinear interaction surfaces provides a system for correcting for the effects of interfering factors as gaze position and baseline within the statistical analysis (in contrast to a separate correction on the data, e.g. Gagl et al. 2011), but could also be used for modeling nonlinear properties of items and participants. Especially in hearing research many continuous covariates are collected that describe participants' hearing abilities and cognitive abilities, such as working memory scores, and covariates that describe the signal properties, such as the frequency of the first formant, the pitch height, or the signal-to-noise ratio. GAMMs allow for including these covariates and their potentially nonlinear interactions as continuous covariates.

A comparison between different time course methods is outside the scope of this paper, but it is worth noting that

GAMMs can actually implement Growth Curve Analysis (GCA) and Functional Data Analysis (FDA) (see Wood 2017a, for examples).

The second aim of this paper was to explain the issues that arise with all time course analyses, and to show potential solutions. A very important aspect of all time course analyses is a thorough evaluation of the model fit. Visualizing the residuals is a good starting point for detecting problems with the model fit, such as autocorrelation in residuals and residuals that do not follow a normal distribution. Autocorrelation increases the probability for Type I errors (detecting an effect that is not really there) and may yield conclusions that are not replicable. Therefore, when using time course analyses, it is extremely important to apply model criticism procedures and to report about the model evaluation.

On the basis of our simulations, we argue that two major sources of autocorrelation are *a*) differences between the model fit and the data; and *b*) the nature of the pupil dilation signal, which is a slow signal with a relatively small amount of noise.

The most important method to avoid anti-conservative conclusions is to improve the model fit. A better model fit implies smaller residuals and potential autocorrelation of smaller residuals is less likely to affect the model than autocorrelation of large residuals. In linear mixed-effects modeling a maximum random effects structure would include random intercepts and random slopes for the design predictors for participants and item. However, in time series data, each event (a particular trial for a particular participant) consists of a series of observations with a nonlinear structure. Therefore, we cannot assume that the combined effects of participants and items together define each time series. Rather, in time course analyses a maximum random effects structure for time series data would include a random smooth for each individual time series. This is currently computationally not always possible for larger data sets. When including a random smooth for each time series is not possible, it is not immediately clear what would then be the next best alternative. We propose to include a random intercept and a random slope for each unique time series on top of random smooths for participants and items to provide more flexibility for the model to fit the unique time series.

When improving the model fit is not sufficient to eliminate the presence of autoregressive processes in the data, GAMMs offer the option to include an AR1 model to account for the autocorrelation in the residuals and adjust the estimates and confidence accordingly. However, including an AR1 model may not prevent Type I errors when the model accounts for only a small proportion of the variance in the pupil dilation data.

Note that autocorrelation in residuals is not a particular property of pupil dilation analyses, but also applies to other time scourse measures such as EEG, pitch contours, or articulography (EMA). However, in measurements with a lower signal-to-noise ratio (i.e., comparing the noise amplitude with the signal amplitude of the signal) such as EEG, the autocorrelation is less severe than in pupil dilation. Our simulations show that noise interrupts and reduces the autocorrelation in the residuals that arise when the the model does not provide a good fit for the individual time series.

In conclusion, pupil dilation is a sensitive measure of cognitive processing, but the measure is easily confounded with other factors eliciting pupil dilation. Careful experimental design and analysis are necessary to be able to interpret pupil dilation results. GAMMs are particularly suited for analyzing pupil dilation, because they allow to investigate the time course of pupil dilation directly and to correct for factors such as gaze position and participant differences. Although time course analyses require a thorough evaluation to avoid anti-conservative conclusions, they provide a more complete description of the data than traditional analyses of pupil dilation.

Notes

1. This issue seems to be unknown to many researchers analyzing pupillometric data, because the majority of aforementioned studies that applied one of the time series analyses to their pupillometric data (i.e., Jackson and Sirois 2009; Kuchinsky et al. 2013; van Rij 2012; Vogelzang et al. 2016; Winn et al. 2015) did not report whether this problem did play a role in their analysis and how they corrected their analyses. Therefore, this tutorial is particularly relevant for this field.
2. The Supplementary Materials are available at `https://git.lwp.rug.nl/p251653/analyzing-time-course-pupil-data`.
3. If an eye tracker that is not head mounted reports pupil size in mm, these results should be interpreted with caution until these results are calibrated offline.
4. Binary curves do not need an additional parametric intercept adjustment, as they fit only one regression line. See Supplementary Materials for more information.

Acknowledgements

This research was supported by grants from the Netherlands Organisation for Scientific Research NWO (Veni grant no. 275-70-044, van Rij; Vici grant no. 277-70-005, Hendriks) and the Alexander von Humboldt Foundation (Humboldt Professorship award, no. 1141527, Baayen). We would like to thank the anonymous reviewers for their constructive comments.

References

Baayen RH (2010) The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In: Olson E (ed.) *New impulses in word-formation*, Linguistische Berichte Sonderheft 17. Buske, Hamburg.

Baayen RH, Davidson DJ and Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412.

Baayen RH, van Rij J, de Cat C and Wood SN (to appear) Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In: Speelman D, Heylen K and Geeraerts D (eds.) *Mixed Effects Regression Models in Linguistics*. Berlin, Springer.

Baayen RH, Vasishth S, Kliegl R and Bates D D M (2017) The cave of Shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94: 206 – 234.

Beatty J (1982) Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91(2): 276–292.

Beatty J and Lucero-Wagoner B (2000) The pupillary system. In: Cacioppo JT, Tassinary LG and Berntson GG (eds.) *Handbook of psychophysiology*. New York: Cambridge University Press., pp. 142–162.

Boehm U, van Maanen L, Forstmann B and van Rijn H (2014) Trial-by-trial fluctuations in cnv amplitude reflect anticipatory adjustment of response caution. *NeuroImage* 69: 95–105.

Bradshaw JL (1970) Pupil size and drag state in a reaction time task. *Psychonomic Science* 18(2): 112–113.

Brisson J, Mainville M, Mailloux D, Beaulieu C, Serres J and Sirois S (2013) Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods* 45(4): 1322–1331.

Cooper RM (1974) The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6(1): 84–107.

Engelhardt PE, Ferreira F and Patsenko EG (2010) Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology* 63(4): 639–645.

Gagl B, Hawelka S and Hutzler F (2011) Systematic influence of gaze position on pupil size measurement: analysis and correction. *Behavior Research Methods* 43(4): 1171–1181.

Goldwater BC (1972) Psychological significance of pupillary movements. *Psychological bulletin* 77(5): 340–355.

Hastie T and Tibshirani R (1990) *Generalized Additive Models*. John Wiley & Sons, Inc.

Hayes TR and Petrov AA (2016) Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods* 48(2): 510–527.

Hendrix P, Bolger P and Baayen RH (2016) Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology LMC* 43(1): 128–149.

Hess EH and Polt JM (1960) Pupil size as related to interest value of visual stimuli. *Science* 132(3423): 349–350.

Hess EH and Polt JM (1964) Pupil size in relation to mental activity during simple problem-solving. *Science* 143(3611): 1190–1192.

Hoeks B and Levelt WJM (1993) Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods* 25(1): 16–26.

Hyönä J, Tommola J and Alaja AM (1995) Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology* 48(3): 598–612.

Jackson I and Sirois S (2009) Infant cognition: going full factorial with pupil dilation. *Developmental Science* 12(4): 670–679.

Jainta, Stephanie, Vernet, Marine, Yang, Qing and Kapoula, Zoi (2011) The pupil reflects motor preparation for saccades - even before the eye starts to move. *Frontiers in Human Neuroscience* 5(97). DOI:10.3389/fnhum.2011.00097.

Janisse MP (1977) *Pupillometry: The psychology of the pupillary response.* Series in clinical and community psychology. Halsted Press.

Just MA and Carpenter PA (1993) The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology* 47(2): 310–339.

Kösling K, Kunter G, Baayen RH and Plag I (2013) Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech* 56(4): 529 – 554.

Kuchinke L, Võ MLH, Hoffman M and Jakobs AM (2007) Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology* 65: 132–140.

Kuchinsky SE, Ahlstrom JB, Vaden Jr KI, Cute SL, Humes LE, Dubno JR and Eckert MA (2013) Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology* 50(1): 23–34.

Lõo K, van Rij J, Järvikivi J and Baayen RH (2016) Individual differences in pupil dilation during naming task. In: Papafragou A, Grodner D, Mirman D and Trueswell J (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, pp. 550–555.

Laeng B, Sirois S and Gredebäck G (2012) Pupillometry: A window to the preconscious? *Perspectives on Psychological Science* 7(1): 18–27. DOI:DOI:10.1177/1745691611427305.

Lin X and Zhang D (1999) Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61: 381–400.

Mathôt S (2013) A simple way to reconstruct pupil size during eye blinks. Retrieved from https://doi.org/10.6084/m9.figshare.688001.

Mathôt, Sebastiaan, van der Linden, L, Grainger, J and Vitu, F (2015) The pupillary light response reflects eye-movement preparation. *Journal of Experimental Psychology: Human Perception and Performance* 41(1): 28–35. DOI:10.1037/a0038653.

Milin P, Feldman LB, Ramscar M, Hendrix P and Baayen RH (2017) Discrimination in lexical decision. *PLoS ONE* 12(2): e0171935.

Mirman D, Dixon JA and Magnuson JS (2008) Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59(4): 475–494.

Nixon JS, van Rij J, Li X and Chen Y (2015) Cross-category phonological effects on erp amplitude demonstrate context-specific processing during reading aloud. In: Botinis A (ed.) *ExLing 2015: Proceedings of the International Conference of Experimental Linguistics*. pp. 50–53.

Nixon JS, van Rij J, Mok P, Baayen RH and Chen Y (2016) The temporal dynamics of perceptual uncertainty: eye movement evidence from cantonese segment and tone perception. *Journal of Memory and Language* 90: 103–125.

Pinheiro JC and Bates DM (2000) *Mixed-effects models in S and S-PLUS*. New York: Springer.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ramsay JO and Silverman BW (2002) *Applied functional data analysis: methods and case studies*. Springer series in statistics. New York: Springer.

Ramsay JO and Silverman BW (2005) *Functional Data Analysis*. Springer series in statistics. Springer.

Scheepers C and Crocker MW (2004) Constituent order priming from reading to listening: A visual-world study. In: Carreiras M and Clifton C (eds.) *The On-line Study of Sentence Comprehension: Eyetracking, ERPs, and Beyond*. Psychology Press.

Schluroff M (1982) Pupil responses to grammatical complexity of sentences. *Brain and Language* 17: 133–145.

Schluroff M, Zimmermann TE, Freeman Jr R, Hofmeister K, Lorscheid T and Weber A (1986) Pupillary responses to syntactic ambiguity of sentences. *Brain and Language* 27(2): 322–344.

SR Research Ltd (2005–2010) *EyeLink®1000 User Manual*, 1.5.2 edition.

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM and Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217): 1632–1634.

Tomaschek F, Arnold D, Bröker F and Baayen RH (to appeara) Lexical frequency co-determines the speed-curvature relation in articulation. *Journal of Phonetics* .

Tomaschek F, Tucker B and Baayen RH (to appearb) Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistic Vanguard* .

Tremblay A and Newman AJ (2015) Modeling nonlinear relationships in erp data using mixed-effects regression with r examples. *Psychophysiology* 52: 124–139. DOI:10.1111/psyp.12299.

van Rij J (2012) *Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults.* Phd thesis, University of Groningen.

van Rij J, Hollebrandse B and Hendriks P (2016) Children's eye gaze reveals their use of discourse context in object pronoun resolution. In: Holler A and Suckow K (eds.) *Empirical perspectives on anaphora resolution*. Berlin, Walter de Gruyter.

van Rij J, Vaci N, Wurm LH and Feldman LB (to appear) Alternative quantitative methods in psycholinguistics: Implications for theory and design. In: Pirrelli V, Plag I and Dressler WU (eds.) *Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon*, chapter 3. Berlin: Mouton de Gruyter.

van Rij J, Wieling M, Baayen RH and van Rijn H (2017) itsadug: Interpreting time series and autocorrelated data using gamms. Published on the Comprehensive R Archive Network (CRAN). URL `https://cran.r-project.org/web/packages/itsadug`.

van Rijn H, Dalenberg JR, Borst JP and Sprenger SA (2012) Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS ONE* 7(12): e51134. DOI:https://doi.org/10.1371/journal.pone.0051134.

Verney SP, Granholm E and Marshall SP (2004) Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology* 52: 23–36.

Vogelzang M, Hendriks P and van Rijn H (2016) Pupillary responses reflect ambiguity resolution in pronoun processing. *Language, Cognition and Neuroscience* 31(7): 876–885.

Watson AB and Yellott JI (2012) A unified formula for light-adapted pupil size. *Journal of Vision* 12(10): 1–16.

Wieling M (submitted) Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between l1 and l2 speakers of english. *Journal of Phonetics* .

Wierda S, van Rijn H, Taatgen NA and Martens S (2012) Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *PNAS* 109(22): 8456–8460.

Winn B, Whitaker D, Elliot DB and Phillips NJ (1994) Factors affecting light-adapted pupil size in normal human subjects. *Investigative Ophthalmology & Visual Science* 35(3): 1132–1137.

Winn MB, Edwards JR and Litovsky RY (2015) The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and hearing* 36(4): 153–165.

Wood SN (2006) *Generalized Additive Models: An Introduction with R*. 1st edition edition. Chapman and Hall/CRC.

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1): 3–36.

Wood SN (2017a) *Generalized Additive Models: An Introduction with R*. 2nd edition edition. Chapman and Hall/CRC.

Wood SN (2017b) mgcv: Mixed gam computation vehicle with automatic smoothness estimation. Published on the Comprehensive R Archive Network (CRAN). URL `https://cran.r-project.org/web/packages/mgcv`.

Wood SN, Goude Y and Shaw S (2014) Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Wood SN, Pya N and Saefken B (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111: 1548–1575. DOI: http://dx.doi.org/10.1080/01621459.2016.1180986.

Zekveld AA, Kramer SE and Festen JM (2011) Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear & Hearing* 32(4): 498–510.

Zellin M, Pannekamp A, Toepel U and van der Meer E (2011) In the eye of the listener: Pupil dilation elucidates discourse processing. *International Journal of Psychophysiology* 81(3): 133–141.