SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2017, Vol. 26(6) 2708–2725
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280215607411
journals.sagepub.com/home/smm
SAGE

# Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm

**Jacob J Oleson,[1] Joseph E Cavanaugh,[1] Bob McMurray[2] and Grant Brown[1]**

**Abstract**

In multiple fields of study, time series measured at high frequencies are used to estimate population curves that describe the temporal evolution of some characteristic of interest. These curves are typically nonlinear, and the deviations of each series from the corresponding curve are highly autocorrelated. In this scenario, we propose a procedure to compare the response curves for different groups at specific points in time. The method involves fitting the curves, performing potentially hundreds of serially correlated tests, and appropriately adjusting the overall alpha level of the tests. Our motivating application comes from psycholinguistics and the visual world paradigm. We describe how the proposed technique can be adapted to compare fixation curves within subjects as well as between groups. Our results lead to conclusions beyond the scope of previous analyses.

**Keywords**

Autocorrelation, bootstrap, multiple comparisons, time series

## 1 Introduction

A fundamental problem in the cognitive sciences involves adequately accounting for the temporal dynamics of behavioral, cognitive, and neural processes. Basic human processes such as perception, cognition, and language unfold over the span of milliseconds and seconds, and increasingly sophisticated behavioral and neural recording techniques have been developed to measure these ongoing processes.

Language provides a critical example of this phenomenon. Spoken language unfolds over time, and the cognitive processes that underlie it have their own distinct temporal dynamics. Therefore, accurate characterization of these dynamics, and the use of such characterization to distinguish theories and delineate different populations of language users is of great scientific importance.

Consider the simple task of recognizing a single spoken word like *beaker*. During the first few hundred milliseconds, the listener may have heard only "*bea*"; consequently, they may partially consider a range of potential competitors like *beetle*, *beak*, *beach*, and so forth. As more of the input unfolds, these words compete until only the word that is recognized remains. This dynamic competition process is now well understood and can be characterized by computational models that make explicit claims about the involved mechanisms (e.g. McClellend and Elman,[1] Shortlist[2]). In such models, the competition dynamics are written in terms of differential equations, which give rise to nonlinear functions to describe the process.

Now consider the challenge of distinguishing between two subpopulations of people that only differ in their temporal dynamics over a small fraction of time relative to the entire time course (e.g. typically developing children and children with language impairment: McMurray et al.[3]). Alternatively, consider the evaluation of two experimental conditions that lead to a transient difference in the time course of a behavioral measure such that

[1]Department of Biostatistics, The University of Iowa, Iowa City, Iowa, USA
[2]Department of Psychology, The University of Iowa, Iowa City, Iowa, USA

**Corresponding author:**
Jacob J Oleson, University of Iowa, 145N Riverside Dr, College of Public Health, Iowa City, IA 52242-2007, USA.
Email: jacob-oleson@uiowa.edu

two types of words (e.g. low- and high-frequency words: Dahan et al.[4]) may show different degrees of consideration, but only within a prescribed time window. The ability to isolate and detect such difference is crucial for disentangling theoretical accounts of language. The objective of the present manuscript is to develop a statistical method to determine when we can reasonably declare that these trajectories are significantly different from one another and to demonstrate the applicability of this method to simulated data.

The problem of detecting temporal differences appears in many fields and is of crucial importance for cognitive neuroscientific methods such as fMRI, EEG, and electrocorticography. The problem is particularly prevalent in the psycholinguistics literature which has emphasized the issue of temporal dynamics for many years. We develop and showcase our statistical technique in the context of a particular behavioral method, the visual world paradigm (VWP), which combines eye tracking with a simple interactive environment to measure the ongoing dynamics of language processing.[5] We build this method upon recent work using nonlinear curves to describe such data.[3,6] To briefly summarize the need for such an approach, while standard statistical techniques are well equipped to detect an overall difference between time series, it is much harder to identify the specific time periods during which sufficient evidence exists to declare exactly where the difference occurs while maintaining statistical rigor.

We next introduce some specifics of the VWP and the type of data that are usually generated in this setting. We then briefly describe existing techniques used to analyze such data and present our own approach along with the motivation for it. The technique is validated with an analysis of data from a previous study, and also with a series of Monte Carlo simulated analyses used to evaluate the likelihood of falsely detecting an effect where there is none, and of correctly detecting an effect that is legitimately present. Although we emphasize psycholinguistics applications, the method is more broadly applicable, and we return to possible extensions in the general discussion.

## 1.1 The VWP

The VWP is designed to assess the real-time unfolding of language processing while subjects are engaged in everyday tasks. It is premised on the notion that in order to make an overt response to an object (a visual world), subjects typically must fixate on their intended response prior to initiating an action such as reaching for it, pushing a button, or clicking it with a mouse. Since eye movements can be generated quickly, they can therefore reveal the decision dynamics leading up to this final decision.

In typical instantiations of the VWP, a participant is presented with a spoken instruction to select one of several visual objects either on a computer screen or in the real world (see Salverda et al.[7] for a review). The objects are selected by the experimenter to represent possible competing interpretations of the auditory signal (often interpretations that will be briefly considered before being ruled out). Crucially, in order to complete the tasks, participants typically must fixate one or more objects (i.e. they need to know where the target object is before they can click on it with their mouse). However, fixations can be initiated much faster than a reach, and often begin 200 ms after the beginning of the word, with multiple fixations leading up to the overt response. Consequently, participants' fixations to each object at any given point in time—fixations which are typically initiated far earlier than the overt response—can indicate something about how strongly they are considering that object as a possible response.

Analyses of VWP data usually proceed, at least in part, on the basis of time series which reflect the probability of fixating on a target across a series of small time slices (e.g. milliseconds). On each trial, objects may be chosen to exhibit a number of linguistically interesting relationships to the designated target. Fixations arrive as a series of discrete "looks" to multiple objects on each trial. By averaging across many trials, but within small time slices, researchers can compute an estimate of the probability of fixating on a given object and how it gradually changes over time. While such data are popularly termed "fixation proportions" (or some variant), they are really the proportion of trials in which the participant is fixating on an object at a specific time.

For example, in classic versions of this paradigm used to study the recognition of isolated single words (e.g. McMurray et al.,[3] Allopenna et al.[8]), each participant might hear a word like *lizard* while viewing a screen that contains a picture of a *lizard* (the correct response, designated a target), *liver* (a cohort, which overlaps with the target word at onset), a *wizard* (a rhyme, which overlaps at offset), and a *necklace* (which is unrelated). On each trial, participants might make a series of discrete or ballistic saccades to one or more objects on the screen. Averaged across trials, this yields the typical pattern of data as shown in Figure 1(a); within about 200 ms, participants fixate on both the target and cohort more than the unrelated objects. This delay corresponds to the amount of time it takes to plan and launch an eye movement; consequently, these early fixations are driven by only the earliest portion of the stimulus ("*li-*") and are directed to both objects. Several hundred milliseconds later, more of the word is heard, and the target starts to diverge from the competitor. However, as the complete word

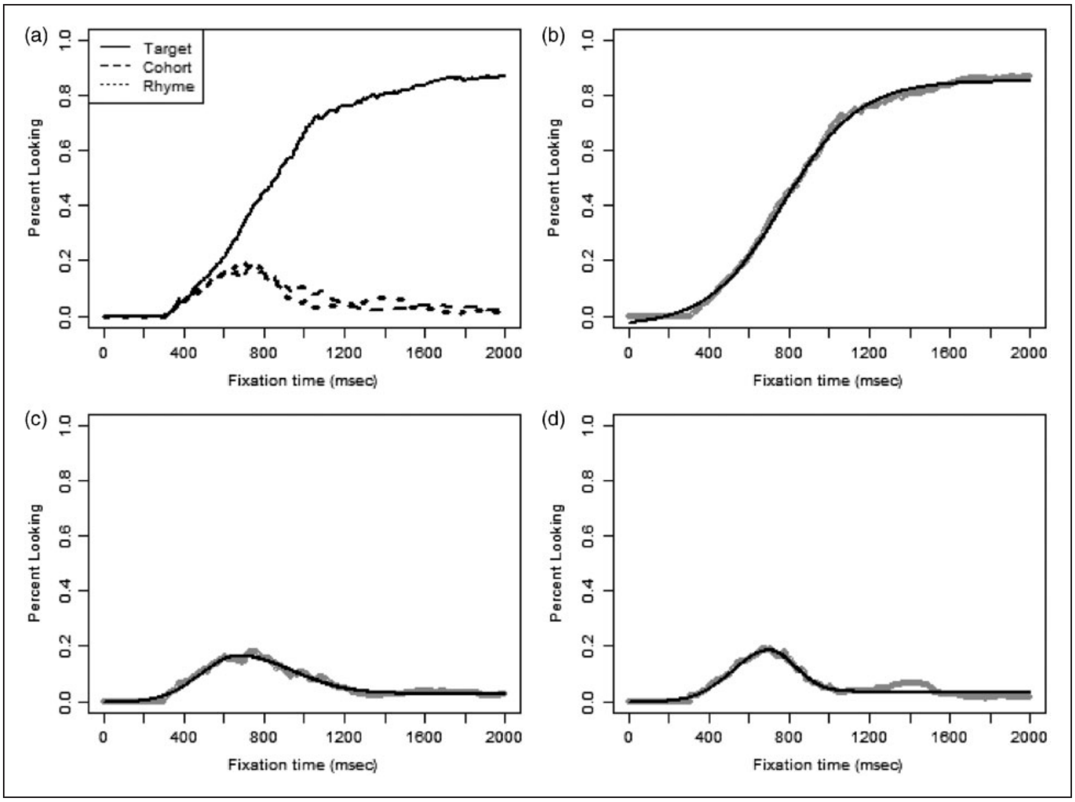

**Figure 1.** This figure contains the fixation proportions across the time span for one randomly selected individual in the cochlear implant dataset. (a, raw data) shows the observed proportions for the (b) target, (c) cohort, and (d) rhyme figures. Figures (b) to (d) display the observed curve in gray and the fitted curve in black. The observed curves in (b) to (d) are the same curves shown in (a).

unfolds, emerging overlap with the rhyme object may lead to some partial fixations. By the end of the time course, typically only one object (the target or referent) is under active consideration.

Within the VWP, multiple experimental designs are possible. Many studies compare the time course of fixations to the target under various conditions; for example, comparing target fixations as a function of the number of competitors[9] or as a function of a word's frequency of occurrence[4] (these are within-subject, between-word comparisons). Studies also compare the fixations to two different types of competitors within the same trial.[4,10] Further, a number of studies compare fixations to a competitor object across different trials, for example after different manipulations of the auditory stimulus (within-subject, within-item[11]), or across entirely different stimuli (within-subject, between-item[12]). Finally, a growing number of studies are using the VWP in correlational designs to evaluate individual differences,[3,13–15] comparing different aspects of the time course of fixations across individuals.

Across all of these designs, a variety of statistical approaches have been applied. The form of the resulting data has a number of important features—many of which are glossed over by existing approaches. First, the data researchers typically analyze are in the form of densely sampled time series. Underlying this time series is an autocorrelation structure. People cannot move their eyes every 4 ms (a typical sampling window for many eye trackers)—fixations usually last 200–300 ms. The smooth gradual changes in the function derive from averaging many series of ballistic saccades, implying that adjacent time windows are almost always, in part, the product of the same physiological events. Second, as is typical in most psycholinguistic paradigms, these functions are a product of sampling from multiple random factors—typically subject and item (words), but often other factors (like the talker). Finally, the data are fundamentally probabilistic (and a product of averaging discrete trials). However, the data are rarely distributed in a binomial fashion, but are instead multinomial; there are typically more than two options on the screen. As we describe in Section 2, most existing analytic approaches for VWP data ignore one or more of these factors in the interest of tractability. Our proposed approach was not developed to address every intricacy of the data; however, an understanding of these features is required to evaluate existing approaches and develop improvements.

## 1.2 Motivating study: Word recognition by cochlear implant (CI) users

We develop and test our approach in the context of a recent study examining individual differences in the time course of word recognition between normal hearing (NH) listeners and hearing impaired listeners who use a CI. CIs are designed to provide access to sound for individuals with severe to profound deafness (see Niparko,[16] for a review). The device receives acoustic signals through an externally worn microphone. These signals are processed to filter and transmit those components of sound critically important for speech perception. The components are transmitted via electrical signals to an array of electrodes implanted directly in the cochlea, which directly stimulates the auditory nerve. Due to limitations of signal processing, implant technology, and biophysical factors (the fact that electrical current "spreads" in the cochlea), the CI does not produce an exact replica of NH. Systematic loss of low-frequency information, and fine-grained frequency differences, among other factors, routinely impact the hearing of CI users. However, with current CI technology and surgical techniques, the majority of postlingually deafened CI recipients score above 80% on high-context sentences in quiet listening conditions, even without visual cues.[17]

In this study, we reanalyze a portion of a previous study[13] which examined the time course of spoken word recognition among adult, postlingually deafened CI users and NH listeners (experiment 1). In that study, listeners heard a word from one of 29 different sets of four common words. Each set contained a target word (*dollar*), a cohort that overlapped at onset (*dolphin*), a rhyme that overlapped at offset (*collar*), and an unrelated baseline word (*hamster*). The goal was to examine real-time spoken word recognition and lexical access under degraded input, by analyzing the time course of spoken word recognition in each group. The authors found a number of differences between groups, such as in the speed at which target fixations grew, and the subsequent loss of fixations to competitors. We demonstrate the utility of our technique on that study and determine at what point in time these specific differences occur.

# 2 Statistical methods for VWP data

The earliest and still most widely used analysis technique for VWP data is the area under the curve (AUC) approach. Here, the dynamics of the time series are typically discarded, and the experimenter computes the average proportion of fixations within some fixed time window for each condition. AUC is quite useful for detecting the existence of an effect, but it has few options for detecting when an effect occurs. Further, it is hampered by the need to specify an arbitrary time window over which the area is computed.

A second, less common, option is to identify specific components of the fixation record for analysis, rather than starting with the time course functions. Examples include the duration of a particular fixation,[18] the latency to look to an object,[19] or the raw number of fixations.[20,21] However, as with AUC, such techniques emphasize whether there was an effect at various times, and not a fine-grained detection of when effects can be seen.

An increasingly popular approach to visual world data is to fit some nonlinear function of time to visualizations of the data. The parameters of the individual specific functions can then be used as descriptors of how the trajectories change over time. The resulting parameters may be analyzed using traditional statistical models (e.g. ANOVAs), or the functions can be integrated into mixed effects models. The first such approach proposed[22] the use of orthogonal polynomials as the basis for the analyses. This approach is advantageous because it is implemented using linear methods. Moreover, orthogonal polynomials can easily be integrated into mixed models, which can capture both subject- and item-specific effects, and which can accommodate variation within individuals in determining the significance of effects. However, polynomial functions do not offer a precise characterization of the typical shape of VWP data. For example, the looks to the target shown in Figure 1(a) would require approximately 6–7 polynomial terms for adequate description, and the parameters themselves do not describe any meaningful aspect of the data. Thus, these functions often provide evidence that there was a temporal effect but do not generally facilitate more precise description. The goal of this project is to develop a statistical tool for detecting not only whether two time series (as in the VWP) differ, but the specific time points at which they differ.

More recent approaches[3,6] have used nonlinear functions designed to capture more meaningful aspects of variation in these time series. For example, target fixations can be fit by a four-parameter logistic function that captures the asymptotes, slope, and crossover point. The function can be represented by the following form

$$p_{it} = A_i + \frac{B_i - A_i}{1 + \exp\left(4{*}D_i \frac{(C_i - t)}{(B_i - A_i)}\right)} \tag{1}$$

In this form, $p_{it}$ denotes the proportion of times out of $M$ trials that individual $i$ is looking at the target curve at time $t$. Then, $A_i$ is the baseline level of the curve, $B_i$ is the maximum proportion reached at the end of the measurement period, $C_i$ is the crossover point of the logistic curve, and $D_i$ is the slope obtained at the crossover point. These parameters represent meaningful descriptors of the data, and the functions tend to approximate the empirical curves well. Note that we do not explicitly restrict $p_{it}$ to be bound between 0 and 1. In some cases, the overall model fit is better if the minimum, $A_i$, is allowed to be slightly negative.

Fixations to the competitor curves take a different form than does the logistic target curve, rising from baseline to a peak and then falling again to a separate baseline level. The functional form we use here has been employed successfully,[3,6] where it is termed the asymmetric Gaussian curve. The function essentially merges a Gaussian curve on the left with a Gaussian curve on the right that shares the same peak but have separate variances and baselines

$$p_{it} = \begin{cases} B_i - A_{1i}) * \exp\left(\frac{(t-\mu_i)^2}{-2\sigma_{1i}^2}\right) + A_{1i} & \text{if } t \leq \mu_i \\ (B_i - A_{2i}) * \exp\left(\frac{(t-\mu_i)^2}{-2\sigma_{2i}^2}\right) + A_{2i} & \text{if } t > \mu \end{cases} \tag{2}$$

Again, $p_{it}$ denotes the probability of looking at that object for individual $i$ at time $t$. Then, $A_{1i}$ is the asymptotic curve height at the beginning of the time span, $A_{2i}$ is the asymptotic curve height at the end of the time span, $B_i$ is the height of the curve at its peak, $\mu_i$ is the point in time when the peak is reached, $\sigma_{1i}^2$ is the variance that determines the incoming slope of the curve, and $\sigma_{2i}^2$ is the variance that determines the outgoing slope of the curve. Again, we do not explicitly restrict $p_{it}$ to be bound between 0 and 1. These two functions have been successfully employed in a number of prior studies with this sort of data.[3,6,13]

VWP researchers often estimate the parameters of these functions for individual subjects, and then analyze the resulting parameter estimates with more traditional, general linear modeling approaches, rather than fitting a larger, unified, nonlinear mixed model. Among other drawbacks, this approach ignores the variability in the parameter estimates. Integrating all of the individual curves together into one model might better incorporate both within-subject variability and other random effects. However, a full nonlinear model is difficult to fit using standard software given the size of the data, the use of nonlinear functions, and multiple random effects. Related Bayesian models for growth curves have been successfully implemented in other fields (e.g. Oleson et al.[23]) and represent an important avenue of recent development. In this work, we instead focus on the individual fits of such functions to each participant.

An alternate class of techniques uses nonparametric functions such as splines or functional mixed effects models.[24,25] Although nonparametric forms have been successfully implemented in other settings, the parameterization used in this paper is well suited to comparison between curves. Importantly, while we focus on the two functions previously outlined, the technique proposed here can be applied to any parameterization of the time course function (including polynomial growth curves).

Both the polynomial and nonlinear approaches offer a reasonable way to characterize the time course of fixations to individual objects, and comparing the parameters of the function across conditions allows researchers to make reasonable inferences about how they differ with respect to curve properties like peak height and growth rate. In most cases, however, these approaches do not permit direct inferences about when conditions differ and can sometimes lead to ambiguous results.

## 3 Development of test statistic

Here we develop a testing method to detect specific points in time at which two time series significantly differ. Our approach handles this task in three steps. First, we fit nonlinear curves to each participant's data. Second, for between-group comparisons, we bootstrap those fitted functions to estimate the mean of the group curves and the variance associated with these estimators. For within-subject comparisons, the bootstrapping is accomplished to assess the variation of the subject-specific curves. Third, we use these estimated curves along with the associated bootstrap measures of accuracy to make statistical comparisons at each point in time. Finally, we apply a novel family-wise error correction that is sensitive to the autocorrelation in the data to ensure appropriately conservative inference.

### 3.1 Motivation

This procedure aims to solve a particular type of multiple comparison problem where researchers can simply compare two groups at each point along the time series and obtain reasonable estimates of the time at which the curves deviate. Our method differs from those previously used in the literature in a number of important ways. First, rather than relying upon the raw data, we use fitted curves, to smooth the data in a way that is consistent with the underlying theoretical behavior of these processes. This smoothing is important as individual subjects' curves can be fairly noisy despite inherent reality, which suggests gradual change. In order to make inferences about the underlying processes, we therefore use curves fitted from a selection of reasonable and interpretable functional forms (see Section 2). This nonlinear approach leaves to the modeler the selection of a functional basis. However, as our inferential approach is based on differences between fitted curves rather than specific parameters, determination of statistically significant regions is not greatly impacted by this choice.

Second, the complexities of eye-movement data, along with the challenges involved in fitting subject- and group-specific curves, leads to difficulties in the determination of accuracy assessments for the estimated curves. For instance, eye-movement data is measured on a proportional scale, and for many of the relevant measures (particularly looks to the competitors) the absolute proportions can be near zero. Consequently, assumptions of normality do not always hold. The use of bootstrapping to assess the variability of estimated curves offers a way to overcome such problems. Moreover, the proportion of fixations always has some degree of within-subject variance. Typical approaches discard within-subject variance, focusing only on the average proportion, or the fitted curve, for each subject. However, if a bootstrap procedure is developed on both the estimated parameters and their corresponding standard errors for the subject-specific curves, then at the group level, the procedure can incorporate both within-subject and between-subject variation into the accuracy assessments for the estimated curves.

Third, the combination of fitted curves and bootstrapping addresses another important problem that cannot be solved with simple means alone. With a nonlinear function such as the aforementioned logistic curve, inference on the grand mean (across subjects) constitutes a population-averaged analytic approach. Researchers in this setting prefer a subject-specific analytic approach. The fitting of curves to individual subjects thus represents a truer depiction of the underlying generating mechanism.

Finally, the preceding innovations are independent of the need to correct for hundreds of autocorrelated multiple comparisons. We will outline some reasons why a family-wise error rate (FWER) approach may be advantageous in the present setting. The central insight is that these comparisons are not independent and should be very highly correlated. We will incorporate an estimate of this correlation into the family-wise error estimate to achieve a more powerful correction than a strict Bonferroni adjustment.

## 3.2 The nonlinear model with autocorrelated errors

To better understand all of the model components, we first examine fixation curves for individuals from the study discussed in Section 1.2. Specifically, consider the curves for the fixations to the target, shown in Figure 1(a). Assume the following model for $i = 1,\ldots, n$ subjects and $t = 1,\ldots, N$ time points

$$p_{it} = f(\theta_{it}) + \varepsilon_{it} \tag{3}$$

where $f(\theta_{it})$ is defined as some linear or nonlinear function of interest, and the $\varepsilon_{it}$ are normally distributed error terms with zero mean and constant variance (i.e. Gaussian white noise). Conveniently, the error terms, $\varepsilon_{it}$, are assumed independent overall $i$ and $t$, although we intend to relax this assumption to allow for autocorrelated errors across $t$ for each subject $i$. For the target curves, $f(\theta_{it})$ is the four-parameter logistic function, given in equation (1). For the competitor curves, this function is the asymmetric Gaussian curve given in equation (2). Each individual's curve is fit separately using the gnls (generalized nonlinear least squares) function from the nlme R-package.[26,27] As shown in Figure 1(b) to (d) for one sample individual, the model estimated curves (in black) appear to fit the raw proportions (gray) well.

It is important to note that because the modeling takes place on the average across many trials rather than on data representing the true sequence of eye movements within a given trial, we expect considerable temporal correlation in the model errors. When the residuals, $e_{it}$, are examined from each fitted model, the lag one autocorrelation is extremely high, ranging from 0.9 to 0.99 per individual. Since the residuals substantially deviate from serial independence, a critical assumption in the traditional least squares model, we should incorporate a time series correlation structure for the model errors.

A variety of methods could be used to account for the correlation in the residuals. Depending on the nature of the sample autocorrelation function (ACF) and the partial autocorrelation function (PACF), an autoregressive moving-average model could be fit to the residuals. For our analysis discussed in Section 5, a check of the ACF and the PACF suggests that an autoregressive model of order one (AR(1)) will adequately account for the temporal correlation of the model errors. The AR(1) model provides a parsimonious and intuitively appealing mechanism for describing temporal dependence. Moreover, an AR(1) process based on a large, positive autoregressive parameter can be used to adequately characterize gradually changing time series. (Of note, the autoregressive parameter corresponds to the lag-one correlation.) Since the underlying process governing eye-tracking data suggests smooth transitions, in our setting, assuming an AR(1) model for the error process is defensible based on both the dynamics of the phenomenon and empirical evidence. We will therefore continue our development assuming an AR(1) structure for the errors, but note that other error structures may be entertained and are easily implemented. The chosen error structure does not affect the development of the test statistic but can impact the fit of the individual curves.

The nonlinear model with autocorrelated errors is identical to the model provided in equation (3), except that the error terms follow the relation

$$e_{it} = \phi e_{i,t-1} + w_{it} \tag{4}$$

where the $w_{it}$ denote independent Gaussian white noise processes over $t$ for each subject $i$.

The effect of incorporating autocorrelation into the estimation process can be seen by comparing fits assuming independent errors with those assuming autocorrelation on four randomly selected individuals. The parameter estimates from these fitted models are each listed in Table 1. Note that the parameter estimates are relatively unchanged by accounting for correlated errors, but the standard errors for each estimator are dramatically influenced; any tests using standard errors which ignore this autocorrelation will be inaccurate (and too conservative). The need to accommodate residual autocorrelation is also evidenced by the substantial differences in the values of the Akaike information criterion (AIC) for each subject-specific fitted model (a difference of two or more in magnitude is generally considered meaningful.) The residual autocorrelation will be different for each subject but tends to be between 0.9 and 0.99.

For the logistic function, correlation appears to have the greatest impact on the crossover point, as the other three parameter estimates are relatively unchanged. Moreover, as the correlation increases, the more the crossover appears to be impacted. It is important to note that this impact is not a systematic bias, since subjects 1 and 4 (the two with the highest crossovers) show estimates that move in different directions using the two methods.

We will further assess the impact of the AR(1) structure in Section 4 of the paper. It is possible that developing a model that directly incorporates the eye-movement dynamics would preclude the need for correlated errors, but such a model would be considerably more difficult to specify. Moreover, as we will demonstrate, our model reasonably simulates this autocorrelation while remaining relatively parsimonious.

## 3.3 Bootstrapping

As previously stated, fitting a large multilevel mixed model that accounts for a population curve and subject specific-curves within a unified structure is a very challenging problem in this setting. Our goal lies in discovering at

**Table 1.** Parameter estimates for four subjects.

| Parameter | Subject 1 | | Subject 2 | | Subject 3 | | Subject 4 | |
|---|---|---|---|---|---|---|---|---|
| | Ind | AR(1) | Ind | AR(1) | Ind | AR(1) | Ind | AR(1) |
| Minimum | −0.0216<br>(0.0022) | −0.0181<br>(0.0195) | −0.0024<br>(0.0008) | −0.0030<br>(0.0045) | −0.0207<br>(0.0014) | −0.0180<br>(0.0070) | −0.0430<br>(0.0028) | −0.0361<br>(0.0177) |
| Maximum | 0.8677<br>(0.0021) | 0.8619<br>(0.0188) | 0.8806<br>(0.0013) | 0.8855<br>(0.0063) | 0.9393<br>(0.0018) | 0.9382<br>(0.0084) | 0.8946<br>(0.0013) | 0.8991<br>(0.0104) |
| Slope | 0.0017<br>(0.00002) | 0.0017<br>(0.00010) | 0.0022<br>(0.00001) | 0.0022<br>(0.00006) | 0.0018<br>(0.00001) | 0.0018<br>(0.00005) | 0.0014<br>(0.00001) | 0.0014<br>(0.00007) |
| Crossover | 769.44<br>(1.67) | 752.41<br>(12.33) | 801.33<br>(0.67) | 802.327<br>(3.47) | 688.04<br>(0.97) | 688.67<br>(4.97) | 660.89<br>(1.80) | 665.41<br>(12.79) |
| Correlation | 0 | 0.989 | 0 | 0.935 | 0 | 0.941 | 0 | 0.976 |
| AIC | −2023.8 | −3401.0 | −2220.1 | −2874.0 | −2020.4 | −2707.3 | −2671.5 | −4178.7 |

which time points that the time series curves of the two groups significantly deviate from each other. In the previous section, we demonstrated how to fit each individual's curve through a nonlinear model with autocorrelated errors. We now showcase how to arrive at a population level curve so that two groups can be compared.

Inference can be made about population level curves through parametric bootstrapping of individual-specific curve fits. Specifically, we implement the following algorithm, which provides an estimate of the population level curve and an associated standard error at every time point.

(1) For each subject, fit the nonlinear function, specifying an AR(1) autocorrelation structure for the model errors. Obtain the estimates and standard errors for the model parameters resulting from each fit. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a normal distribution with a mean corresponding to the point estimate and a standard deviation corresponding to the standard error.
(2) Using the approximate sampling distributions in (1), randomly draw one bootstrap estimate for each of the model parameters on every subject.
(3) Once a bootstrap estimate has been collected for each parameter and for every subject, for each parameter, find the mean of the bootstrap estimates across individuals.
(4) Use the mean parameter estimates to determine the predicted population level curve, which provides the average population response (e.g. proportion of looks) at each time point.
(5) Perform steps (2) through (4) a large number of times (e.g. 1000), thereby obtaining a bootstrap collection of hypothetical population curves.
(6) At every time point, the bootstrap sample of population curves in (5) can be used to produce an average response and a standard deviation.

Performing the bootstrapping on the individual-specific parameter estimates in this manner effectively accounts for subject-specific variation that is often ignored in analyses.

## 3.4 Comparing curves

We perform the preceding parametric bootstrapping algorithm separately for the two groups that we wish to compare. At each time point, a test statistic for a two-sample t-test is then constructed using the estimated functions and their standard errors. The null hypothesis is that the proportion of looks for group 1 equals the proportion of looks for group 2 at time $t$. The test statistic $T_t$ can be written as

$$T_t = \frac{(\bar{p}_{1t} - \bar{p}_{2t})}{\sqrt{s_{1t}^2 + s_{2t}^2}} \tag{5}$$

where $\bar{p}_{1t}$ is the average of all the bootstrap population curves at time $t$ for group 1, with $s_{1_t}^2$ denoting the corresponding variance of the bootstrapped estimates. Similarly, $\bar{p}_{2t}$ and $s_{2_t}^2$ are the mean and variance for group 2. The result leads to a t-test at each measurement point in time.

Another important test to conduct in practice is a within-subject comparison. For the VWP, much of the understanding of language processing comes from examining competitor curves. A within-subject comparison allows for a comparison of the cohort word to the rhyme word or even to the unrelated word, with individuals serving as their own controls. Since each study participant views cohort, rhyme, and unrelated words as part of the testing procedure, we will be able to determine at what points in time, the proportion of looks between these competitors differs.

The aforementioned comparison can be performed through a paired t-test approach. The bootstrapping method is very similar to the approach previously outlined. The main steps are as follows.

(1) For each of the two fixation curves, fit the nonlinear function, and obtain the resulting estimates and standard errors. For each pair of parameter estimates from the two fitted curves, approximate the sampling distribution by a bivariate normal distribution, with means corresponding to the point estimates, standard deviations corresponding to the standard errors, and the correlation set at the empirical estimate obtained across the sample of subjects.
(2) Using the approximate sampling distributions in (1), randomly draw pairs of bootstrap estimates.

(3) Once pairs of bootstrap estimates have been collected for each parameter, find the two corresponding fixation curves determined by each set of estimates.
(4) Perform steps (2) through (4) a large number of times (e.g. 1000), thereby obtaining a bootstrap collection of matched hypothetical fixation curves.

At each time point, we then compute a bootstrapped subject-specific difference score based on the two fixation curves (e.g. by taking the percentage viewing the cohort minus the percentage viewing the rhyme word). Denoting the average of the difference scores as $\bar{p}_{Dt}$ and the standard deviation of the differences as $s_{Dt}$, the test statistic for the paired t-test can be written as

$$T_t = \frac{\bar{p}_{Dt}}{\sqrt{s_{Dt}^2}} \tag{6}$$

## 3.5 Multiple comparisons

In the present setting, using the approach previously outlined involves performing over 500 two-sample t-tests, assuming a standard analysis window of 0–2000 ms, and 4 ms sampling frame. With a larger analysis window, or a more fine-grained sampling at a higher frequency, the number of tests would be even more extreme. As one would expect given the high correlation between observations close in time, the tests are also very highly autocorrelated. Interpretation of so many simultaneous hypothesis tests requires at least some consideration of multiple comparisons procedures. Investigators must, for each analysis, determine whether strict control of the FWER or control of the false discovery rate (FDR)[28] is appropriate and at what level adjustments should be performed. A number of methods have been proposed to adjust for multiple comparisons. Some of the most popular are the modified Bonferroni step-down procedures REGWQ, REGWF, SNK, Duncan, Bonferroni–Holm, and Sidak–Holm. However, these procedures are general and are therefore not tailored to the setting at hand, i.e. conducting tests sequentially in time based on test statistics that are highly autocorrelated. As a consequence, they may be overly conservative.

We note that the development of multiplicity adjustments for correlated tests is an area of considerable interest in the statistical genetics community. Most of the tests in that literature tend to focus on controlling the FDR. More recent developments are discussed in Hastie et al.[29] In the setting where the clinical implications of false positive results are serious, or where the commission of type one error is relatively more severe than type two errors, strict FWER control is desirable. FDR discards significance tests with the highest p-values first. In dense time series analysis such as this one, these will always occur at the edges of a significant region, creating an uneven standard of significance at exactly the places where it is needed. Additionally, FDR is relatively insensitive to the number of comparisons as the FDR holds constant as a percentage of the comparisons. Consequently, whether the time series is sampled at 4 ms, or more sparsely at 100 ms, the power is likely the same. In contrast, with a FWER metric, power will be (in part) a function of the number of comparisons. These are two critical reasons why controlling the FWER as opposed to the FDR may be advantageous. In this work, we examine the relative performance of an FDR-based approach and compare it to the strict control of the FWER that we develop next.

The FWER is defined as the probability of rejecting the null hypothesis for at least one test when all null hypotheses are true. Existing multiple comparisons procedures are overly conservative in the present setting due to their inability to deal with such a large number of highly autocorrelated tests. Since the test statistics at adjacent time points tend to be quite similar, if the groups at one time point are significantly different, there is a high probability that the groups will be significantly different at the next time point (or conversely if they are not different). An alternative adjustment to the alpha level for multiple comparisons can be motivated by considering a probabilistic description of the sequence of t-tests themselves.

Our development will be based on a sequence of autocorrelated test statistics assumed to have a null standard normal distribution. We characterize the evolution of this sequence using an AR(1) process

$$T_t = \rho T_{t-1} + \epsilon_t; \qquad t = 1, \ldots, N$$

Here, $\epsilon_t \sim N\big(0, \big(1 - \rho^2\big)\big)$, and $\rho$ is the autoregressive parameter, which corresponds to the lag-one correlation (and is therefore bounded between -1 and 1). We note that the variance of $T_t$ is governed by the assumption that the variance of $T_t$ is one.

In general, when fitting nonlinear curves with autocorrelated errors to densely sampled time series, we expect the test statistics to evolve slowly, exhibiting minor perturbations from one time period to the next as opposed to abrupt shifts in values. As previously noted, AR(1) processes based on a large, positive autoregressive parameter can be used to provide a parsimonious, intuitively appealing characterization of a gradually changing time series. The assumption of an AR(1) sequence of statistics is not only defensible but also facilitates a tractable approach for FWER control, as outlined in what follows.

Based on the preceding AR(1) model, we have

$$\begin{aligned} &T_t|T_{t-1} \sim N(\rho T_{t-1}, 1-\rho^2),\\ &T_t \sim N(0,1) \end{aligned}$$

It follows that the joint distribution $[T_t, T_{t-1}]$ is bivariate normal with mean zero, variances equal to one, and correlation $\rho$. Therefore, for a given cut-off $q(\alpha^*)$, the actual FWER $\alpha$ can be calculated under the null hypothesis as

$$1 - P\left(\bigcap_{t=1}^{N}(I_t)\right) = 1 - P(I_1)\prod_{t=2}^{N} P(I_t|I_{t-1}) = 1 - P(I_1)P(I_t|I_{t-1})^{N-1} \tag{7}$$

where $I_t$ is the event that $|T_t| < z_{\left(1-\frac{\alpha^*}{2}\right)}$. Here, $z_{\left(1-\frac{\alpha^*}{2}\right)}$ denotes the $100\left(1-\frac{\alpha^*}{2}\right)$ percentile of the standard normal distribution.

The next step is to find a nominal significance level $\alpha^*$ which produces the desired effective FWER $\alpha$. When the correlation between tests is perfect, there is no need for a multiple comparisons adjustment: the formulation above reduces to a single comparison. Conversely, under the assumption of independent tests, this method produces the usual Bonferroni correction. Given a specified level of autocorrelation and a desired $\alpha$, the quantity $\alpha^*$ can be easily computed using the mvtnorm R-package. The impact of the correlation on the nominal level $\alpha^*$, and how it compares to a strict Bonferroni adjustment, is shown in Figure 2. Specific examples of required nominal significance levels are presented in Table 2.

In some settings, strict control of the FWER may not be necessary. Because the specified autocorrelation structure meets the required positivity criterion,[30] the standard FDR procedure based on ranked p-values controls the FDR at the desired value. This can be applied to the bootstrapped estimates, preserving many of the advantages of this approach but with a different multiplicity correction. In Section 4, we compare our version of strict FWER with the FDR approach.

## 4 Simulations

### 4.1 General setting

We conducted a simulation study to evaluate the performance of the proposed methods. All sets of simulations are based on 20 subjects per group and 400 time points (measured every 4 ms = 1600 ms), 1000 sets of bootstrap estimates, and 1000 simulated datasets. With $\alpha$ set at 0.05, we tabulated the relative frequency, at each recorded time point (out of the 1000 simulated datasets), that the null hypothesis is rejected for some specified set of parameters. We refer to this relative frequency as the empirical power if the generating group curves are different, or the per comparison error rate if the curves are the same. When the curves are the same, the FWER controls the probability that the two group curves are declared significantly different for at least one time point along the entire time span of the study. All simulations are performed in R.

We begin with a very simple model that only includes an intercept and autocorrelated errors with $\phi = 0.8$ and $\sigma^2 = 0.025$. Note that although the motivating dataset yields autocorrelations greater than 0.9, we set $\phi = 0.80$ for the simulations for the sake of computational stability (recall that an AR(1) process is stationary if and only if $|\phi| < 1$). Also, the variance is set at a relatively small value to assure simulations stability. This results in a setting where small differences are detectable between the groups. We apply our testing procedure to each time point and use the resulting sequence of test statistics to detect when the two groups are determined to significantly differ. The null hypothesis is rejected 3.4% of the time in this simple case. The 3.4% is both the per comparison error rate and the FWER.

### 4.2 Four-parameter logistic model (target curve)

We next evaluate the performance of our procedure using the four-parameter logistic model in equation (3), where the errors follow an AR(1) process (4) with $\phi = 0.8$ and $\sigma^2 = 0.025$. While generating the simulated data directly from the logistic function ignores some well-known dynamics of eye movement, it reasonably approximates the experimental process and facilitates the evaluation of the functionality of the model. Initially, data are simulated for two groups assuming equal population curves. The parameters are set to values that are realistic based on observed data, with A = 0, B = 0.75, D = 0.0025, and C = 200. Again, the dataset for each subject is comprised of 400 total time points.

As previously noted, accommodating the residual autocorrelation is crucial due to its impact on the standard errors. The importance of this feature of our model can be further demonstrated by evaluating the per comparison error rate of our testing procedure if autocorrelation is ignored. When we ignore the AR(1) error process, and treat the errors as independent, we reject the null hypothesis an average of 75% of the time (median = 0.77), with a minimum of 35% at the mid time point and 100% of the time the final quarter of the time span. That is, we frequently conclude that the two groups are different at various time points, when in fact, the population curves are identical.

When we include the AR(1) error structure in the estimation process, the per comparison error rate (e.g. at any given time slice) has a median of 0.042 with a minimum of 0.0140 (again at the mid-point) and a maximum of 0.059. The first quartile is 0.025 and the third quartile equals 0.050. However, with no multiple comparisons adjustment and 400 comparisons, the FWER is too high at 17.4%.

Thus, we turn to the adjusted nominal significance level for correlated tests. The value relies upon the correlation of the test statistics. As there is a slightly different correlation for each of the 1000 simulated datasets, we choose the maximum of those autocorrelations, 0.9995, to achieve maximum power of the tests. This choice gives $\alpha^* = 0.0024$, which we use throughout all of the tests based on simulations. When $\alpha^*$ is used on the target curve, the resulting family error rate is 0.006. We note this adjustment is still conservative and using a correlation of 0.99995 would have yielded a family error rate of approximately 0.05. Using the FDR procedure yields a corresponding family error rate of 0.018 and a median per comparison error rate of 0.02.

Now that it has been demonstrated that the alpha level correction works well under the null hypothesis, we turn to the case where the curves differ to evaluate empirical power. For the function parameters, we keep A = 0, B = 0.0025, and C = 200 the same for the two groups, yet set the maxima to be one standard deviation apart (based on an empirical estimate obtained using all subjects), with D = 0.750 for group 1 and D = 0.773 for group 2. This would be considered a small clinical effect. To increase the effective difference between the groups, we could increase $\sigma^2$ so that a one standard deviation difference is larger, but the resulting conclusions should be not be impacted. Using the modified alpha level, we obtain 80% empirical power to detect a difference between the groups at time 247 and beyond. Without the adjustment, 80% empirical power is attained at time 234 as shown in Figure 3(a). The FDR method yielded a difference at time 238.

We repeat the same procedure by keeping A = 0, B = 0.80, D = 0.0025, yet allowing for different crossover points that are one standard deviation apart. Specifically, we set the crossover for group 1 at C = 200, and the crossover for group 2 at C = 211. With our multiplicity adjustment, we obtain 80% empirical power to detect a difference between groups at time 190. By time 217, the empirical power drops to less than 80% (Figure 3(b)). With no adjustment, the difference would be detected between times 173 and 234, whereas with the FDR adjustment, differences would be detected between times 184 and 223.

## 4.3 Asymmetric Gaussian model (competitor curve)

Data were also simulated using the asymmetric Gaussian formulation of equation (2), where the errors follow an AR(1) process (4) with $\phi = 0.80$ and $\sigma^2 = -0.005$. As before, the parameters are set to values that are realistic based on observed data, with $A_1 = 0$, $A_2 = 0.02$, $B = 0.1$, $\mu = 200$, $\sigma_1^2 = 25^2$ and $\sigma_2^2 = 40^2$, and the dataset for each subject is comprised of 400 time points.

With no multiple comparisons adjustment, the median per comparison error rate was 0.049, with a minimum of 0.000 (at the upward slope and the downward slope) and a maximum of 0.083 (when the height is reached). However, without this adjustment, the FWER is too high at 23.3%. Using the multiplicity adjustment with correlation 0.9995, the resulting FWER is 0.014 with a corresponding FDR of 0.036.

Again, to assess power, we examine settings where the two curves differ in one parameter. Initially, we consider allowing the means to differ (x-axis shift) by one standard deviation with all other parameters equal. Specifically, we set $A_1 = 0$, $A_2 = 0.02$, $B = 0.1$, $\sigma_1^2 = 25^2$, $\sigma_2^2 = 40^2$, $\rho = 0.80$, $\sigma^2 = 0.005$, and set $\mu_1 = 200$ and $\mu_2 = 203.1$. Using the modified alpha level, we obtain 80% empirical power to detect a difference between groups at time 139. The empirical power drops below 80% between time points 199 and 212 and rises to over 80% from time 212 to 283. For FDR, the corresponding differences are detected between times 173 and 199 and again from 210 to 294.

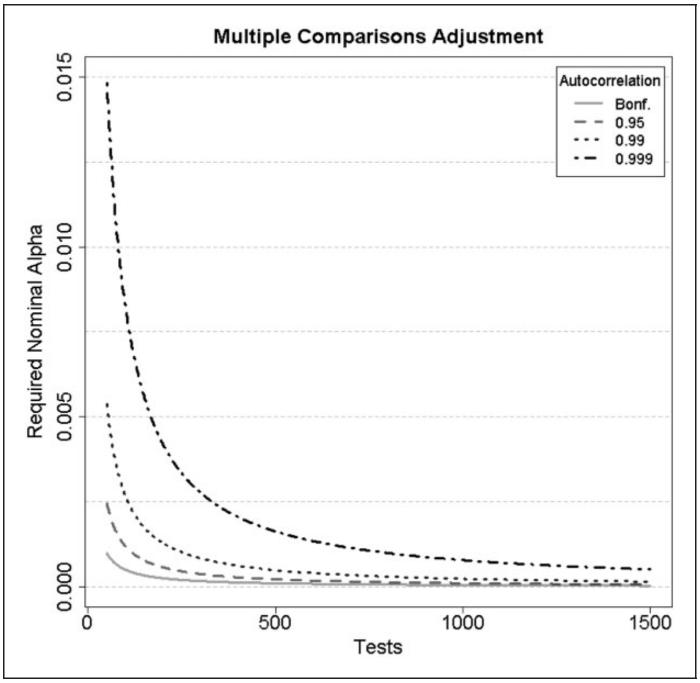

**Figure 2.** Display of the required nominal alpha level by number of tests being conducted for three specific autocorrelation values (0.95, 0.99, 0.999). As the correlation approaches one, the required nominal alpha level increases.

Thus, when one group has a shift at time 200 but the other group has a shift at time 203, a difference is detected as early as time 139, and the difference remains significant until time 283 which can be seen in Figure 3(c). At their peaks, the empirical power briefly drops below 80% for 14 time points.

The next simulation set adjusts the peak, B, by one standard deviation but keeps all other parameters the same. Here, $A_1 = 0$, $A_2 = 0.02$, $\mu = 200$, $\sigma_1^2 = 25^2$, $\sigma_2^2 = 40^2$, $\rho = 0.80$, $\sigma^2 = 0.005$, and the peaks are set at $B_1 = 0.1$ and $B_2 = 0.103$. Using the multiplicity adjustment, we obtain at least 80% empirical power to detect a between-group difference between time 188 and 236 (Figure 3(d)). The FDR approach achieves 80% power between 184 and 242.

## 5 Results

We now return to the CI study outlined in Section 1.2. A complete description is given in Farris-Trimble et al.[13] The overall dataset of 55 participants is comprised of measurements based on 29 CI participants and 26 NH participants. All subjects are observed for 2000 ms with observations every 4 ms, yielding 500 time points per participant. We will compare the time points of detected group deviations using no multiplicity correction, our FWER procedure, and a FDR set to 5%.

The analysis was carried out in R 3.20 using a Windows 7 SP1 machine with Intel Core i7-2600 @3.4 GHz. Using parallel processing, 55 logistic curves were fit using three cores in 16.23 s and 1000 bootstrap iterations took 6.23 s. Then, 55 asymmetric Gaussian curves were fit in 16.15 s with 1000 bootstrap iterations taking 6.92 s. An R package to perform these analyses is available from the first author.

The first comparison of interest is determining when the target curves for NH participants deviate from the target curve for CI patients. The curve of each participant is fit with the nlme command in R. A histogram of the residuals yields an approximately normal shape. A summary of the parameter estimates is given in Table 3. The resulting curves are featured in Figure 4. The proportions generated from the bootstrap estimates satisfy the normality assumption for the t-test. After computing the test statistics, the ACF and PACF figures were examined and they indicate that an AR(1) model adequately describes the correlation of the test statistics. For the target curves, it appears that the maximum threshold values differ (as was found in Farris-Trimble et al.[13] comparing the parameters of the logistic function using ANOVA). We use 1000 bootstrap draws to generate the estimated population curves. In applying our testing procedure, we detect a difference starting at 408 ms and lasting until the end of the trial.

**Table 2.** Required nominal alpha.

| Correlation | Number of tests | Desired alpha | Required nominal alpha | Correlation |
|---|---|---|---|---|
| 0 | 100 | 0.01 | 0.0001 | 0 |
| 0 | 100 | 0.05 | 0.00051 | 0 |
| 0 | 500 | 0.01 | 0.00002 | 0 |
| 0 | 500 | 0.05 | 0.0001 | 0 |
| 0 | 1000 | 0.01 | 0.00001 | 0 |
| 0 | 1000 | 0.05 | 0.00005 | 0 |
| 0.5 | 100 | 0.01 | 0.0001 | 0.5 |
| 0.5 | 100 | 0.05 | 0.00053 | 0.5 |
| 0.5 | 500 | 0.01 | 0.00002 | 0.5 |
| 0.5 | 500 | 0.05 | 0.0001 | 0.5 |
| 0.5 | 1000 | 0.01 | 0.00001 | 0.5 |
| 0.5 | 1000 | 0.05 | 0.00005 | 0.5 |
| 0.9 | 100 | 0.01 | 0.00016 | 0.9 |
| 0.9 | 100 | 0.05 | 0.00086 | 0.9 |
| 0.9 | 500 | 0.01 | 0.00003 | 0.9 |
| 0.9 | 500 | 0.05 | 0.00016 | 0.9 |
| 0.9 | 1000 | 0.01 | 0.00001 | 0.9 |
| 0.9 | 1000 | 0.05 | 0.00008 | 0.9 |
| 0.99 | 100 | 0.01 | 0.00046 | 0.99 |
| 0.99 | 100 | 0.05 | 0.00266 | 0.99 |
| 0.99 | 500 | 0.01 | 0.00009 | 0.99 |
| 0.99 | 500 | 0.05 | 0.00049 | 0.99 |
| 0.99 | 1000 | 0.01 | 0.00004 | 0.99 |
| 0.99 | 1000 | 0.05 | 0.00023 | 0.99 |

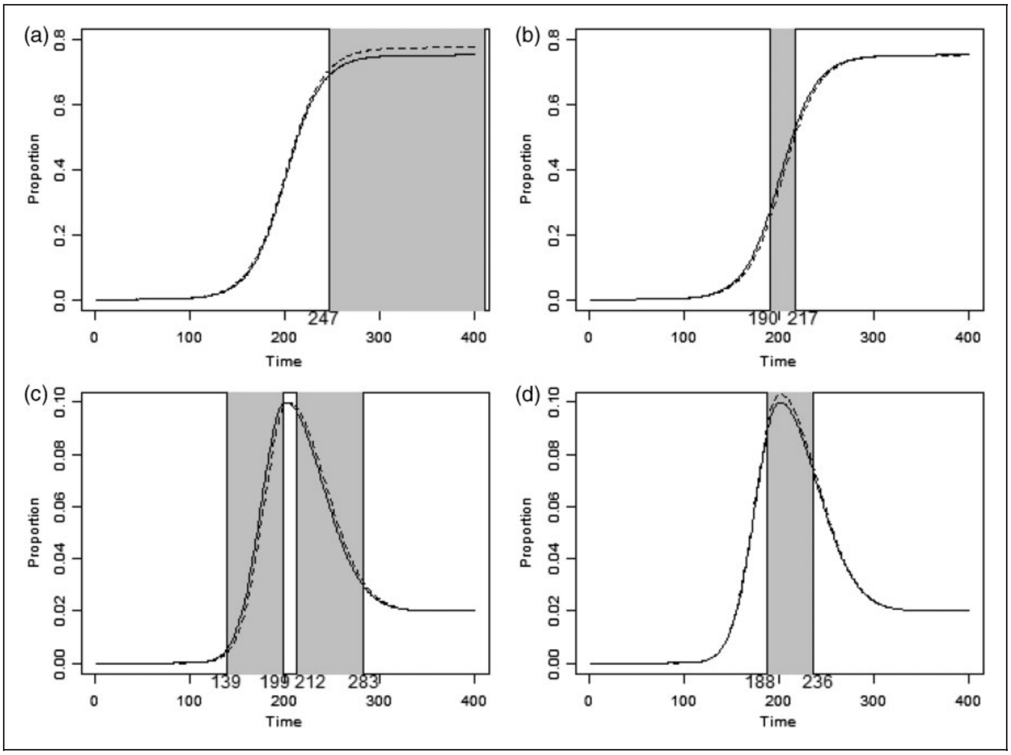

**Figure 3.** Curves generated from simulations. The shaded region demonstrates where significant differences were found between the curves with at least 80% power. The maximum level was shifted by one standard deviation in (a). The crossover point was shifted by one standard deviation in (b). The mean time was shifted by one standard deviation in (c). The height was shifted by one standard deviation in (d).

**Table 3.** Summarized results from model fits.

| Curve | Group | Parameter | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Target | NH | *A* (min) | −0.025 | 0.008 | −0.042 | −0.006 |
| Target | NH | *B* (max) | 0.964 | 0.063 | 0.68 | 0.983 |
| Target | NH | *C* (crossover) | 690.3 | 41.7 | 615.5 | 791.0 |
| Target | NH | *D* (slope) | 0.0019 | 0.0003 | 0.001 | 0.0025 |
| Target | CI | *A* (min) | −0.021 | 0.017 | −0.058 | 0.020 |
| Target | CI | *B* (max) | 0.86 | 0.113 | 0.489 | 0.976 |
| Target | CI | *C* (crossover) | 762 | 59.9 | 649.2 | 921.3 |
| Target | CI | *D* (slope) | 0.0014 | 0.0004 | 0.0008 | 0.0024 |
| Cohort | NH | $A_1$ | −0.002 | 0.002 | −0.0066 | 0.0016 |
| Cohort | NH | $A_2$ | 0.0047 | 0.0042 | −0.002 | 0.0183 |
| Cohort | NH | $B$ | 0.18 | 0.046 | 0.073 | 0.243 |
| Cohort | NH | $\mu$ | 601.58 | 55.16 | 495.56 | 713.66 |
| Cohort | NH | $\sigma_1$ | 124.098 | 23.785 | 94.084 | 189.024 |
| Cohort | NH | $\sigma_2$ | 200.483 | 72.896 | 108.523 | 474.304 |
| Cohort | CI | $A_1$ | −0.0032 | 0.0035 | −0.0096 | 0.0089 |
| Cohort | CI | $A_2$ | 0.024 | 0.015 | 0.003 | 0.060 |
| Cohort | CI | $B$ | 0.181 | 0.047 | 0.024 | 0.250 |
| Cohort | CI | $\mu$ | 661.617 | 65.116 | 550.835 | 822.085 |
| Cohort | CI | $\sigma_1$ | 153.889 | 46.538 | 2.228 | 225.989 |
| Cohort | CI | $\sigma_2$ | 238.598 | 70.376 | 104.1 | 378.102 |
| Rhyme | NH | $A_1$ | −0.0032 | 0.0056 | −0.028 | 0.0008 |
| Rhyme | NH | $A_2$ | 0.008 | 0.006 | −0.0003 | 0.029 |
| Rhyme | NH | $B$ | 0.13 | 0.044 | 0.052 | 0.236 |
| Rhyme | NH | $\mu$ | 574.626 | 51.585 | 490.813 | 693.058 |
| Rhyme | NH | $\sigma_1$ | 117.225 | 24.513 | 79.936 | 170.670 |
| Rhyme | NH | $\sigma_2$ | 205.322 | 62.343 | 65.484 | 327.582 |
| Rhyme | CI | $A_1$ | −0.009 | 0.0139 | −0.0439 | 0.0013 |
| Rhyme | CI | $A_2$ | 0.037 | 0.032 | 0.002 | 0.146 |
| Rhyme | CI | $B$ | 0.135 | 0.052 | 0.017 | 0.231 |
| Rhyme | CI | $\mu$ | 620.733 | 88.533 | 498.059 | 977.332 |
| Rhyme | CI | $\sigma_1$ | 131.378 | 28.403 | 84.35 | 190.247 |
| Rhyme | CI | $\sigma_2$ | 231.047 | 111.277 | 32.696 | 511.574 |
| Unrelated | CI | $A_1$ | −0.0035 | 0.0045 | −0.0237 | 0.0006 |
| Unrelated | CI | $A_2$ | 0.025 | 0.014 | 0.005 | 0.055 |
| Unrelated | CI | $B$ | 0.136 | 0.047 | −0.001 | 0.213 |
| Unrelated | CI | $\mu$ | 629.029 | 114.27 | 496.004 | 1122.605 |
| Unrelated | CI | $\sigma_1$ | 141.925 | 53.07 | 58.311 | 351.098 |
| Unrelated | CI | $\sigma_2$ | 240.287 | 95.344 | 57.548 | 455.282 |

Although it is clear that the target curves deviate, they offer different information from competitors, and typically a complete examination of all of the competitor curves is necessary to evaluate theoretical claims. Thus, the cohort curves are fit using the asymmetric Gaussian function in equation (2); parameter estimates are summarized in Table 3. All of the estimated parameter values, apart from A1, appear distinct, suggesting that the two curves do have different shapes. This can be seen in Figure 4(b). However, it is not clear at which points in time the two curves deviate from each other. The test statistics have an estimated autocorrelation of 0.998, meaning that our adjusted alpha for individual testing should be 0.001. In applying our testing procedure, we can say that the cohort curves are significantly different, beginning at 444 ms through the end of the trial while no adjustment and the FDR both detect a difference slightly earlier at 412 ms. However, the lines crossover at the maximum values and no significant difference is found between 632 and 648 ms without an adjustment, between 632 and 652 ms for the FDR, and between 624 and 656 ms for our adjustment method.

The rhyme curves are similar to the cohort curves; the parameter estimates are again summarized in Table 3. All of the estimated parameter values, apart from A1, appear to be different, suggesting that the two curves do have different shapes. This can be seen in Figure 4(c). The test statistics have an estimated autocorrelation of 0.998, leading to an adjusted alpha of 0.001. Applying our procedure, we can conclude that the rhyme curves are

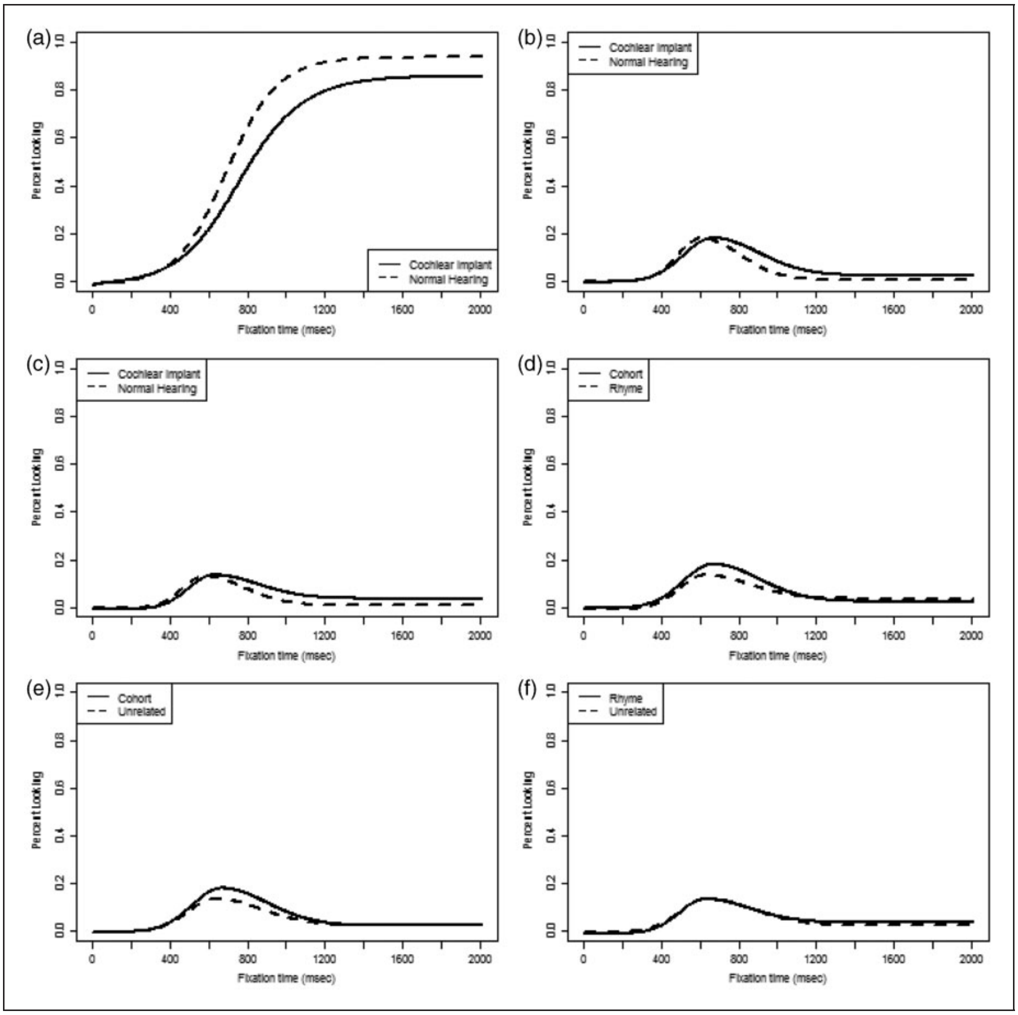

**Figure 4.** Estimated population level curves from data analysis. (a) is the target, (b) is the cohort, (c) is the rhyme, (d) is cohort versus rhyme for the Cochlear Implant group, (e) is cohort versus unrelated for the Cochlear Implant group, and (f) is rhyme versus unrelated for the Cochlear Implant group.

significantly different, beginning at 312 ms with no adjustment, at 324 ms with FDR, and at 424 ms using our method. The difference lasts throughout the trial, except where the curves crossover at the maximum values between 572 and 620 ms for both no adjustment and FDR, and between 552 and 652 ms for our method.

The preceding comparisons focused on between-subject comparison between groups on the same measure (e.g. fixations to cohort, etc.). The previous study also found differences between the CI and NH groups, by comparing the estimated parameters for each subject, although it was unable to characterize the time course precisely. Another important research question which could not be addressed by the previous methods concerns when the cohort and rhyme curves do, and do not, differ significantly from one another. This is a challenging comparison as it is not directly indicated by any of the parameters of the fitted function.

Here, we implement the paired t-test version of our procedure for a within-subject comparison. All of the parameter values, apart from A1, appear to be different, suggesting that the two curves do have different shapes. This can be seen in Figure 4(d). The test statistics have an estimated autocorrelation of 0.999, resulting in an adjusted alpha of 0.002. We conclude that the cohort curve is significantly different from the rhyme curve at 340 ms which lasts until 1044 ms when the curves converge. Then, by 1308 ms, the cohort curve has a slightly, but significantly, lower offset baseline than the rhyme curve. The FDR approach detects differences from 148 to 1080 ms and after 1232 ms, while using no adjustment detects differences from 116 to 1080 ms and after 1228 ms.

We are also interested in comparing the cohort and rhyme words with the unrelated words. Typically, unrelated words are included as a baseline measure of randomness. Additionally, particularly with different subject populations, looking dynamics can vary making the difference between cohorts (or rhymes) and unrelated objects a useful measure. Again, the paired t-test approach is used to make a within-subject comparison. When comparing the cohort fixation curve with the unrelated fixation curve, we first examine the parameter estimates in Table 3. It is interesting that the mean incoming baseline value (A1) and outgoing baseline value (A2) are nearly identical. The other four parameter means are all smaller for the unrelated word than they are for the cohort word. This indicates that the unrelated curve rises more slowly, reaches a lower maximum earlier, and falls at a slower rate than does the cohort word. The test statistics have an estimated autocorrelation of 0.997, resulting in an adjusted alpha of 0.0013. When the tests are performed, the two groups are significantly different from each other from 484 to 1224 ms with no adjustment, from 500 to 1192 ms with FDR, and from 524 to 1136 ms using our adjustment.

The unrelated curve appears to have a different relationship to the rhyme curve than it does to the cohort curve. Again, focusing first on the parameter estimates in Table 3, the outgoing baseline values appear different (A2), and perhaps the incoming slopes as well ($\sigma_1$). The test statistics have an estimated autocorrelation of 0.997, leading to an adjusted alpha of 0.0013. Our test procedure finds differences at the end of the time course with the differences appearing at 1076, 1104, and 1200 ms for unadjusted, FDR, and autocorrelated FWER adjustments, respectively, after which they remain significantly different throughout the remainder of the time course. However, while our autocorrelated FWER adjustment finds no difference early in the time course, the other two methods do ranging from 108 to 380 ms for unadjusted, and 252 to 324 ms for FDR. This result showcases an excellent example of the appropriateness of the autocorrelated FWER adjustment because there should not be a realistic difference between those two curves at those early stages.

## 6 Discussion

The goal of this work was to provide a method to determine when two nonlinear time series curves significantly differ from each other. This was accomplished by fitting individual specific curves and estimating the population level curves through bootstrapping. Specifically, at each time point, the bootstrapped values are used to determine point estimates and variability assessments, which are subsequently utilized in either a two-sample t-test or a paired t-test. The bootstrap approach has the advantage of incorporating subject-specific variability that is missing in many other approaches to analyzing such datasets. Since this approach results in highly autocorrelated test statistics, we have developed a unique multiple comparisons adjustment to control the FWER. Depending on the needs of researchers, we have presented a method for strict FWER control and have established the efficacy of the FDR in cases where small numbers of false positive tests are not a concern.

With the data analysis, we were not only able to conclude that the curves were different between the two groups for all of the various curve types, but we were also able to determine at what points in time these curves significantly deviated from each other. Although this is an important step forward in the analysis of VWP data, some important issues remain that are not accounted for by our approach. It is well known that it takes 100–200 ms for eyes to move, and as a result the raw data are a series of short and ballistic units, not a continuous probabilistic sample of the fixation proportions. The inclusion of the highly autocorrelated AR(1) error process does appear to incorporate the salient dynamics when we consider the averages of such processes across trials.

Traditional results regarding stationary time series do tend to break down for AR(1) processes for correlations approaching one. For series with very high autocorrelation, processes are often treated as random walks. If a random walk is assumed for the sequence of test statistics, then the derivation of the multiplicity correction is changed substantially. Simulations show that the test statistics do not really behave like a random walk, because the variance of the test statistics is not monotonically increasing over time. Instead, series of test statistics tend to approach a steady state by the end of the trial period. Thus, a multiplicity correction under the random walk framework is inappropriate and does not capture the nature of the test statistics that we see in practice. On the other hand, an AR(1) process with high correlation does appear to reflect the essence of the test statistics that we see in practice and allows us to effectively control the FWER.

It may be reasonable to reduce the number of comparisons by selecting time points that are further apart and ignoring some intermediate time points. Considering an analytical framework for a time series that is less fine grained would raise a number of implications. First of all, in our setting, the same time course trajectory could reasonably be fit with the same function (i.e. the estimated function parameters should not be greatly impacted). Second, there would be fewer tests for which to adjust. Third, since the correlation is so high between tests, the correlation between tests should be relatively unaffected with the time points being further apart. These factors could ultimately lead to a more powerful procedure because we could drastically reduce the number of tests being performed without greatly impacting the fitted functions or reducing the autocorrelation of the series of test statistics. Finally, the FWER control approach may be particularly useful when researchers are able to identify a particular time scale on which such control would be helpful, providing both statistical and clinical benefits.

One of the peculiarities of applying this to psycholinguistics is the fact that many common designs have crossed random effects; for example, words are randomly sampled from the language and this is crossed with participants. Currently, this can be handled elegantly by mixed models.[31] Although, crossed random effects are not directly incorporated in the present scheme, it is quite possible to use a classic by-subject/by-item approach, in which the analysis is performed twice, once with the data grouped by participants, and once by items.[32]

In the future, it may be worthwhile to devise an approach for estimating all of the curves simultaneously, since, for each subject, the target and competitor looks are not independent of each other.

## Acknowledgements

The authors wish to express their appreciation to two anonymous referees for valuable feedback which helped to improve the original version of this paper.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the National Institutes on Deafness and Other Communication Disorders, the National Institutes of Health (grant number 2P50DC000242-26A1), from the General Clinical Research Centers Program, Division of Research Resources, National Institutes of Health (grant number RR00059), the Lions Clubs International Foundation, and the Iowa Lions Foundation.

## References

1. McClellend JL and Elman J. The TRACE model of speech perception. *Cogn Psychol* 1986; **18**: 1–86.
2. Shortlist ND. A connectionist model of continuous speech recognition. *Cognition* 1994; **52**: 189–234.
3. McMurray B, Samelson VS, Lee SH, et al. Individual differences in online spoken word recognition: implications for SLI. *Cogn Psychol* 2010; **60**: 1–39.
4. Dahan D, Magnuson JS and Tanenhaus MK. Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cogn Psychol* 2001; **42**: 317–367.
5. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, et al. Integration of visual and linguistic information in spoken language comprehension. *Science* 1995; **268**: 1632–1634.
6. Farris-Trimble A and McMurray B. Test/retest reliability for the visual world paradigm as a measure of real-time lexical processes. *J Speech Lang Hear Res* 2013; **56**: 1328–1345.
7. Salverda AP, Brown M and Tanenhaus MK. A goal-based perspective on eye movements in visual world studies. *Acta Psychol* 2011; **137**: 172–180.
8. Allopenna P, Magnuson JS and Tanenhaus MK. Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *J Memory Lang* 1998; **38**: 419–439.
9. Magnuson JS, Dixon J, Tanenhaus MK, et al. The dynamics of lexical competition during spoken word recognition. *Cogn Sci* 2007; **31**: 1–24.
10. Toscano JC, Anderson ND and McMurray B. Reconsidering the role of temporal order in spoken word recognition. *Psychon Bull Rev* 2013; **20**: 1–7.
11. McMurray B, Tanenhaus MK and Aslin RN. Gradient effects of within-category phonetic variation on lexical access. *Cognition* 2002; **86**: B33–B42.
12. Apfelbaum KS, Blumstein SE and McMurray B. Semantic priming is affected by real-time phonological competition: evidence for continuous cascading systems. *Psychon Bull Rev* 2011; **18**: 141–149.
13. Farris-Trimble A, McMurray B, Cigrand N, et al. The process of spoken word recognition in the face of signal degradation: cochlear implant users and normal-hearing listeners. *J Exp Psychol Hum Percept Perfor* 2014; **40**: 308–327.
14. Mirman D, Yee E, Blumstein SE, et al. Theories of spoken word recognition deficits in aphasia: evidence from eye-tracking and computational modeling. *Brain Lang* 2011; **117**: 53–68.

15. Revill KP and Spieler DH. The effect of lexical frequency on spoken word recognition in young and older listeners. *Psychol Aging* 2012; **27**: 80.
16. Niparko J. *Cochlear implants: principles and practices*, 2nd ed. Philadelphia: Lippincott, Williams, and Wilkins, 2009.
17. Wilson BS. Cochlear implant technology. In: Niparko J, Kirk K, Mellon N, et al (eds) *Cochlear implants: principles and practices*. New York: Lippincott, Williams, and Wilkins, 2000, pp.109–118.
18. McMurray B, Aslin RN, Tanenhaus MK, et al. Gradient sensitivity to within-category variation in words and syllables. *J Exp Psychol Hum Percep Perform* 2008; **34**: 1609–1631.
19. McMurray B, Tanenhaus MK and Aslin RN. Within-category VOT affects recovery from "lexical" garden paths: evidence against phoneme-level inhibition.
*J Memory Lang* 2009; **60**: 65–91.
20. Blumenfeld HK and Marian V. Constraints on parallel activation in bilingual spoken language processing: examining proficiency and lexical status using eye-tracking. *Lang Cogn Process* 2007; **22**: 633–660.
21. Ju M and Luce PA. Falling on sensitive ears: constraints on bilingual lexical activation. *Psychol Sci* 2004; **15**: 314–318.
22. Mirman D, Dixon J and Magnuson JS. Statistical and computational models of the visual world paradigm: growth curves and individual differences. *J Memory Lang* 2008; **59**: 475–494.
23. Oleson JJ, Cavanaugh JE, Tomblin JB, et al. Combining growth curves when a longitudinal study switches measurement tools. *Stat Methods Med Res* 2016; **25**: 2925–2938.
24. Guo W. Functional mixed effects models. *Biometrics* 2002; **58**: 121–128.
25. Kliethermes SA and Oleson JJ. A Bayesian approach to functional mixed-effects modeling for longitudinal data with binomial outcomes. *Stat Med* 2014; **33**: 3130–3146.
26. Pinheiro J, Bates D, DebRoy S, et al. *nlme: linear and nonlinear mixed effects models* (2015, accessed 5 June 2015). R package version 3.1-119, http://CRAN.R-project.org/package=nlme.
27. R Core Team. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2008, accessed 5 June 2015).
28. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995; **57**: 289–300.
29. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2009.
30. Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001; **29**: 1165–1188.
31. Baayen RH, Davidson DJ and Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. *J Memory Lang* 2008; **59**: 390–412.
32. Clark HH. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J Verb Learn Verb Behav* 1973; **12**: 335–359.